



Deep learning segmentation of the mandibular canal on cone-beam computed tomography: a systematic review

Segmentación del conducto mandibular en tomografía computarizada de haz cónico mediante aprendizaje profundo: una revisión sistemática

Greta Cordovi Vallongo¹  

Youris Samón Mayan¹

¹ Universidad de Ciencias Médicas de la Habana. La Habana, Cuba

Submitted: 06-07-2025 | Revised: 25-08-2025 | Accepted: 29-10-2025 | Published: 31-10-2025

Corresponding Author: Greta Cordovi Vallongo 

ABSTRACT

Introduction: The mandibular canal houses the inferior alveolar nerve, and its accurate localization on cone-beam computed tomography is essential to avoid nerve injury during third molar surgery and implant placement. We synthesized the accuracy and validation practices of deep-learning models that segment the mandibular canal on cone-beam computed tomography. **Methods:** We searched PubMed, OpenAlex, and Crossref for studies published from 2013 to 2025, screened 245 records after deduplication, and included studies that evaluated a deep-learning model for mandibular canal segmentation against manual annotation. **Segmentation accuracy and reporting practices were extracted. Results:** Sixteen studies were included, published between 2020 and 2025 across nine countries. Eleven studies reported the Dice similarity coefficient, which ranged from 0.55 to 0.95, with the highest values above 0.85 and the lowest below 0.6. Accuracy varied widely with dataset, annotation protocol, and the inclusion of anatomical variants such as bifid canals and the anterior loop. Few studies used external validation, and public datasets were the exception rather than the rule. **Conclusions:** Deep-learning models can segment the mandibular canal on cone-beam computed tomography with Dice values that approach expert annotation, but the evidence is limited by small and mostly private datasets, heterogeneous reporting, and scarce external validation, so clinical deployment still requires prospective multicentre evaluation.

Keywords: cone-beam computed tomography; mandibular nerve; deep learning; image segmentation; artificial intelligence.

RESUMEN

Introducción: El conducto mandibular aloja el nervio alveolar inferior, y su localización precisa en tomografía computarizada de haz cónico es esencial para evitar lesiones nerviosas en cirugía de terceros molares y colocación de implantes. Sintetizamos la precisión y las prácticas de validación de los modelos de aprendizaje profundo que segmentan el conducto mandibular en tomografía computarizada de haz cónico. **Métodos:** Buscamos en PubMed, OpenAlex y Crossref estudios publicados de 2013 a 2025, cribamos 245 registros tras eliminar duplicados e incluimos estudios que evaluaran un modelo de aprendizaje profundo para segmentar el conducto mandibular frente a anotación manual. **Se extrajeron la precisión de segmentación y las prácticas de comunicación. Resultados:** Se incluyeron dieciséis estudios publicados entre 2020 y 2025 en nueve países. Once estudios comunicaron el coeficiente de similitud de Dice, que osciló entre 0,55 y 0,95, con los valores más altos por encima de 0,85 y los más bajos por debajo de 0,6. La precisión varió ampliamente según el conjunto de datos, el protocolo de anotación y la inclusión de variantes anatómicas como los conductos bífidos y el asa anterior. Pocos estudios usaron validación externa, y los conjuntos públicos fueron la excepción. **Conclusiones:** Los modelos de aprendizaje profundo pueden segmentar el conducto mandibular en tomografía computarizada de haz cónico con valores de Dice próximos a la anotación experta, pero la evidencia está limitada por conjuntos pequeños y mayoritariamente privados, comunicación heterogénea y escasa validación externa, por lo que el despliegue clínico requiere evaluación prospectiva multicéntrica.

Palabras clave: tomografía computarizada de haz cónico; nervio mandibular; aprendizaje profundo; segmentación de imágenes; inteligencia artificial.

INTRODUCTION

The mandibular canal carries the inferior alveolar nerve and its accompanying vessels through the body of the mandible, and its position is the single most important anatomical fact in much of oral surgery.

During third molar removal, the roots of the lower wisdom tooth frequently lie against or wrap around the canal, and a misjudged instrument can bruise, stretch, or transect the nerve, producing the numbness of the lower lip and chin that is the most common permanent complication of that operation.

The same injury can follow implant placement, when a drill or fixture invades the canal and produces sensory disturbance, hemorrhage, or an implant that must be removed. The stakes are therefore high, and the difficulty is that the canal is a thin, tortuous, low-contrast tube whose course varies from patient to patient and whose walls are often incomplete, especially near the mental foramen, where the canal bends and its anterior loop and bifid variants blur its outline.(1-3)

Its position relative to the buccal cortex, the molar apices, and the mandibular morphology differs systematically across skeletal patterns, so there is no single location that a surgeon can assume, and every patient must be assessed individually.(2,4)

The consequence of getting it wrong is not hypothetical: inferior alveolar nerve injury is a recognized and potentially permanent harm, and its avoidance is a central reason that three-dimensional imaging and careful planning have become standard of care, which in turn is the reason an automated, reliable trace of the canal would have immediate value.

Locating the canal reliably is therefore a problem of three-dimensional anatomy that a two-dimensional panoramic radiograph can only approximate, which is why cone-beam computed tomography is now the standard when the canal is judged to be close to the planned surgery, and why the comparison of the two modalities consistently favors the three-dimensional view for the critical decision.(3,5)

The clinical value of an accurate trace is concrete and twofold: it tells the surgeon where not to drill, and it shortens the preoperative planning that currently consumes the clinician's time slice by slice. Both benefits scale with the volume of surgery, and third molar removal and implant placement are among the most frequent procedures in oral surgery, so even a modest gain in the speed and reliability of canal tracing would translate into a large aggregate saving of time and a measurable reduction in avoidable nerve injury, and this is the practical motivation for the automated approaches that this review examines.

