# Diagnostic accuracy and measurement sensitivity of digital models for orthodontic purposes: A systematic review



*Published version:*

DOI:10.1016/j.ajodo.2015.06.029



(Article begins on next page)

**Diagnostic accuracy and measurement sensitivity of digital models for orthodontic purposes: A systematic review**

Gabriele Rossini, Simone Parrini, Tommaso Castroflorio, Andrea Deregibus, and Cesare L. Debernardi

**Summary**

Introduction: Our objective was to assess the accuracy, validity, and reliability of measurements obtained from virtual dental study models compared with those obtained from plaster models. Methods: PubMed, PubMed Central, National Library of Medicine Medline, Embase, Cochrane Central Register of Controlled Clinical trials, Web of Knowledge, Scopus, Google Scholar, and LILACs were searched from January 2000 to November 2014. A grading system described by the Swedish Council on Technology Assessment in Health Care and the Cochrane tool for risk of bias assessment were used to rate the methodologic quality of the articles. Results: Thirty-five relevant articles were selected. The methodologic quality was high. No significant differences were observed for most of the studies in all the measured parameters, with the exception of the American Board of Orthodontics Objective Grading System. Conclusions: Digital models are as reliable as traditional plaster models, with high accuracy, reliability, and reproducibility. Landmark identification, rather than the measuring device or the software, appears to be the greatest limitation. Furthermore, with their advantages in terms of cost, time, and space required, digital models could be considered the new gold standard in current practice.

During the past 10 years, models and facial scanning, as well as cone-beam computed tomography (CBCT) technologic advancements, have permitted the complete virtualization of the orthodontic patient, with more accurate 3-dimensional (3D) reconstructions of teeth, bones, and soft tissues.[1] Plaster models are the gold standard in dental diagnosis and treatment procedures. However, they require rigorous archiving and massive physical storage space. Moreover, plaster models are not practical in the long term because of breakage and degradation issues.[2]

Digital study models were introduced commercially in late 1990s. Different technologies can be used to generate digital study casts.[1] This is why standardization

issues are still important. Furthermore, different technologies might account for the differences between conventional plaster and digital models.[2]

The diagnostic accuracy and measurement sensitivity of digital models compared with plaster models are the most investigated issues.[2-4] In 2011, Fleming et al[1] performed a systematic review of the literature focused on the comparisons between measurements on digital models and measurements with digital calipers on plaster models. The authors stated that “digital models offer a high degree of validity when compared to direct measurement on plaster models.” However, the overall quality of the selected studies was variable, with generally inadequate descriptions of the sample populations and rare reports of confidence intervals and standard errors between different techniques.[1] Another review by Luu et al[5] published in 2012 analyzed intrarater reliabilities in terms of mean differences, intraclass correlation coefficients (ICCs), and Pearson correlation coefficients (PCCs) of measurements of digital models compared with gypsum casts. The authors agreed with Fleming et al, stating that validity and reliability for all parameters showed clinically nonsignificant differences. Furthermore, as stated by the authors, only quantitative linear measurements were analyzed, excluding from the review all articles treating qualitative ordinal measures such as orthodontic indexes or scales (ie, Peer Assessment Rating [PAR], American Board of Orthodontics [ABO] Objective Grading System, and Index of Complexity, Outcome, and Need [ICON]).

Considering the velocity of the technologic advancements in scanning and digital models in recent years, the aims of our study were to conduct a systematic review to update the data in these 2 reviews[1,5] and to find answers to a clinical research question related to the use of digital study models in orthodontic practice: What are the accuracy, validity, and reliability of measurements obtained from virtual dental study models compared with those obtained from plaster models?

To try to answer this question, articles about orthodontics indexes or scales were included in our review.

## MATERIAL AND METHODS

### Eligibility criteria

The inclusion and exclusion criteria for this systematic review (Table I) were based on the type of study and were dependent on the clinical research questions. Case reports, reviews, abstracts, author debates, summary articles, and animal studies were excluded from the review process. However, the reference lists of those articles were perused and followed up.

### Information sources, search strategy, and study selection

On November 1, 2014, a systematic search in the medical literature produced between January 2000 and November 2014 was performed to identify all peer-reviewed articles potentially relevant to our questions to be included in the review. The research was performed in the following databases: PubMed, PubMed Central, National Library of Medicine Medline, Embase, Cochrane Central Register of Controlled Clinical Trials, Web of Knowledge, Scopus, Google Scholar, and LILACS.

The same search strategy was used and adapted to the syntax of the different databases. An example of the string used on PubMed is provided in Table II.

A hand search was performed for additional articles in the medical library of the University of Turin in Italy, the authors' personal libraries, and the references of the selected articles. Titles and abstracts were screened to select articles for full-text retrieval.

If there was disagreement between the investigators, inclusion of the study was confirmed by mutual agreement.The studies were selected for inclusion independently by 2 authors (G.R. and S.P.). All decisions on the definitive inclusion of a potentially relevant article were made by consensus.

From the selected articles, the investigators independently extracted data answering the clinical research questions.

