Accuracy and reliability of automatic three-dimensional cephalometric landmarking

Dot, Gautier[1]*; Rafflenbeul, Frédéric*[2]*; Arbotto, M.*[1]*; Gajny, Laurent*[1]*; Rouch, Philippe*[1]*; Schouman, Thomas*[1]

1 Institut de Biomecanique Humaine Georges Charpak [IBHGC]
2 Université de Strasbourg [UNISTRA]

ABSTRACT

The aim of this systematic review was to assess the accuracy and reliability of automatic landmarking for cephalometric analysis of 3D craniofacial images. We searched for studies that reported results of automatic landmarking and/or measurements of human head CT or CBCT scans in MEDLINE, EMBASE and Web of Science until march 2019. Two authors independently screened articles for eligibility. Risk of bias and applicability concerns for each included study were assessed using the QUADAS-2 tool. Eleven studies with test dataset sample sizes ranging from 18 to 77 images were included. They used knowledge-, atlas- or learning-based algorithms to landmark 2 to 33 points of cephalometric interest. Ten studies measured mean localization errors between manually- and automatically-detected landmarks. Depending on the studies and the landmarks, mean errors ranged from <0.50 mm to >5 mm. The two best-performing algorithms used a deep learning method and reported mean errors <2 mm for every landmark, approximating results of operator variability in manual landmarking. Risk of bias regarding patient selection and implementation of the reference standard were found, therefore the studies might have yielded overoptimistic results. The robustness of these algorithms needs to be more thoroughly tested in challenging clinical settings. PROSPERO registration number: CRD42019119637.

INTRODUCTION

Cephalometric analysis (or cephalometry) is a standardized diagnostic and treatment evaluation method used daily by orthodontists and maxillofacial surgeons. The analysis is based on linear and angular measurements performed on radiographic images. The gold standard for this procedure is a manual detection and landmarking of meaningful anatomical structures on lateral or frontal skull radiographs called cephalograms[1]. This X-ray technique is a two-dimensional (2D) projection of three-dimensional (3D) craniofacial structures, which leads to superimposition of bilateral structures and distortion of images, with enlargement in some areas and reduction in others[2].

To overcome the downsides of cephalograms, several authors have proposed 3D cephalometric analysis, based on 3D craniofacial images provided by computed tomography (CT) or cone beam computed tomography (CBCT) imaging techniques[3,4]. For now, there is no globally recognized 3D analysis or validated list of landmarks. Most of the proposed analyses have been 3D adaptations of previous 2D techniques, relying on landmarks localized on the bone surface of the skull[3,4]. These landmarks are then used to provide cephalometric results in the form of linear (Euclidian distance between two points), angular (angle between three points or two planes) and ratio (between two linear values) measurements. An example of a set of 3D landmarks localized on a skull model is shown in Fig. 1. It is suggested that 3D cephalometry could improve treatment outcomes for difficult cases (e.g. patients with craniofacial syndromes, major asymmetries/craniofacial anomalies or undergoing orthognathic surgery) when compared to traditional 2D cephalometry[5–7]. *In vitro* 3D craniofacial measurements are proven to be highly reliable, validating the possible use of CT or CBCT scans for 3D cephalometry[8].

Manual landmarking of 3D volumes requires time and a high level of expertise and experience[9]. Hassan et al. reported durations up to 14 minutes to place 22 landmarks[10]. Thorough training of the operators aims at reducing their identification errors in order to keep interobserver variation at a clinically acceptable level[9]. Reproducibility studies have shown that some landmarks are more reliable than others, with midsagittal plane landmarks usually showing greater reliability than bilateral landmarks[11]. Depending on the points and the studies, inter-operator variability ranges from less than 0.5 mm to more than 2 mm[5,8,11]. As a result, for the time being, this 3D technique is barely used in clinical settings and there is a lack of evidence as to which patients would benefit from it.

Automatization of the 3D cephalometric landmarking process could greatly facilitate access to this diagnostic tool. It would save time and enable untrained clinicians to use 3D cephalometry on a daily basis. Automatic 3D cephalometry could be more accurate than manual landmarking by learning to average out landmarking errors[12]. Various numerical methods have been proposed, including knowledge-based, atlas-based and learning-based algorithms[13,14]. The studies rely on different reporting methods, making it difficult to compare them. To our knowledge, neither a review of this research, nor an analytic comparison between results of different automatic 3D landmarking methods have been reported.

The aim of this systematic review is to assess the current evidence on the accuracy and reliability of automatic landmarking in comparison to manual landmarking for cephalometric analysis of 3D craniofacial images (CT or CBCT scans).

To this aim, our systematic review details the various techniques used and answers the following research questions:

1) What is the accuracy of automatic 3D landmarking when compared to manual landmarking?
2) How reliable are linear and angular measurements obtained through automatic landmarking when compared to manual landmarking?

MATERIALS AND METHODS

Protocol and registration

This systematic review is reported based on the PRISMA extension for Diagnostic Test Accuracy (DTA) guidelines. In accordance with the guidelines, our protocol was registered with the International Prospective Register of Systematic Reviews (PROSPERO) on the 28th of January 2019 (registration number CRD42019119637).

Eligibility criteria

Studies were selected according to the criteria outlined below:

- Study designs: we included *in vitro* and *in vivo* prospective and retrospective studies (clinical trials, comparative studies, validation studies or evaluation studies). We excluded book chapters, animal studies, case reports, epidemiologic studies, narrative reviews and author opinion articles.
- Population: we included studies examining the general human population, with no age limit.
- Index test: the index test of interest was automatic landmarking and/or measurements of 3D cranio-facial CT or CBCT scans. Several skeletal or dental landmarks with cephalometric interest needed to be localized in the maxillofacial area. "Automatic" meant that the landmarking or the measurements were performed by an algorithm, with minimal intervention by the operator (e.g. reorientation of the volume or manual localization of a few landmarks to run the

procedure). Detailed definitions of landmarks and/or measurements needed to be provided, as well as detailed definition of the algorithm used.

- Sample: for the index test, a sample size of at least 10 images needed to be provided.
- Reference standard: manual landmarking of 3D craniofacial CT or CBCT scans.
- Timing: there was no restriction in the search period.
- Language: we included articles reported in English, French and German.

Information sources

Our search was performed in the following databases: MEDLINE via Pubmed, EMBASE, Web of Science and Cochrane Central Register of Controlled Clinical Trials (CENTRAL). We searched the grey literature through OpenGrey database and Google Scholar, for which we considered the first 300 results for inclusion. To ensure literature saturation, we scanned the reference lists of included studies or relevant reviews identified through the search, and handsearched for studies citing included studies. No limit has been applied as to the date of publication, and our last search was performed on the 14th of March 2019.