### Data items and collection

To extract data from the selected articles, we used a table to report for every article sample size, measurements evaluated, mean differences, P values, standard errors, and confidence intervals (Appendix 1). A separate table was used to evaluate the results of articles analyzing the ABO Objective Grading System score (Appendix 2).

Reliability indexes were included in a separate table for every article: sample size, ICC values, PCCs, and other reliability methods, if calculated (Appendix 3).

All studies were assessed separately; in cases of divergent assessments with regard to the assignment of strengths and weaknesses, consensus was reached by discussion.

**Risk of bias and quality assessment in the studies**

According to the Centre for Reviews and Dissemination at the University of York in the United Kingdom[6] and the preferred reporting items for systematic reviews and meta-analyses (PRISMA)[7] statements, evaluation of methodologic quality gives an indication of the strength of the evidence in the study because flaws in the design or the conduct of a study can result in biases. The Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool was used to rate the methodologic quality of the articles and to assess the level of evidence for the conclusions of this review (Tables III and IV, Fig 1).[42,43]

**Summary measures and approach to synthesis**

Clinical heterogeneity of the included studies was evaluated by assessing the treatment protocols: participants and settings, index tests, and measurement techniques. For accuracy of measurements, mean differences, with standard errors and 95% confidence intervals, were reported when available. For reliability, ICC values and PCCs were extracted from the studies.

## RESULTS

**Study selection and characteristics**

A search with the terms shown in Table II gave the following results: PubMed yielded 475 publications; PubMed Central, 2880 publications; Cochrane Central Register, 11 publication; Web of Knowledge, 392 publications; Scopus, 458 publications; and LILACS, 15 publications. In addition, 15 articles were identified through hand searching. The selection process is illustrated in the PRISMA flow diagram (Fig 2).

Overlapping data among the databases were obtained. Application of the inclusion and exclusion criteria and follow-up identified 35 relevant publications.[8-41,44]

Different methods were used to evaluate the reliability of digital models: 31 studies[8,9,11-18,20-23,25-41,44] compared digital models with plaster models, 2 studies[10,19] used CBCT, 1 study[24] analyzed digital models obtained by dedicated software, and 1 study[17] used both CBCT and plaster models.

**Risk of bias in studies**

Quality assessment was performed independently by 2 investigators (G.R. and S.P.). Disagreements were resolved by discussions among all authors. The researchers were not blinded to the authors or the results of the research. According to the QUADAS-2 tool (Tables III and IV; Fig 2), the overall quality of these studies was high, and any applicability concern was identified. The subject selection method may have introduced high bias in several studies in which recruitment was not randomized[8,9,11,12,14-24,29,32-35,37,39,41,44] or it was influenced by inclusion criteria that might have altered the results of the measurements.[8,10,12,13,16-18,21,29,33,37,38,40,41] The methods used to perform the measurements were described well enough to allow the reviewer to answer the review question in all studies. However, only 1 study declared that all examiners were blinded to the identity of the models by assigning them a new random number for each measurement series.[15] Thus, the lack of a proper blinding procedure was the most recurrent possible source of bias in the sample.

**Results of individual studies, meta-analysis, and additional analyses**

Quantitative measurements between CBCT-obtained casts or digital casts and plaster casts were grouped into 7 categories: analysis of transverse dimensions, other miscellaneous linear measurements, tooth size, Bolton ratio, arch length and crowding and irregularity index, interarch occlusal features, and occlusal indexes (Appendix 1). The ABO Objective Grading System scoring was analyzed independently (Appendix 2). Repeatability and reproducibility coefficients were also investigated (Appendix 3).

**Transverse dimensional measurements**

Five hundred two models in 8 studies were measured for various transverse dimensions.[9,12,19,20,22,28,31,38] These studies assessed the agreement between transverse dimensional measurements obtained from digital and plaster models. One study analyzed CBCT images.[19] The dimensions included mandibular and maxillary intercanine, interpremolar, and intermolar widths. Mean discrepancies between the approaches ranged from 0.02 mm, measured at the mandibular first premolars,[19] to1.46mm,measuredatthemandibularfirst molars.[12] No significant differences were found between the digital and plaster models when considering the transverse dimensions of the mandibular and maxillary arches.

**Tooth size**

Differences in individual tooth sizes with digital and direct methods were measured in the mesiodistal and vertical dimensions in 11 studies.[9,10,15,22,23,28,29,31-33,40] Mean differences in tooth dimensions of 0.01 to 0.45 mm were reported. A difference in the measurements of canine and molar heights of 0.1 mm was found by Keating et al.[31]

**Bolton ratio**

Comparisons of Bolton tooth size analyses were performed between the plaster and digital models[9,15-17,34,35,41] and between the CBCT and digital models.[21] Naidu and Freer[15] reported discrepancies of 0.91 for the anterior Bolton ratio and 0.21 for the overall Bolton ratio; these were statistically significant. Nalcaci et al[16] showed statistically significant differences between the 2 methods (1.8 and 1.6 for anterior maxillary and mandibular Bolton ratios, respectively; 4.3 and 4.5 for overall maxillary and mandibular Bolton ratios, respectively). Hajeer[9] found differences between plaster and digital models of 0.83 for anterior Bolton ratio and 0.87 for overall Bolton ratio. Wiranto et al[17] showed a difference of 0.4 mm for anterior Bolton ratio, whereas Stevens et al[35] recorded a difference of 0.6 mm for the same measurement.

Regarding overall Bolton measurements, Mullen et al[34] reported a mean difference of 0.05 mm, similar to the differences found by Wiranto et al[17] (0.75 mm with Digimodels [Orthoproof, Nieuwegein, The Netherlands], and 1.03 mm with Lava [3M ESPE, Seefeld, Germany]) and Stevens et al[35] (0.38 mm using emodels), whereas Tomassetti et al[41] found differences of about 1.02 to 1.2 mm using OrthoCad (CADENT Inc, Fairview, NJ). Tarazona et al[21] obtained similar results comparing measurements on CBCT models and a 2-dimensional digital measuring technique (anterior Bolton ratio, 0.15 mm; overall Bolton ratio, 0.06 mm).

**Arch length, crowding, and irregularity index**

Differences from 12 studies[11,12,14,22,24-26,29,32,34,35,38] (636 models) regarding space analysis, crowding, and irregularity index even statistically significant differences were not clinically significant except for arch perimeter[12] and available mandibular space.[38] Regarding arch length, differences between the techniques ranged from 0.0712 to 1.47 mm.[34] The differences between the measurement of crowding varied from 0.19 to 1.19 mm.[29] Stevens et al[35] showed a high discrepancy in the irregularity index of 3.7 mm.

**Interarch occlusal features**

Agreement between overjet and overbite measurements from digital and plaster models was considered in 6 studies.[12,25,28,35,38,40] One study compared overjet and overbite measurements between CBCT images and digital models obtained by 3D scans of traditional impressions.[24] All studies showed excellent agreement for overjet and overbite measurements, with the results ranging between 0.01[38] and 0.50 mm.[12] Stevens et al[35] confirmed these results for the measurements of centerline discrepancy and posterior crossbite.

**Occlusal indexes**

Different occlusal indexes, such as the PAR, ICON, and ABO scoring system, were evaluated in 7 studies.[12,27,30,35-37,44] Comparing manual and digital measurements, high agreement was recorded with respect to both PAR[35,36] and ICON.[37] In relation to the ABO score, 4 studies reported minimal differences between the measurements on the 2 kind of models.[12,30,37,44] Three studies[12,30,44] reported significant differences with respect to occlusal contacts, and Okunami et al[44] showed significant discrepancies also in occlusal relationships. Alignment,[30,37] overjet,[12,30] and total score[12,44] measurements were significantly different in several studies. Costalos et al[37] observed differences also for buccolingual inclination scores.

**Miscellaneous linear measurements**

Miscellaneous linear measurements were assessed by 6 studies.8,12,22,25,31,39 Im et al,12 Nouri et al,8 and Sjogren et al[25] considered tip, rotation, and torque of various teeth. The main discrepancies between plaster and digital models were observed for the rotation and the tip of the maxillary lateral incisors.[25] Measurements of torque were made by Im et al[12]: they showed a greater discrepancy for the measurement related to the maxillary lateral incisor, whereas the most reliable measure was that related to the mandibular central incisor. Bell et al[39] and Keating et al[31] assessed miscellaneous linear measurements defined by occlusal landmarks. The results from these studies showed good accuracy of measurements, with mean discrepancies of 0.14 and 0.27 mm, respectively. El-Zanaty et al[22] evaluated the depth of the palate and found no significant differences between plaster and digital models.

**Examiners' and techniques reliability**

ICC parameters were analyzed in 14 studies8,10-15,17-21,29,36; the Houston coefficient was analyzed in 1 study[9]; and in another study, the Cronbach a and the McNemar test values were evaluated.[16]

Intraexaminer reliability

Three studies analyzed intraexaminer reliability.[16,20,36] Nalcaci et al[16] considered both the Cronbach a and the McNemar test results, and reported high intraobserver reliability in measuring the Bolton index. The systematic errors observed by Sousa et al[20] related to arch width and arch length measurements were not statistically or clinically significant. Mayers et al[36] obtained similar results for PAR scores in both techniques.

**Interexaminer reliability**

Regarding interexaminer reliability, De Waard et al,[10] using ICCs, reported very good reliability for almost every analyzed measurement on 3 types of digital models. The worst results were recorded for intercanine and intermolar mandibular distances on CBCT casts. Radeke et al,[13] performing a Bland-Altman test, showed no significant differences in interexaminer reliability comparing 3 categories of operators. High correlation coefficients were also reported by Naidu and Freer[15] and Wiranto et al.[17]

**Intratechnique reliability**

High ICC values were obtained by Im et al,[12] whereas Abizadeh et al[18] observed statistically significant variance ratios for the intratechnique systematic errors. Plaster values were superior in reproducibility for 8 measurements, and digital ones were superior for 3 measurements.