Search strategy

The publications were searched electronically by one author, using controlled index terms and relevant specific free text words. After the MEDLINE strategy was finalized, it was adapted to the syntax and subject headings of the other databases (see Supplementary Table S1 for detailed search strategy). Duplicate articles were removed after importing the lists into a reference management software (Zotero v.5.0.62).

Study selection

Two reviewers independently screened the resulting collection of titles and abstracts. Studies that did not pertain to the review topic were excluded. When a title or abstract was considered to be relevant by only one of the reviewers, the publication was not excluded. The full texts of the remaining publications were then retrieved and reviewers independently assessed them to decide whether these met the inclusion criteria or not. Additional information was sought from study authors where necessary to resolve questions about eligibility. Disagreements between reviewers were solved through discussion and reasons for exclusion were recorded. When the same research team had published several articles on the refinement of an algorithm, the most recent paper was included. Neither of the reviewers were blinded to the journal titles, study authors or institutions.

Data collection process

One author extracted data into a standardized form subsequently checked by a second author. Disagreement was resolved through discussion. Additional information was sought from study authors where necessary, but articles were excluded whenever there were three or more unanswered requests.

Risk of bias and applicability

To evaluate the risk of bias and applicability of each study, information was collected using a tailored checklist based on the QUADAS-2 tool[15] and recommendations from the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy[16] (Supplementary Table S2). If there was insufficient detail reported in the study, the risk of bias was judged as "unclear". These judgements were made independently by two review authors. Disagreements were resolved first by discussion and then by consulting a third author for arbitration.

Diagnostic accuracy measures

The diagnostic accuracy measures reported were the mean differences and standard deviations expressed in mm (Euclidean distances), in degrees (angles) or in ratios (proportional measurements) between the automatic and the manual methods.

Synthesis of results

A systematic narrative and qualitative synthesis was provided with information presented in the text and tables to summarize and explain the characteristics and findings of the included studies. The narrative synthesis explored the relationship and findings both within and between the included studies. A meta-analysis was not possible because of the heterogeneity of the methodologies used in the selected studies.

Data availability

All the data generated or analysed during this study is included in this published article (and its Supplementary Data file).

RESULTS

Study selection

The flow chart of the selection process for inclusion of articles in this study is outlined in Fig. 2. A total of 654 manuscripts were selected for the screening phase (see Supplementary Table S1 for detailed results for each database) and 599 studies were excluded following abstract/title assessment. Following full-text review, a further 44 papers were excluded and the reasons for it were recorded, leaving 11 studies as eligible for inclusion in the qualitative synthesis. Among them, 10 studies[13,14,17–24] were related to our research question 1, and 1 study[25] was related to our research question 2. No studies were included in a quantitative synthesis.

Study characteristics

All of the selected studies for the qualitative analysis were published in English between 2014 and 2019 and were based on a retrospective selection of *in vivo* CBCT or CT scans. The sample size of the training dataset ranged from 24 to 201 images, and from 18 to 77 images for the test dataset.

None of the articles provided detailed descriptions of sample characteristics (e.g. gender, age, inclusion criteria, exclusion criteria, main craniofacial characteristics) nor provided calculation of the sample size. Four studies reported the use of a random method for population selection, but none reported the details of the randomization. A various number of skeletal and dental landmarks were localized, ranging from 2 to 33 points of cephalometric interest.

Three main types of algorithms were used for the automatic 3D landmarking: knowledge-based, atlas-based and learning-based methods. A synthesis of the principles, advantages and limitations of these algorithms is provided in Table 1. The computational power needed and running time of the algorithms was stated in only 2 studies[21,24]. For all studies, the reference standard used was manual landmarking. The reference landmarks were usually obtained by calculating the mean of landmarks provided by several observers. Summaries of the descriptive characteristics of the included articles are provided in Table 2 for research question 1 and Table 3 for research question 2.

Risk of bias and applicability

Using a tailored QUADAS-2 tool, three studies were assessed as being at overall low risk of bias[13,23,25] and two were at low concern regarding applicability[14,23].

Regarding patient selection, 7 studies showed an unclear risk of bias[17–22,24] and 8 showed applicability concerns[13,17,19–22,24,25]. This was mainly due to a lack of testing on random or consecutive patients and a lack of description of the population sample. Furthermore, 5 studies were at unclear or high risk of bias regarding the implementation of the reference standard[14,18,20,21,24]. They lacked a proper reference standard, or failed to report inter and intra-operator reproducibility results. Risk of bias and applicability assessment is summarized in Fig. 3.

Results of individual studies: research question 1

For research question 1, the results were separated according to the method used for the automatization algorithm.

Knowledge-based methods

Knowledge-based methods use mathematical descriptions (e.g. peak, lowest point…) to localize the landmarks on the anatomical contours of the images. Two studies used this method. Detailed description of the algorithm and mathematical entities of the points were provided.

A first study by Gupta et al.[13] tested 20 landmarks adapted from 2D cephalometry on a dataset of 30 CBCT scans. The initialization of the algorithm was based on the automatic search of a “seed point” using a template matching method on a segmented part of the images. A volume of interest (VOI) was defined around a point detected through distance vector from the “seed point”. Then, landmarks were detected on the contours identified on the anatomical structures of VOI. The overall mean error was 2.01 mm (standard deviation 1.23 mm).

Neelapu et al.[22] aimed at improving the results and robustness of the aforementioned method. After segmentation of the images, algorithm initialization was based on the automatic localization of the mid-sagittal plane using symmetry features of the skull. The image was then cropped into four quadrants and landmarks were detected on the anatomical contours. The algorithm showed slightly better results than the previous study, and was said to be more robust for deformed cases. The overall mean error was 1.88 mm (standard deviation 1.10 mm) for the 20 landmarks.

Atlas-based methods

Atlas-based methods use an atlas of one or more reference images, with landmarks manually placed by experts. In order to localize landmarks on a new image, one of these reference images is automatically registered (fitted) on the test image and the landmarks are transferred. Two studies used this method.

Shahidi et al.[17] used 8 manually-annotated CBCT scans to generate the head atlas. The algorithm was tested with 14 landmarks on 20 CBCT scans. Depending on the age of the subject, one image of the atlas was automatically selected and fitted on the test image. The algorithm used feature and voxel similarity-based registration before scaling, rotation, and translation of the test image. The overall mean error was 3.40 mm.

Codari et al.[19] tested the automatic localization of 21 landmarks on 18 CBCT scans of healthy adult Caucasian women. One manually-annotated CBCT scan was used for the head atlas. After automatic segmentation of the test image using k-means clustering, the atlas image was automatically fitted on the test image, using first an affine (linear) intensity-based image registration technique and then an elastic (nonlinear) one. The overall mean error was 2.39 mm (standard deviation 1.73 mm).

Learning-based methods

Learning-based methods include various methods which rely on a training sample of images. Two sub-types can be described: statistical and machine learning methods. Statistical methods (active shape model and Elastic Bunch Graph Matching) correlate a shape with deformation modes, or a graph representation, extracted from the training images, with the test image. Machine learning methods (random forest and deep learning) use the training data in order to learn where to localize the landmarks without being explicitly programmed to perform this task.

Montúfar et al.[21] used a combination of learning-based and knowledge-based methods. First, 2 active shape models (ASM) were trained on digitally reconstructed 2D radiographs for a holistic automatic 2D landmark approximation. Then, the 3D coordinates of the points were computerized and segmentation of the images' sub-volumes was performed around the points. Finally, a knowledge-based method was used to localize the landmarks precisely on the anatomical contours. The ASM was trained on 24 CBCT scans, and the overall localization results were tested on the same set of images (leave-one-out test). The mean error was 2.51 mm (standard deviation 1.6 mm). In terms of processing time, Montúfar et al.'s[21] method was compared to Gupta et al.[13] and Codari et al.[19]. Reported processing times were 49, 126.25 and 2,892.2 seconds, respectively.

De Jong et al.[20] used an Elastic Bunch Graph Matching-based (EBGM) method. The training dataset consisted of 39 CBCT scans which were manually segmented and landmarked once by one operator. A total of 33 landmarks were localized. The segmented skulls were projected on a 2D plane and a large set of features was derived from this data. EBGM method was used to search for a maximum correlation between the test image and a graph representation extracted from the training images. A leave-one-out test was used to evaluate the algorithm, and 10 landmarks had a mean error inferior to 2 mm.

Zhang et al.[18] used a random forest method to automatically localize 15 landmarks on 41 CBCT scans in a 5-fold cross validation test. The method was based on a regression voting strategy, using image segmentation to remove uninformative voxels. Then, a partially-joint model was used to localize landmarks separately based on the coherence of their positions. The training dataset consisted of 41 CBCT and 30 CT scans, which were labelled once by one experienced operator. The overall mean error of the automatic localization was 1.44 mm, with all the landmarks having a mean error inferior to 2 mm.

Three studies used a deep learning method. O'Neil et al.[24] used a shallow fully convolutional neural network (FCN) and *atlas location* autocontext in order to localize 22 landmarks in the head. *Atlas location* autocontext was described in this work as "iteratively feeding the coordinate in atlas space, according to the output of a model, to a subsequent model". Two of these 22 landmarks had a cephalometric interest. A total of 170 CT scans were used for training, 31 for validation and 20 for testing. These images contained "pathology, inclusive of haemorrhage, tumours and age-related change". The data was manually labelled once by two observers, but the mean localization of the manual points was not used. The overall mean error of the automatic localization for the two points was 2.45 mm (standard deviation 2.53 mm) for observer A and 3.49 mm (standard deviation 2.88 mm) for observer B.

Zhang et al.[14], in a second study, automatically localized 15 landmarks on 77 CBCT scans in a 5-fold cross validation test. Two fully convolutional neural networks (FCN-1 and FCN-2) with a U-Net architecture were used. FCN-1 was used to learn the *displacement maps* for multiple landmarks in order to model the spatial context information in the whole image. Then, FCN-2 performed bone segmentation and landmark localization using both FCN-1 results and the original image as input. The training dataset consisted of 77 CBCT and 30 CT scans. The overall mean error of the automatic localization was 1.10 mm (standard deviation 0.71 mm).

Torosdagli et al.[23] used an adapted fully convolutional DenseNET network (also called Tiramisu network) for image segmentation, followed by an improved Zhang et al.[14] U-Net network to localize sparsely-spaced landmarks. Then, a long short-term memory (LSTM) network was used to localize mid-sagittal closely-spaced landmarks near the "Menton" point. The training dataset consisted of 50 CBCT scans of mandibles including subjects with “congenital deformities fading to extreme developmental variations” and artefacts. The algorithm was tested on the same dataset using a 5-fold cross validation test. Eight out of the 9 mandibular landmarks tested were localized with a mean error inferior to 0.5 mm. The 9th one, “Pogonion”, had a mean error of 1.55 mm.

We collected the detailed results of the mean errors and standard deviations for each cephalometric landmark. These mean errors were computed as mean distances (in mm) between the automatically-detected test landmarks and the manually-detected reference landmarks (the latter being the mean of several observers, except for Zhang et al.[14,18], de Jong et al.[20] and O’Neil et al.[24]). Table 4 provides detailed results for the 19 most reported landmarks. The entire list of results can be found as Supplementary Table S3.

Results of individual studies: research question 2

For research question 2, the only study was performed by Gupta et al.[25] following the same knowledge-based method as their other study. Linear, angular and ratio measurements were computerized using the manually-placed or automatically-placed landmarks. Then, the difference between the measurements was calculated as mean error. The unpaired t-test (95% level of significance) showed no statistically significant differences. For the linear measurements (Euclidian distance between two points), the highest error was 2.63 mm (mean standard deviation between 0.35 and 2.46 mm). For the angular measurements (angle between three points or two planes), the highest error was 2.12° (mean standard deviation between 0.46 and 2.40°). For the ratios (proportional measurements between two linear measurements), the highest error was 0.03 (mean standard deviation between 0.01 and 0.03).

DISCUSSION

Our systematic review revealed that automatic landmarking of 3D craniofacial images is an active and current research field, as 5 out of 11 of our included studies were published in 2018 or 2019. Only one among the selected studies answered our research question 2 about the reliability of linear and angular 3D measurements obtained through automatic landmarking. This is quite surprising considering that diagnostic value of cephalometric analysis rests on linear and angular measurements, not merely on landmarks. Although these measurements are based on landmarks, overall measurement errors cannot be deduced systematically from landmark localization errors. Depending on landmark coordinate values, the overall measurement error can be reduced or increased, thus modifying the clinical significance of the results[8,11,25]. Therefore, there is a lack of evidence about the diagnostic accuracy of automatic 3D cephalometry[26].

Concerning our research question 1, the best localization results were obtained by two studies that used a deep learning method to automatically localize the landmarks[14,23]. More specifically, these two studies used fully convolutional neural network with a U-Net architecture. Similarly, two of the best performing algorithms for automatic 2D cephalometry used a machine learning-based algorithm[12,27].