**Intertechnique reliability**

Eight studies evaluated the reliability between plaster and digital casts.[8,9,11,14,18,19,21,29] Hajeer[9] calculated the Houston coefficient of reliability between methods, with results between 0.96 and 0.99 (high reliability). Burns et al[11] analyzed the Pearson test and ICC values for contact point displacements and found high reliability. Akyalcin et al[14] assessed a 0.99 ICC for crowding in both the maxillary and the mandibular arches comparing caliper measurements with digital models and CBCT casts. High reliability was confirmed by Lightheart et al,[19] Tarazona et al,[21] and Goonewardene et al[29] analyzing, respectively, arch diameters, Bolton index, and space analysis and irregularity index. Abizadeh et al[18] found good reliability between methods except for mandibular intercanine width, where digital values were significantly higher than plaster ones. Nouri et al[8] stated that the reliability of the measurements performed on 3D casts was higher than that for measurements on plaster models.

## DISCUSSION

**Summary of evidence**

This systematic review showed high reliability and reproducibility, as well as the absence of clinically significant differences, of measurements on digital and CBCT models compared with plaster casts. With respect to the previous review by Fleming et al,[1] this review is an update with more evaluated articles with a lower risk of bias.

We included both retrospective and prospective studies, of which 11 were randomized.[10,13,25-28,30,31,36,38,40] According to the QUADAS-2 tool (Tables III and IV; Fig 2), the overall quality of the sample was high.[42,43] All studies, except one, had a strong limitation of the absence of clues about the blinding procedures of the sample and the investigators.[15] However, the design of these studies and the particular subjects investigated made it virtually impossible to blind the operators.

The analyzed articles had great variability and heterogeneity of measuring methods. Another source of variability is represented by the sample selection process covering a period of 14 years, during which a great evolution involved both scanning devices and measuring software. Regarding the selection procedures, the most important source of bias was the definition of inclusion criteria that may influence the results of measurements.[8,10,12,13,16-18,21,29,33,37,38,40,41] However, any concerns about the applicability of these procedures were registered by the authors to answer the review questions.

No significant differences were observed for most of the studies in all measured parameters. The comparisons between digital and plaster models resulted in significant differences for only a few studies. The significant differences were related to mandibular first and second intermolar widths,[12,28,38] tooth size discrepancies,[10,15,28,40] arch perimeter,[12,26,29,34,35,38] Bolton analysis,[15,17] and torque, tip, and rotation.[8,12] All of these alterations could be due to the low precision of proximal surfaces, which makes the positioning of landmarks more difficult.[12] Regarding the Bolton analysis, registered differences were not clinically significant, agreeing with previous studies that reported, even on plaster models, that intraexaminer variability for the Bolton measurements was greater than 1.5 mm.[45] Other reasons leading to significant differences between plaster and digital casts could be (1) a more accentuated correction of tooth position adjustment and (2) the increased accuracy of the virtual setup compared with the manual one, especially when the casts include teeth with altered crown morphology. Furthermore, it is possible that superimposition of moving objects may alter the shapes of the virtual models, unlike plaster models for which this is impossible.[12,38] However, in the vast majority of the sample, the magnitude of differences was less than 0.5 mm, which could be statistically significant but clinically irrelevant.[45]

Regarding the ABO Objective Grading System, the analyzed studies reported that digital casts cannot substitute for plaster ones for calculation of this index.[12,30,37,44] The main reasons were incorrect articulation of the digital casts, bias in the software used to calculate the ABO Objective Grading System, and difficulties in landmark identification on the digital casts.

In the most recent study on this topic in 2014, Im et al[12] observed an improvement in the ABO Objective Grading System scoring with digital models, even if there were still significant problems. The most troublesome parameters were occlusal contacts,[12,30,44] overjet, and alignment.[30] The total Objective Grading System scores were significantly different in these 3 studies in relation to the amount of discrepancies in the single features. However, differences between plaster and digital casts appeared to be irrelevant when considering their influence on the possibility to succeed with the ABO Objective Grading System. Furthermore, as stated by Murakami et al[46] in 2007, an important source of error for grading models with the ABO scale could be the lack of training for designated examiners.

The findings in this systematic review related to examiner and technique reliability agree with the statements of Luu et al[5] in their review from 2012.

Both the randomized trials[10,13,36] and the nonrandomized trials[8,9,11,12,14-21,29] observed differences between the digital and plaster casts in reproducibility and reliability in the normal range of accepted errors.

Only De Waard et al[10] observed relevant differences in reliability between measurements. More precisely, the authors stated that models from CBCT are not sufficiently reliable in reconstructing the occlusal surfaces when producing 3D casts. Thus, digital models obtained by an intraoral scanner or a 3D scan of plaster models should be better than CBCT models.

Digital models have shown high accuracy, efficacy, and effectiveness when compared with the gold standard evaluation.