These results need to be compared to those obtained through manual landmarking. Reproducibility studies of manual landmarking report variable results depending on the landmarks. Intra-operator results usually show mean differences smaller than 1 mm, and inter-operator variability ranges from less than 0.5 mm to more than 2 mm[5,8,11]. At the moment, there is no clear threshold for clinical significance of inter-observer variability. Depending on the authors, the limit could be 0.5 mm, 1 mm, 2 mm or even more[9,11,25]. This questions the use of manual landmarking as the reference standard to test automated landmarking, but for now there is no other choice than to consider landmark localization by the mean of experts as the gold standard[12,13]. A way to reduce uncertainty with this reference standard is to use the mean of manual landmarks obtained by several independent observers at different times.

When compared to the aforementioned body of literature, the localization results of the automated methods are very promising. Nonetheless, most of the algorithms were tested on a small set of cephalometric points or localized unconventional landmarks, as showed in Table 4 and Supplementary Table S3. This jeopardize the clinical application of most of these methods, which cannot be used to perform a complete 3D cephalometric analysis. In the detailed point-by-point results of the two best performing studies, some points show larger standard deviations than others. It is particularly noticeable in the results of Zhang et al.[14] for points “Gonion Left” and “Gonion Right”, and in the results of Torosdagli et al.[23] for points “Pogonion” and “Gnathion”. It is difficult to know what explains this phenomenon without detailed directional results for the errors. These landmarks are localized on curved structures with no clear boundaries, which are also known to be difficult to localize precisely in manual landmarking[11].

The performance of the learning-based algorithms entirely depends on the quality, size and variability of their training datasets[12]. The robustness of these algorithms needs to be more thoroughly tested in challenging and actual clinical sets, and time cost of the methods should be considered. These tests should primarily focus on the main target population of 3D cephalometry, difficult cases (e.g. patients with craniofacial syndromes, major asymmetries/craniofacial anomalies or undergoing orthognathic surgery)[5,6]. As it has been done for automated 2D cephalometry, it would be interesting to gather a public and unbiased labelled set of images for the benchmarking of the algorithms[28]. It would allow the training and testing of the algorithms with a consistent evaluation method, thus helping the direct comparison between the results. In order to minimise the radiation dose of the patients, the algorithms should also be trained and tested on images acquired through low-dose protocols[8,29].

Several studies showed risk of bias or applicability concerns regarding patient selection and implementation of the reference standard, mainly because the risks were assessed as unclear. Insufficient data has been reported in the included studies, therefore it cannot be ruled out that patients might have inappropriately been excluded or that the manual landmarking step might have failed to correctly detect the reference landmarks[15]. Overall, some of the included studies might have yielded overoptimistic results. Interestingly, the study that provided the best results was also the only one that was assessed as being at overall low risk of bias and low concern regarding applicability[23]. However, it only focused on a set of mandibular landmarks and was validated on a rather small dataset.

The studies could have reported their results in other forms. Only three studies reported the percentage of points successfully located within a radius of 1 mm, 2 mm and 3 mm from the reference point. This data is needed to compute successful detection rates of the algorithms[28]. Moreover, mean error might not be the most relevant result to assess distribution when error distributions are asymmetrical, as it is frequently the case with the algorithms used in the included studies. Median error and interquartile range should be used in that case[19]. Finally, the error results were given as Euclidian distances in all included studies, without referring to the x- y- z-axis. Detailed directional results are necessary to identify the points that are prone to error in one plane more than the others and thus are of different clinical significance[1,9].

Finally, the reliability of landmarking does not necessarily translate into meaningful implications and clinically relevant results[11]. The same limitation applies for now to manual 3D cephalometry[5,8]. More studies on diagnostic thinking efficacy and therapeutic efficacy[26] of automatic 3D cephalometry are needed in order to know in which cases this technique is useful for diagnosis and treatment planning[29].

DECLARATIONS

Funding: This work was funded in part by the "Association Française pour le Développement de la Stomatologie". The sponsor had no involvement in any aspect of the study design nor in the manuscript preparation and submission.

Competing Interests: The authors declare no competing interests. All authors contributed to revising the manuscript critically and approved it for submission.

Ethical Approval: Not applicable.

Patient Consent: Not required.

APPENDIX A. SUPPLEMENTARY DATA

Supplementary data associated with this article (Supplementary Table S1 to S3) can be found in the supplementary data file.

REFERENCES

1. Leonardi R., Giordano D., Maiorana F., Spampinato C. Automatic Cephalometric Analysis: A Systematic Review. *Angle Orthod* 2008;**78**(1):145–51. Doi: 10.2319/120506-491.1.
2. Gribel BF., Gribel MN., Frazão DC., McNamara JA., Manzi FR. Accuracy and reliability of craniometric measurements on lateral cephalometry and 3D measurements on CBCT scans. *Angle Orthod* 2011;**81**(1):26–35. Doi: 10.2319/032210-166.1.
3. Olszewski R., Cosnard G., Macq B., Mahy P., Reychler H. 3D CT-based cephalometric analysis: 3D cephalometric theoretical concept and software. *Neuroradiology* 2006;**48**(11):853–62. Doi: 10.1007/s00234-006-0140-x.
4. Lee S-H., Kil T-J., Park K-R., Kim BC., Kim J-G., Piao Z., et al. Three-dimensional architectural and structural analysis--a transition in concept and design from Delaire's cephalometric analysis. *Int J Oral Maxillofac Surg* 2014;**43**(9):1154–60. Doi: 10.1016/j.ijom.2014.03.012.
5. Pittayapat P., Limchaichana-Bolstad N., Willems G., Jacobs R. Three-dimensional cephalometric analysis in orthodontics: a systematic review. *Orthod Craniofac Res* 2014;**17**(2):69–91. Doi: 10.1111/ocr.12034.
6. Kapila SD., Nervina JM. CBCT in orthodontics: assessment of treatment outcomes and indications for its use. *Dentomaxillofacial Radiol* 2015;**44**(1):20140282. Doi: 10.1259/dmfr.20140282.
7. Swennen GRJ., Schutyser FAC., Hausamen J-E. *Three-Dimensional Cephalometry: A Color Atlas and Manual*. Berlin Heidelberg: Springer-Verlag; 2006.
8. Smektała T., Jędrzejewski M., Szyndel J., Sporniak-Tutak K., Olszewski R. Experimental and clinical assessment of three-dimensional cephalometry: A systematic review. *J Cranio-Maxillofac Surg* 2014;**42**(8):1795–801. Doi: 10.1016/j.jcms.2014.06.017.
9. Lagravère MO., Low C., Flores-Mir C., Chung R., Carey JP., Heo G., et al. Intraexaminer and interexaminer reliabilities of landmark identification on digitized lateral cephalograms and formatted 3-dimensional cone-beam computerized tomography images. *Am J Orthod Dentofacial Orthop* 2010;**137**(5):598–604. Doi: 10.1016/j.ajodo.2008.07.018.
10. Hassan B., Nijkamp P., Verheij H., Tairie J., Vink C., van der Stelt P., et al. Precision of identifying cephalometric landmarks with cone beam computed tomography in vivo. *Eur J Orthod* 2013;**35**(1):38–44. Doi: 10.1093/ejo/cjr050.
11. Sam A., Currie K., Oh H., Flores-Mir C., Lagravére-Vich M. Reliability of different three-dimensional cephalometric landmarks in cone-beam computed tomography: A systematic review. *Angle Orthod* 2018;**89**(2):317–32. Doi: 10.2319/042018-302.1.
12. Lindner C., Wang C-W., Huang C-T., Li C-H., Chang S-W., Cootes TF. Fully automatic system for accurate localisation and analysis of cephalometric landmarks in lateral cephalograms. *Sci Rep* 2016;**6**:33581.
13. Gupta A., Kharbanda OP., Sardana V., Balachandran R., Sardana HK. A knowledge-based algorithm for automatic detection of cephalometric landmarks on CBCT images. *Int J Comput Assist Radiol Surg* 2015;**10**(11):1737–52. Doi: 10.1007/s11548-015-1173-6.
14. Zhang J., Liu M., Wang L., Chen S., Yuan P., Li J., et al. Joint Craniomaxillofacial Bone Segmentation and Landmark Digitization by Context-Guided

Fully Convolutional Networks. *Med Image Comput Comput-Assist Interv MICCAI Int Conf Med Image Comput Comput-Assist Interv* 2017;**10434**:720–8. Doi: 10.1007/978-3-319-66185-8_81.

15. Whiting PF. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. *Ann Intern Med* 2011;**155**(8):529. Doi: 10.7326/0003-4819-155-8-201110180-00009.

16. Reitsma JB., Rutjes AWS., Whiting P., Vlassov VV., Leeflang MMG., Deeks JJ. Assessing methodological quality. *Cochrane handbook for systematic reviews of diagnostic test accuracy version*, vol. 1. 2009.

17. Shahidi S., Bahrampour E., Soltanimehr E., Zamani A., Oshagh M., Moattari M., et al. The accuracy of a designed software for automated localization of craniofacial landmarks on CBCT images. *BMC Med Imaging* 2014;**14**:32. Doi: 10.1186/1471-2342-14-32.

18. Zhang J., Gao Y., Wang L., Tang Z., Xia JJ., Shen D. Automatic Craniomaxillofacial Landmark Digitization via Segmentation-Guided Partially-Joint Regression Forest Model and Multiscale Statistical Features. *IEEE Trans Biomed Eng* 2016;**63**(9):1820–9. Doi: 10.1109/TBME.2015.2503421.

19. Codari M., Caffini M., Tartaglia GM., Sforza C., Baselli G. Computer-aided cephalometric landmark annotation for CBCT data. *Int J Comput Assist Radiol Surg* 2017;**12**(1):113–21. Doi: 10.1007/s11548-016-1453-9.

20. de Jong MA., Gül A., de Gijt JP., Koudstaal MJ., Kayser M., Wolvius EB., et al. Automated human skull landmarking with 2D Gabor wavelets. *Phys Med Biol* 2018;**63**(10):105011.

21. Montúfar J., Romero M., Scougall-Vilchis RJ. Hybrid approach for automatic cephalometric landmark annotation on cone-beam computed tomography volumes. *Am J Orthod Dentofac Orthop Off Publ Am Assoc Orthod Its Const Soc Am Board Orthod* 2018;**154**(1):140–50. Doi: 10.1016/j.ajodo.2017.08.028.

22. Neelapu BC., Kharbanda OP., Sardana V., Gupta A., Vasamsetti S., Balachandran R., et al. Automatic localization of three-dimensional cephalometric landmarks on CBCT images by extracting symmetry features of the skull. *Dento Maxillo Facial Radiol* 2018;**47**(2):20170054. Doi: 10.1259/dmfr.20170054.

23. Torosdagli N., Liberton DK., Verma P., Sincan M., Lee JS., Bagci U. Deep Geodesic Learning for Segmentation and Anatomical Landmarking. *IEEE Trans Med Imaging* 2019;**38**(4):919–31. Doi: 10.1109/TMI.2018.2875814.

24. O'Neil AQ., Kascenas A., Henry J., Wyeth D., Shepherd M., Beveridge E., et al. Attaining Human-Level Performance with Atlas Location Autocontext for Anatomical Landmark Detection in 3D CT Data. In: Leal-Taixé L, and Roth S, editors. *Computer Vision – ECCV 2018 Workshops*. Springer International Publishing; 2019. p. 470–84.

25. Gupta A., Kharbanda OP., Sardana V., Balachandran R., Sardana HK. Accuracy of 3D cephalometric measurements based on an automatic knowledge-based landmark detection algorithm. *Int J Comput Assist Radiol Surg* 2016;**11**(7):1297–309. Doi: 10.1007/s11548-015-1334-7.

26. Fryback DG., Thornbury JR. The efficacy of diagnostic imaging. *Med Decis Mak Int J Soc Med Decis Mak* 1991;**11**(2):88–94. Doi: 10.1177/0272989X9101100203.

27. Vandaele R., Aceto J., Muller M., Péronnet F., Debat V., Wang C-W., et al. Landmark detection in 2D bioimages for geometric morphometrics: a multi-resolution tree-based approach. *Sci Rep* 2018;**8**(1). Doi: 10.1038/s41598-017-18993-5.

28. Wang C-W., Huang C-T., Lee J-H., Li C-H., Chang S-W., Siao M-J., et al. A

benchmark for comparison of dental radiography analysis algorithms. *Med Image Anal* 2016;**31**:63–76. Doi: 10.1016/j.media.2016.02.004.

29. SEDENTEXCT project. *Cone Beam CT for dental and maxillofacial radiology (evidence based guidelines)*. European Commission; 2012.