**Limitations**

The main limitation is represented by point identification rather than measuring device or software. Therefore, with enhancement of direct digital superimposition techniques and digital point recognition, digital modeling may replace plaster models as the gold standard.[30]

The methodologic quality of these studies was variable but moderate on average; thus, considering the size of the analyzed sample, the overall evidence from our review should be considered high. Differences in the impression procedures and model reconstruction processes may have contributed to some inconsistencies reported in these trials. However, the mean discrepancy between measurements on digital and plaster models was significantly low. In almost all analyzed studies, the differences were not considered clinically significant.

Digital models have several advantages in terms of cost, time, and space required with respect to plaster models.[47] A further potential advantage of digital models is the possibility of performing 3D measurements of tooth positions. In particular, the evaluation of the inclination of every tooth on plaster models is unreliable and cumbersome. Digital models may be virtually sectioned to permit a more reliable estimation of long axis positions. Furthermore, 3D mapping of tooth movements may be possible by superimposing dental changes on stable reference structures with nondestructive digital manipulation and sectioning techniques.

## CONCLUSIONS

1. Our sample had overall good quality, with many articles, including randomized studies. The most recurrent sources of bias were related to the patient selection processes and the lack of proper blinding procedures. However, a high level of evidence can be obtained by this review.
2. The most recurrent sources of error for measurements on digital models were landmark positions and the low accuracy of interproximal surfaces, but these did not influence the clinical outcome.
3. Digital models are still lacking in accuracy regarding the ABO Objective Grading System measurements, even if this lower accuracy has no influence on its grading success.
4. Digital models are as reliable as traditional plaster models, with high accuracy, reliability, and reproducibility. Furthermore, with their advantages in terms of cost, time, and space required, they could be considered the new gold standard in current practice.

## REFERENCES


1. Fleming PS, Marinho V, Johal A. Orthodontic measurements on digital study models compared with plaster models: a systematic review. Orthod Craniofac Res 2011;14:1-16.
2. Akyalcin S, Cozad BE, English JD, Colville CD, Laman S. Diagnostic accuracy of impression-free digital models. Am J Orthod Dentofacial Orthop 2013;144:916-22.
3. Wriedt S, Schmidtmann I, Niemann M, Wehrbein H. Digital 3D image of bimaxillary casts connected by a vestibular scan. J Orofac Orthop 2013;74:309-18.
4. Kasparova M, Grafova L, Dvorak P, Dostalova T, Prochazka A, Eliasova H, et al. Possibility of reconstruction of dental plaster cast from 3D digital study models. Biomed Eng Online 2012; 12:49.
5. Luu NS, Nikolcheva LG, Retrouvey JM, Flores-Mir C, El-Bialy T, Carey JP, et al. Linear measurements using virtual study models. Angle Orthod 2012;82:1098-106.
6. Centre for Reviews and Dissemination. Systematic reviews – CRD's guidance for undertaking reviews in health care. York, United Kingdom: University of York; 2009.
7. Moher D, Liberati A, Tetzlaff J, Altman DG. PRISMA group. Preferred reporting items for systematic reviews and metaanalyses: the PRISMA statement. PLoS Med 2009;6:e1000097.
8. Nouri M, Abdi AH, Farzan A, Mokhtarpour F, Baghban AA. Measurement of the buccolingual inclination of teeth: manual technique vs 3-dimensional software. Am J Orthod Dentofacial Orthop 2014;146:522-9.
9. Hajeer MY. Assessment of dental arches in patients with Class II division 1 and division 2 malocclusions using 3D digital models in a Syrian sample. Eur J Paediatr Dent 2014;15:151-7.
10. De Waard O, Rangel FA, Fudalej PS, Bronkhorst EM, KuijpersJagtman AM, Breuning KH. Reproducibility and accuracy of linear measurements on dental models derived from cone-beam computed tomography compared with digital dental casts. Am J Orthod Dentofacial Orthop 2014;146:328-36.
11. Burns A, Dowling AH, Garvey TM, Fleming GJ. The reliability of Little's irregularity index for the upper dental arch using three dimensional (3D) digital models. J Dent 2014;42:1320-6.
12. Im J, Cha JY, Lee KJ, Yu HS, Hwang CJ. Comparison of virtual and manual tooth setups with digital and plaster models in extraction cases. Am J Orthod Dentofacial Orthop 2014;145:434-42.

13. Radeke J, von der Wense C, Lapatki BG. Comparison of orthodontic measurements on dental plaster casts and 3D scans. J Orofac Orthop 2014;75:264-74.
14. Akyalcin S, Dyer DJ, English JD, Sar C. Comparison of 3-dimensional dental models from different sources: diagnostic accuracy and surface registration analysis. Am J Orthod Dentofacial Orthop 2013;144:831-7.
15. Naidu D, Freer TJ. Validity, reliability, and reproducibility of the iOC intraoral scanner: a comparison of tooth widths and Bolton ratios. Am J Orthod Dentofacial Orthop 2013;144:304-10.
16. Nalcaci R, Topcuoglu T, Ozturk F. Comparison of Bolton analysis and tooth size measurements obtained using conventional and three-dimensional orthodontic models. Eur J Dent 2013;7(Suppl 1):S66-70.
17. Wiranto MG, Engelbrecht WP, Tutein Nolthenius HE, van der Meer WJ, Ren Y. Validity, reliability, and reproducibility of linear measurements on digital models obtained from intraoral and cone-beam computed tomography scans of alginate impressions. Am J Orthod Dentofacial Orthop 2013;143:140-7.
18. Abizadeh N, Moles DR, O'Neill J, Noar JH. Digital versus plaster study models: how accurate and reproducible are they? J Orthod 2012;39:151-9.
19. Lightheart KG, English JD, Kau CH, Akyalcin S, Bussa HI Jr, McGrory KR, et al. Surface analysis of study models generated from OrthoCAD and cone-beam computed tomography imaging. Am J Orthod Dentofacial Orthop 2012;141:686-93.
20. Sousa MV, Vasconcelos EC, Janson G, Garib D, Pinzan A. Accuracy and reproducibility of 3-dimensional digital model measurements. Am J Orthod Dentofacial Orthop 2012;142:269-73.
21. Tarazona B, Llamas JM, Cibrian R, Gandıa JL, Paredes V. Evaluation of the validity of the Bolton Index using cone-beam computed tomography (CBCT). Med Oral Patol Oral Cir Bucal 2012;17: e878-83.
22. El-Zanaty HM, El-Beialy AR, Abou El-Ezz AM, Attia KH, ElBialy AR, Mostafa YA. Three-dimensional dental measurements: an alternative to plaster models. Am J Orthod Dentofacial Orthop 2010;137:259-65.
23. Horton HM, Miller JR, Gaillard PR, Larson BE. Technique comparison for efficient orthodontic tooth measurements using digital models. Angle Orthod 2010;80:254-61.
24. Kau CH, Littlefield J, Rainy N, Nguyen JT, Creed B. Evaluation of CBCT digital models and traditional models using the Little's index. Angle Orthod 2010;80:435-9.

25. Sjogren AP, Lindgren JE, Huggare JA. Orthodontic study cast analysis reproducibility of recordings and agreement between conventional and 3D virtual measurements. J Digit Imaging 2010; 23:482-92.
26. Leifert MF, Leifert MM, Efstratiadis SS, Cangialosi TJ. Comparison of space analysis evaluations with digital models and plaster dental casts. Am J Orthod Dentofacial Orthop 2009;136:16.e1-4:discussion 16.
27. Veenema AC, Katsaros C, Boxum SC, Bronkhorst EM, KuijpersJagtman AM. Index of complexity, outcome and need scored on plaster and digital models. Eur J Orthod 2009;31:281-6.
28. Watanabe-Kanno GA, Abr~ao J, Miasiro H Junior, Sanchez-Ayala A, Lagravere MO. Reproducibility, reliability and validity of measurements obtained from Cecile3 digital models. Braz Oral Res 2009; 23:288-95.
29. Goonewardene RW, Goonewardene MS, Razza JM, Murray K. Accuracy and validity of space analysis and irregularity index measurements using digital models. Aust Orthod J 2008;24: 83-90.
30. Hildebrand JC, Palomo JM, Palomo L, Sivik M, Hans M. Evaluation of a software program for applying the American Board of Orthodontics objective grading system to digital casts. Am J Orthod Dentofacial Orthop 2008;133:283-9.
31. Keating AP, Knox J, Bibb R, Zhurov AI. A comparison of plaster, digital and reconstructed study model accuracy. J Orthod 2008; 35:191-201.
32. Redlich M, Weinstock T, Abed Y, Schneor R, Holdstein Y, Fischer A. A new system for scanning, measuring and analyzing dental casts based on a 3D holographic sensor. Orthod Craniofac Res 2008;11: 90-5.
33. Cha BK, Choi JI, Jost-Brinkmann PG, Jeong YM. Applications of three-dimensionally scanned models in orthodontics. Int J Comput Dent 2007;10:41-52.
34. Mullen SR, Martin CA, Ngan P, Gladwin M. Accuracy of space analysis with emodels and plaster models. Am J Orthod Dentofacial Orthop 2007;132:346-52.
35. Stevens DR, Flores-Mir C, Nebbe B, Raboud DW, Heo G, Major PW. Validity, reliability, and reproducibility of plaster vs digital study models: comparison of Peer Assessment Rating and Bolton analysis and their constituent measurements. Am J Orthod Dentofacial Orthop 2006;129:794-803.
36. Mayers M, Firestone AR, Rashid R, Vig KW. Comparison of peer assessment rating (PAR) index scores of plaster and computer based digital models. Am J Orthod Dentofacial Orthop 2005; 128:431-4.