Table 1. Principles, advantages and limitations of the algorithms used in the included articles

General Method	Specific Method	Principle	Advantages	Limitations
Knowledge-based		1. Mathematical entities are associated with the landmark locations (e.g. peak, lowest point…) 2. The landmarks are automatically localized on each contour of the test image based on the definitions	- Applies the concept of manual plotting based on pre-agreed definitions	- Detection of contours is the vulnerable step - Landmarks placed on curved structures are hard to localize - Robustness can be challenged with severely deformed cases
Atlas-based		1. A reference image atlas is created, with landmarks placed manually by experts 2. One image of this atlas is automatically registered (fitted) on a test image 3. The landmarks are transferred on the test image	- Simple method with low amount of a priori information needed - Can be customized easily	- Atlases have to be accurate and match biological variations (for sex, age, ethnicity…) - Highly dependent on registration technique which can be computationally expensive - Robustness can be challenged with severely deformed cases
Learning-based	Active shape model (ASM)	1. The landmarks are placed manually by experts on the training sample images 2. A statistical model (mean shape) is created by scaling, rotating and translating the training shapes so that they correspond as closely as possible 3. The model is iteratively deformed to fit the test image and automatically localize the landmarks	- Well-described and thoroughly studied method - Low sensitivity to artifacts and noise in the image	- 2-dimensional technique - Needs large training sample size to match biological variations - Needs accurate training data - Robustness can be challenged with severely deformed cases
	Elastic Bunch Graph Matching (EBGM)	1. The landmarks are placed manually by experts on the training sample images 2. A large set of 2D features is derived from the training data, using image filtering 3. The landmarks are automatically detected on the test image based on a maximum correlation search between the test image and a graph representation extracted from the training images	- Does not need a large training sample	- 2-dimensional technique - Needs accurate training data - Sensitive to artifacts and noise in the image
	Random forest	1. The landmarks are placed manually by experts on the training sample images 2. Visual features are chosen and a multitude of decision trees is constructed from the training data 3. All these decision trees are automatically combined to vote for the most probable position of the landmarks on the test image	- Well-described and thoroughly studied method - Low sensitivity to artifacts and noise in the image	- Needs a large training sample size with artifacts and anatomical variations - Robustness can be challenged with severely deformed cases
	Deep learning	1. The landmarks are placed manually by experts on the training sample images 2. A deep neural network is trained with the sample data 3. The landmarks are automatically detected on the test image	- Can accommodate strong anatomical variations - Low sensitivity to artifacts and noise in the image - Highly dynamic research field	- Needs a very large training sample size with artifacts and anatomical variations - Needs accurate training data - Training phase is computationally expensive - Downsampling of images might be needed, which can increase uncertainty in results - Neural network parameters have to be determined empirically

Table 2. Summary characteristics of included articles – Research question 1

Article	Population and selection method	Acquisition voxel size	Number of landmarks tested	Index test – Automatic 3D landmarking			Reference standard – Manual landmarking		**Main Results** Total mean difference ± SD between index test and reference standard
				Algorithm used	Training dataset	Test dataset	Observers Repetitions	Intraobserver Interobserver results	
Shahidi et al. 2014[17]	Random retrospective selection from private practice images, without "significant fractures or severe skeletal anomalies" Age 10-43	*Unknown*	14	Atlas-based method	n/a	20 CBCT scans	3 observers 2 times	Intraobserver – ICC = 0.89 Interobserver - ICC + 95% CI = 0.87 [0.82-0.93]	3.40 mm
Gupta et al. 2015[13]	Random retrospective selection from postgraduate orthodontic clinic database "irrespective of age, gender and ethnicity"	Isometric 0.25-0.40 mm	20	Knowledge-based method	*n/a*	30 CBCT scans	3 observers 1 time	Interobserver – ICC > 0.9	2.01 ± 1.23 mm
Zhang et al. 2016[18]	Retrospective selection Non-syndromic dentofacial deformity Skeletal Class II and Class III patients	CBCT scans : isometric 0.4 mm CT scans: 0.488 × 0.488 × 1.25 mm^3	15	Random forest-based method	41 CBCT scans 30 CT scans	41 CBCT scans (same as training - 5-fold cross validation)	1 observer 1 time	*n/a*	1.44 mm
Codari et al. 2017[19]	Retrospective selection from private practice database "Adult healthy Caucasian women" Age 37-74	*Unknown*	21	Atlas-based method	n/a	18 CBCT scans	"Team of expert users" 1 time	Interobserver – ICC = 0.98	2.39 ± 1.73 mm
Zhang et al. 2017[14]	Retrospective selection from private practice database Non-syndromic dentofacial deformities Even distribution between skeletal classes	CBCT scans: isometric 0.3 or 0.4 mm CT scan: 0.488 × 0.488 × 1.25 mm^3	15	Deep learning-based method	77 CBCT scans 30 CT scans	77 CBCT scans (same as training - 5-fold cross validation)	2 observers (on different images) 1 time	*n/a*	1.10 ± 0.71 mm
de Jong et al. 2018[20]	Retrospective selection from orthodontic clinic database Non-syndromic cohort Age 16-54	Slice thicknesses between 0.3 and 1 mm	33	Elastic Bunch Graph Matching-based (EBGM) method	39 CBCT scans	39 CBCT scans (same as training - leave-one-out test)	1 observer 1 time	*n/a*	Mean error <2 mm for 10 landmarks
Montúfar et al. 2018[21]	Random selection from public repository (Virtual Skeleton Database from the Medical Image Repository of the Swiss Institute for Computer Assisted Surgery)	Isometric 0.4 mm	18	Active shape model (ASM) + Knowledge-based method on subvolumes	24 CBCT scans	24 CBCT scans (same as training - leave-one-out test)	2 observers 2 times	Intraobserver: "12 of 18 landmarks reproducible within a 1.0-mm standard deviation"	2.51 ± 1.6 mm
Neelapu et al. 2018[22]	Retrospective selection from postgraduate orthodontic clinic database "irrespective of age, gender and ethnicity"	Isometric 0.25-0.40 mm	20	Knowledge-based method	*n/a*	30 CBCT scans	3 observers 1 time	Interobserver – ICC > 0.9	1.88 ± 1.10 mm
Torosdagli et al. 2019[23]	Retrospective selection from hospital database, including "congenital deformities fading to extreme developmental variations in CMF bones" and artifacts	Isometric 0.29 or 0.377 mm (before resampling)	9[a]	Deep learning-based method	50 CBCT scans	50 CBCT scans (same as training - 5-fold cross validation)	3 observers 2 times for 2 observers 1 time for 1 observer	Interobserver – ICC = 0.92	Mean error ≤0.5 mm for 8 landmarks
O'Neil et al. 2019[24]	Retrospective selection from hospital database, containing "pathology, inclusive of haemorrhage, tumours and age-related change"	"Range of resolutions and slice thicknesses"	2[b]	Deep learning-based method	170 CT scans for training 31 CT scans for validation	20 CT scans	2 observers 1 time	Interobserver – mean=2.20mm / median=1.48mm	Observer A: 2.45 ± 2.53 mm Observer B: 3.49 ± 2.88 mm

CT, computed tomography; CBCT, cone-beam computed tomography; ICC, intraclass correlation coefficient

[a] Only mandibular landmarks [b] Only 2 out of the 22 studied landmarks had cephalometric interest