37. Costalos PA, Sarraf K, Cangialosi TJ, Efstratiadis S. Evaluation of the accuracy of digital model analysis for the American Board of Orthodontics objective grading system for dental casts. Am J Orthod Dentofacial Orthop 2005;128:624-9.
38. Quimby ML, Vig KW, Rashid RG, Firestone AR. The accuracy and reliability of measurements made on computer-based digital models. Angle Orthod 2004;74:298-303.
39. Bell A, Ayoub AF, Siebert P. Assessment of the accuracy of a three dimensional imaging system for archiving dental study models. J Orthod 2003;30:219-23.
40. Santoro M, Galkin S, Teredesai M, Nicolay OF, Cangialosi TJ. Comparison of measurements made on digital and plaster models. Am J Orthod Dentofacial Orthop 2003;124:101-5.
41. Tomassetti JJ, Taloumis LJ, Denny JM, Fischer JR Jr. A comparison of 3 computerized Bolton tooth-size analyses with a commonly used method. Angle Orthod 2001;71:351-7.
42. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, ReitsmaJB,etal.QUADAS-2:arevisedtoolforthequalityassessment of diagnostic accuracy studies. Ann Intern Med 2011;155:529-36.
43. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Medical Research Methodology 2003;3:25.
44. Okunami TR, Kusnoto B, BeGole E, Evans CA, Sadowsky C, Fadavi S. Assessing the American Board of Orthodontics objective grading system: digital vs plaster dental casts. Am J Orthod Dentofacial Orthop 2007;131:51-6.
45. Shellhart WC, Lange DW, Kluemper GT, Hicks EP, Kaplan AL. Reliability of the Bolton tooth-size analysis when applied to crowded dentitions. Angle Orthod 1995;65:327-34.
46. Murakami K, Deguchi T, Hashimoto T, Imai M, Miyawaki S, Takano-Yamamoto T. Need for training sessions for orthodontists in the use of the American Board of Orthodontics objective grading system. Am J Orthod Dentofacial Orthop 2007; 132:427.e1-6.
47. Hajeer MY, Millett DT, Ayoub AF, Siebert JP. Applications of 3D imaging in orthodontics: part II. J Orthod 2004;31: 154-62.

**Tables**

Table I. Study selection criteria

| Inclusion criteria | Exclusion criteria |
|---|---|
| Prospective and retrospective original studies analyzing treated and untreated orthodontic patients with or without malocclusion | Studies of patients with genetic syndromes and severe facial malformations |
| Studies analyzing measurements made on digital and plaster models | Studies with fewer than 10 patients |
| Studies with adequate statistical analysis | Case reports |
| | Reviews |
| | Abstracts |
| | Author debates |
| | Summary articles |
| | |

Table II. Search strategy

| Database | | Search strategy |
|---|---|---|
| PubMed, PubMed Central, Scopus, Web of Knowledge, Embase, National Library of Medicine Medline | | ((3d OR digital OR intraoral OR electronic or computer* OR software) AND (impression* OR model* OR cast* OR scanner* OR cad/cam OR cad cam OR cad-cam)) AND (orthodontics OR orthod*) AND (accuracy OR precision OR effic* OR limitat*) |

*The asterisk is a PubMed operator for optimizing the search query.

Table III. Risk of bias according to the QUADAS-2 tool

| | Risk of bias | | | |
|---|---|---|---|---|
| | Patients selection | Index test | Reference standard | Flow and timing |
| Nouri et al,[8] 2014 | - | ? | ? | + |
| Hajeer,[9] 2014 | + | ? | ? | + |
| De Waard et al,[10] 2014 | - | ? | ? | + |
| Burns et al,[11] 2014 | ? | ? | ? | + |
| Im et al,[12] 2014 | - | ? | ? | + |
| Radeke et al,[13] 2014 | - | + | + | + |
| Akyalcin et al,[14] 2013 | ? | ? | ? | + |
| Naidu and Freer,[15] 2013 | + | ? | ? | + |
| Nalcaci et al,[16] 2013 | - | ? | ? | + |
| Wiranto et al,[17] 2013 | - | ? | ? | + |
| Abizadeh et al,[18] 2012 | - | ? | ? | + |
| Lightheart et al,[19] 2012 | ? | ? | ? | + |
| Sousa et al,[20] 2012 | ? | ? | ? | + |
| Tarazona et al,[21] 2012 | - | ? | ? | + |
| El-Zanaty et al,[22] 2010 | ? | + | ? | + |
| Horton et al,[23] 2010 | ? | ? | ? | + |
| Kau et al,[24] 2010 | ? | ? | ? | + |
| Sjogren et al,[25] 2010 | + | ? | ? | + |
| Leifert et al,[26] 2009 | + | ? | ? | + |
| Veenema et al,[27] 2009 | + | ? | ? | + |
| Watanabe-Kanno et al,[28] 2009 | + | ? | ? | + |
| Goonewardene et al,[29] 2008 | - | ? | ? | + |
| Hildebrand et al,[30] 2008 | + | ? | ? | + |
| Keating et al,[31] 2008 | + | ? | ? | + |
| Redlich et al,[32] 2008 | ? | ? | ? | + |
| Cha et al,[33] 2007 | - | ? | ? | + |
| Mullen et al,[34] 2007 | ? | ? | ? | + |
| Stevens et al,[35] 2006 | + | ? | ? | + |
| Mayers et al,[36] 2005 | + | ? | ? | + |
| Costalos et al,[37] 2004 | - | ? | ? | + |
| Okunami et al,[30] 2004 | ? | ? | ? | + |
| Quimby et al,[38] 2004 | - | ? | ? | + |
| Bell et al,[39] 2003 | ? | ? | ? | + |
| Santoro et al,[40] 2003 | - | ? | ? | + |
| Tomassetti et al,[41] 2001 | - | ? | ? | + |