Table 3. Summary characteristics of included articles – Research question 2

Article	**Population and selection method**	**Acquisition voxel size**	**Number of measurements tested**	**Index test – Automatic 3D landmarking**			**Reference standard – Manual landmarking**		**Main Results**
				Algorithm used	Training dataset	Test dataset	Observers Repetitions	Intraobserver Interobserver results	**Deviations of measurements**
Gupta et al. 2016[25]	Random selection from postgraduate orthodontic clinic database "irrespective of age, gender and ethnicity"	Isometric 0.25-0.40 mm	28 linear 16 angular 7 ratios	Knowledge-based method	*n/a*	30 CBCT scans	3 observers 1 time	Interobserver – ICC > 0.9	- Linear measurements – highest error 2.63mm; mean standard deviation between 0.35 and 2.46 mm - Angular measurements – highest error 2.12°; mean standard deviation between 0.46 and 2.40° - Ratios – highest error 0.03; mean standard deviation between 0.01 and 0.03

CBCT, cone-beam computed tomography; ICC, intraclass correlation coefficient

Table 4. Mean localization errors ± standard deviations (in mm) for the 19 most reported landmarks

	Shahidi et al. 2014[17]	Gupta et al. 2015[13]	Zhang et al. 2016[18a]	Codari et al. 2017[19a]	Zhang et al. 2017[14a,b]	De Jong et al. 2018.[20]	Montúfar et al. 2018[21]	Neelapu et al. 2018[22]	Torosdagli et al. 2019[23a,c]	O'Neil et al. 2019[24a]
Anterior Nasal Spine (ANS)	3.12 ± 0.80	1.42 ± 0.73		2.58 ± 1.50		5.6 ± 8.1	1.72 ± 0.91	1.03 ± 0.62		Observer A: 2.57 ± 3.37 Observer B 2.84 ± 3.49
Condylar Left (CdL)		3.20 ± 2.49						3.78 ± 2.77	0.34 ± 0.60	
Condylar Right (CdR)		2.38 ± 1.71						3.34 ± 2.47	0.08 ± 0.24	
Frontozygomatic Left (FzL)		1.47 ± 0.86		2.84 ± 2.36		2.0 ± 1.2				
Frontozygomatic Right (FzR)		1.60 ± 0.71		2.54 ± 1.76		1.5 ± 1.1				
Gnathion (Gn)	3.77 ± 2.69	1.62 ± 0.62					2.10 ± 1.06	1.64 ± 0.68	0.49 ± 1.42	
Gonion Left (GoL)		2.04 ± 1.47	1.59 ± 0.88	3.92 ± 2.38	1.51 ± 1.00	2.6 ± 2.0	2.33 ± 1.62	2.02 ± 1.09		
Gonion Right (GoR)		2.47 ± 1.37	1.61 ± 1.11	3.20 ± 1.96	1.79 ± 0.65	4.8 ± 5.7	2.45 ± 1.76	2.10 ± 1.18		
Lateral Zygomatic Left (LatzL)		2.80 ± 1.63				2.1 ± 1.1		1.74 ± 1.01		
Lateral Zygomatic Right (LatzR)		2.83 ± 2.05				1.7 ± 1.0		1.48 ± 1.05		
Menton (Me)	3.59 ± 1.79	1.21 ± 0.58	1.02 ± 0.73	1.76 ± 0.83	0.81 ± 0.71		2.28 ± 1.15	1.57 ± 0.54	0.04 ± 0.12	
Nasion (N)	3.20 ± 1.64	1.17 ± 0.49	1.62 ± 0.82	3.19 ± 3.33	0.96 ± 0.69	3.0 ± 2.5	2.14 ± 1.04	0.95 ± 0.69		Observer A: 2.35 ± 1.48 Observer B: 4.04 ± 2.10
Orbitale Left (OrL)		1.78 ± 1.36	1.55 ± 0.70	1.74 ± 1.08	1.08 ± 0.53	1.9 ± 2.5	3.12 ± 2.70			
Orbitale Right (OrR)		2.37 ± 2.23	1.58 ± 0.85	1.69 ± 1.28	0.97 ± 0.56	3.7 ± 3.4	3.46 ± 2.13			
Pogonion (Pog)	3.00 ± 1.02	1.53 ± 0.79	1.03 ± 0.53	2.88 ± 1.52	0.93 ± 0.47	4.6 ± 8.4	2.59 ± 0.98	1.77 ± 0.96	1.55 ± 1.98	
Point A	3.11 ± 0.74	1.73 ± 0.80		1,80 ± 0,86			1.46 ± 0.75	1.91 ± 0.94		
Point B	3.86 ± 1.41	2.08 ± 1.09		2,66 ± 1,33			2.53 ± 0.56	1.78 ± 0.91	0.34 ± 0.72	
Posterior Nasal Spine (PNS)	3.60 ± 1.35	2.08 ± 1.29		1.64 ± 1.18			2.17 ± 1.27	1.60 ± 1.15		
Sella (S)	3.45 ± 1.82	1.52 ± 0.75		1.44 ± 0.73			2.67 ± 2.05	2.19 ± 0.91		

[a] Unpublished data shared by the authors [b] Results for "JSD" method [c] Results for "max pool without dropout" method

CAPTIONS TO ILLUSTRATIONS

Fig. 1. Example of 3D landmarks localized on a skull model, lateral right and frontal views (dotted points show approximate projections of intra-cranial landmarks)

Fig. 2. Flow chart of data searches using PRISMA guidelines

Fig. 3. Bias and applicability assessment of included studies using tailored QUADAS-2 tool

FIGURE 1

Example of 3D landmarks localized on a skull model, lateral right and frontal views

(dotted points show approximate projections of intra-cranial landmarks)

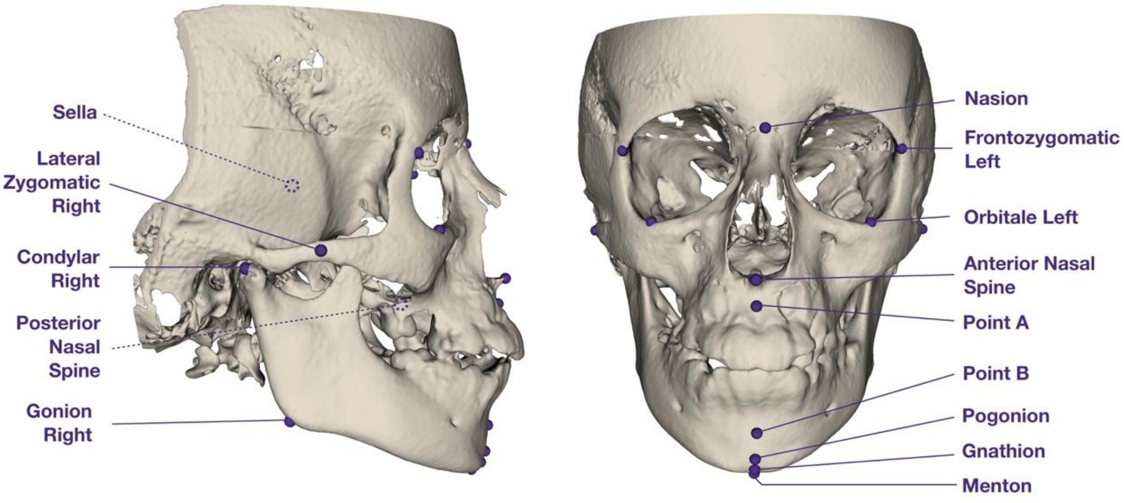
Sella
Lateral Zygomatic Right
Condylar Right
Posterior Nasal Spine
Gonion Right
Nasion
Frontozygomatic Left
Orbitale Left
Anterior Nasal Spine
Point A
Point B
Pogonion
Gnathion
Menton