-, High risk of bias; +, Low risk of bias; ?, Unclear

Table IV. Applicability concerns according to the QUADAS-2 tool

| | Applicability concerns | | |
|---|---|---|---|
| | Patient selection | Index test | Reference standard |
| Nouri et al,[8] 2014 | + | + | + |
| Hajeer,[9] 2014 | + | + | + |
| De Waard et al,[10] 2014 | + | + | + |
| Burns et al,[11] 2014 | + | + | + |
| Im et al,[12] 2014 | + | + | + |
| Radeke et al,[13] 2014 | + | + | + |
| Akyalcin et al,[14] 2013 | + | + | + |
| Naidu and Freer,[15] 2013 | + | + | + |
| Nalcaci et al,[16] 2013 | + | + | + |
| Wiranto et al,[17] 2013 | + | + | + |
| Abizadeh et al,[18] 2012 | + | + | + |
| Lightheart et al,[19] 2012 | + | + | + |
| Sousa et al,[20] 2012 | + | + | + |
| Tarazona et al,[21] 2012 | + | + | + |
| El-Zanaty et al,[22] 2010 | + | + | + |
| Horton et al,[23] 2010 | + | + | + |
| Kau et al,[24] 2010 | + | + | + |
| Sjogren et al,[25] 2010 | + | + | + |
| Leifert et al,[26] 2009 | + | + | + |
| Veenema et al,[27] 2009 | + | + | + |
| Watanabe-Kanno et al,[28] 2009 | + | + | + |
| Goonewardene et al,[29] 2008 | + | + | + |
| Hildebrand et al,[30] 2008 | + | + | + |
| Keating et al,[31] 2008 | + | + | + |
| Redlich et al,[32] 2008 | + | + | + |
| Cha et al,[33] 2007 | + | + | + |
| Mullen et al,[34] 2007 | + | + | + |
| Stevens et al,[35] 2006 | + | + | + |
| Mayers et al,[36] 2005 | + | + | + |
| Costalos et al,[37] 2004 | + | + | + |
| Okunami et al,[30] 2004 | + | + | + |
| Quimby et al,[38] 2004 | + | + | + |
| Bell et al,[39] 2003 | + | + | + |
| Santoro et al,[40] 2003 | + | + | + |
| Tomassetti et al,[41] 2001 | + | + | + |

## Figures

Fig 1. QUADAS-2 charts.

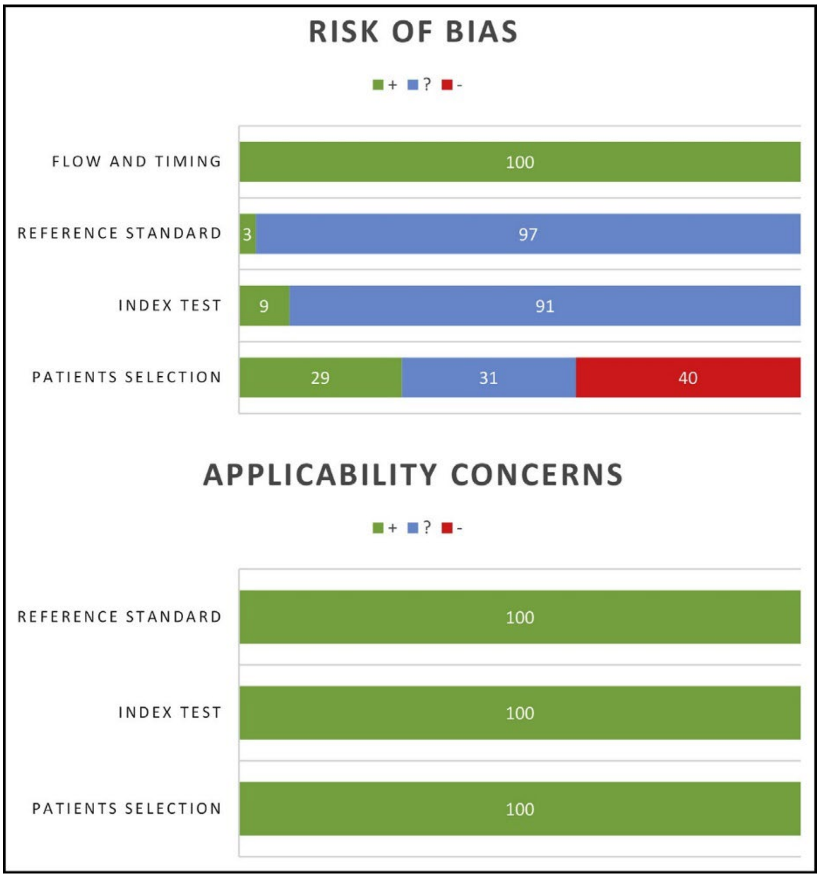

Fig 2, PRISMA flow chart