FIGURE 2

Flow chart of data searches using PRISMA guidelines

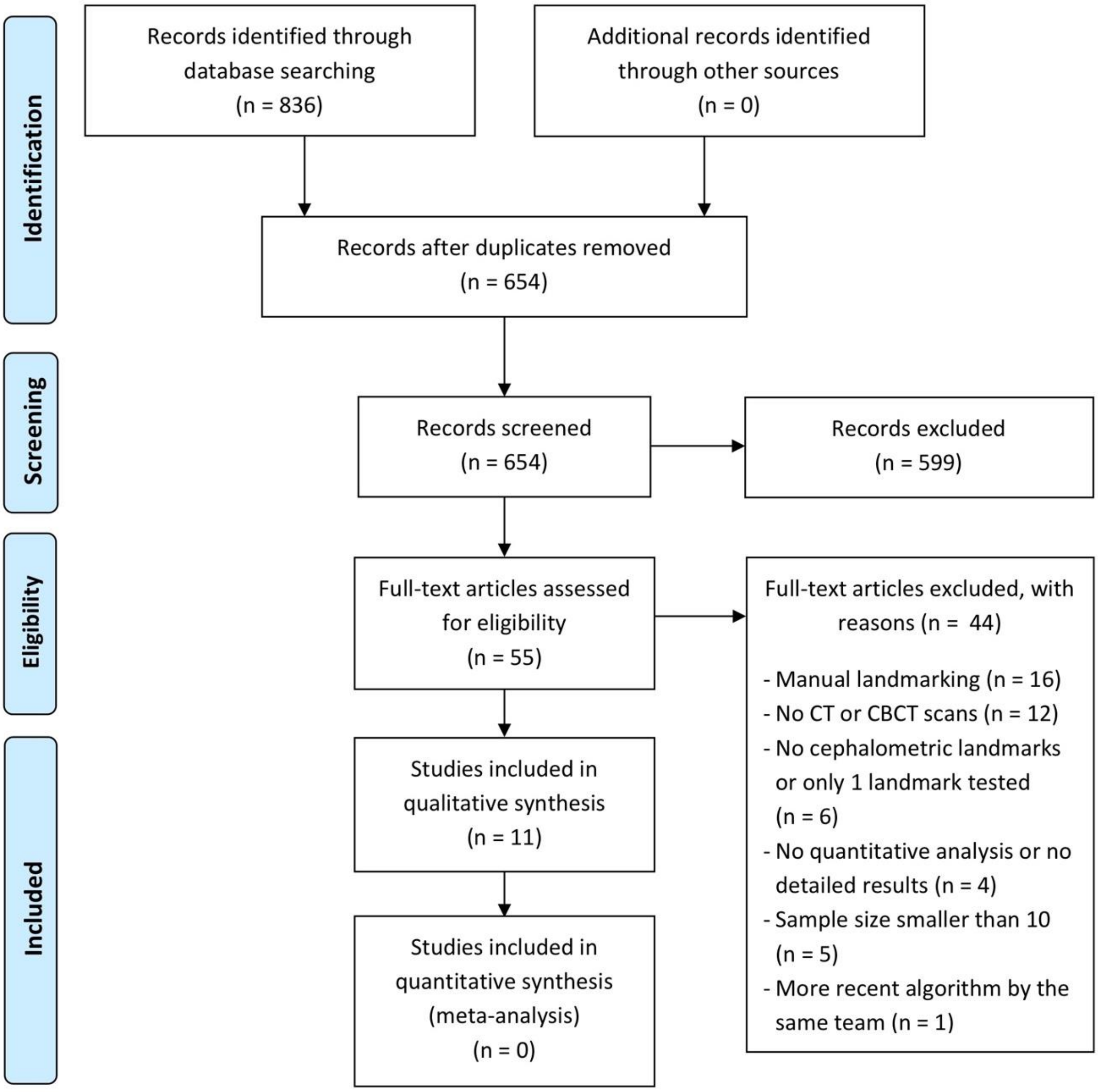
Identification
Screening
Eligibility
Included
Records identified through database searching (n = 836)
Additional records identified through other sources (n = 0)
Records after duplicates removed (n = 654)
Records screened (n = 654)
Records excluded (n = 599)
Full-text articles assessed for eligibility (n = 55)
Full-text articles excluded, with reasons (n = 44)
- Manual landmarking (n = 16)
- No CT or CBCT scans (n = 12)
- No cephalometric landmarks or only 1 landmark tested (n = 6)
- No quantitative analysis or no detailed results (n = 4)
- Sample size smaller than 10 (n = 5)
- More recent algorithm by the same team (n = 1)
Studies included in qualitative synthesis (n = 11)
Studies included in quantitative synthesis (meta-analysis) (n = 0)

FIGURE 3

Bias and applicability assessment of included studies using tailored QUADAS-2 tool

Study	RISK OF BIAS				APPLICABILITY CONCERNS		
	PATIENT SELECTION	INDEX TEST	REFERENCE STANDARD	FLOW AND TIMING	PATIENT SELECTION	INDEX TEST	REFERENCE STANDARD
Shahidi et al. 2014	?	☺	☺	☺	?	☺	☺
Gupta et al. 2015 & 2016	☺	☺	☺	☺	?	☺	☺
Zhang et al. 2016	?	☺	☹	☺	☺	☺	☺
Codari et al. 2017	?	☺	☺	☺	☹	☺	☺
Zhang et al. 2017	☺	☺	☹	☺	☺	☺	☺
De Jong et al. 2018	?	☺	☹	☺	?	☺	☺
Montúfar et al. 2018	?	☺	?	☺	?	☺	☺
Neelapu et al. 2018	?	☺	☺	☺	?	☺	☺
Torosdagli et al. 2019	☺	☺	☺	☺	☺	☺	☺
O'Neil et al. 2019	?	☺	?	☺	☹	☺	☺

☺Low Risk ☹High Risk ? Unclear Risk