Cone-beam computed tomography solves the imaging problem but creates a new one: the clinician must trace the canal through dozens or hundreds of slices, a task that is time-consuming, operator-dependent, and known to differ between observers and even between a single observer's repeated attempts.

The canal is also a hard target for automated methods specifically, because it occupies a tiny fraction of the image volume, which produces a severe class imbalance, and because its boundaries are indistinct, so a small error in the placement of the wall is easy to make and hard to detect. This is a segmentation task, and it is exactly the kind of task for which deep convolutional networks have become the method of choice in medical imaging, following the self-configuring nnU-Net framework that removed much of the manual architecture engineering and set a reproducible baseline for three-dimensional segmentation.(6)

The same framework has already been extended to whole craniomaxillofacial scans for orthognathic planning, which demonstrates that multi-structure segmentation of the facial skeleton is feasible with a general-purpose network rather than a purpose-built one.(7)

Within dentistry specifically, deep learning has moved quickly from the early demonstration that a network could detect third molars and the mandibular nerve on panoramic radiographs to full three-dimensional segmentation of the jaws and teeth on cone-beam computed tomography, and the field is now documented by several reviews of deep learning applications on CBCT.(8-10)

A body of work has accumulated since 2020 that applies these networks to the mandibular canal itself, often reporting Dice similarity coefficients near or above expert agreement, and several reviews have begun to catalog it, most often as part of broader surveys of artificial intelligence in maxillofacial imaging or of tooth and jaw segmentation.(11-13)

The specific problem of the canal has also attracted dedicated reviews, including a systematic literature review of mandibular canal segmentation in digital radiographs and a state-of-the-art review of deep learning for the mandibular nerve canal.(14,15) What these reviews converge on is that the field is growing rapidly, that the reported accuracy is generally high, and that the evidence for generalization beyond the training institution is thin, but they differ in scope, in the imaging

modalities covered, and in how recently they were conducted.

What the field still lacks is a focused, current synthesis of how accurately the canal itself is segmented on cone-beam computed tomography, which architectures and training strategies are used, and, crucially, whether the reported numbers come from internal or external validation and from public or private data.

This review therefore asks how accurately deep-learning models segment the mandibular canal on cone-beam computed tomography, and how their performance is reported and validated. The answer is organized around three tables: the characteristics of the included studies, their reported segmentation accuracy, and their reporting and validation practices.

The aim is not only to summarize the numbers but to show where they come from, because a Dice value computed on a small private dataset with internal splitting does not mean the same thing as one computed on an external, multicentre set, and the difference between those two kinds of evidence is exactly the difference between a laboratory result and a clinical tool.

Two subsidiary questions follow from that aim: which architectural and training choices, if any, are consistently associated with better accuracy, and how much of the reported variation reflects the method as opposed to the data. Answering them honestly requires resisting the temptation to rank the models against one another, because the studies are not comparable enough for a ranking to be meaningful, and the review therefore emphasizes the conditions under which each figure was produced rather than the figure alone.

METHODS

The review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) and was not registered prospectively, which is stated rather than concealed. Studies were eligible if they developed or evaluated a deep-learning model for segmentation of the mandibular canal or inferior alveolar canal on human cone-beam computed tomography, used a manual expert annotation as the reference standard, and reported at least one quantitative segmentation metric.

We excluded studies using panoramic or other two-dimensional radiographs, studies that only classified or detected the relationship between the third molar and the canal without producing a segmentation, studies of other anatomical structures such as the teeth, the temporomandibular joint, or the maxilla, reviews and editorials, and non-English reports, and we restricted the window to publications from 2013 through 2025.

Eligible designs included model-development studies with an internal test set, and studies with external or multicentre validation were retained but distinguished in the synthesis, because the two designs provide different strength of evidence.

PubMed, OpenAlex, and Crossref were searched with terms for deep learning ("deep learning", "convolutional neural network", "U-Net", "segmentation"), the target structure ("mandibular canal", "inferior alveolar canal", "inferior alveolar nerve"), and the modality ("cone-beam computed tomography", "CBCT").

A representative string combined these groups with Boolean operators, for example ("deep learning" OR "convolutional neural network" OR "U-Net") AND ("mandibular canal" OR "inferior alveolar canal") AND ("cone-beam computed tomography" OR "CBCT"), and the complete search log is reproducible from the review records. Records were pooled across the three sources, deduplicated, and screened first by title and abstract and then by full text where a legitimate open-access version was available.

Screening was performed by a single reviewer assisted by artificial intelligence and then verified, and this is reported honestly rather than described as independent dual review; disagreements were resolved by rereading the record against the criteria.

For each included study we extracted the first author and year, country, dataset size and source, the model architecture, the reference standard, and the principal segmentation metric with its point estimate, and no field was completed without evidence, so where a study did not report a quantity it was left blank rather than imputed. Risk of bias was appraised across four domains adapted from QUADAS-2, namely patient selection, the index test, the reference standard, and flow and timing.

Patient selection was judged high risk when the study used a convenience or case-control sample, the index test was judged high or unclear when the model was evaluated on the same data used for training without an independent set, the reference standard was judged on whether the manual annotation was performed by experts and on how many observers were involved, and flow

and timing was judged on whether the interval and exclusion of poor-quality scans were reported.

The category “unclear” was used liberally and a judgment of low risk was never assigned on the basis of missing information. A meta-analysis was not performed because the studies reported different metrics, different reference annotations, and different datasets, and any pooled estimate would have been uninformative; the synthesis is therefore narrative and qualitative, with no summary receiver operating characteristic or pooled Dice value.

The searches were executed during the review period, and the year restriction was applied to both the search and the screening so that no study outside the 2013-2025 window could enter the review.

RESULTS

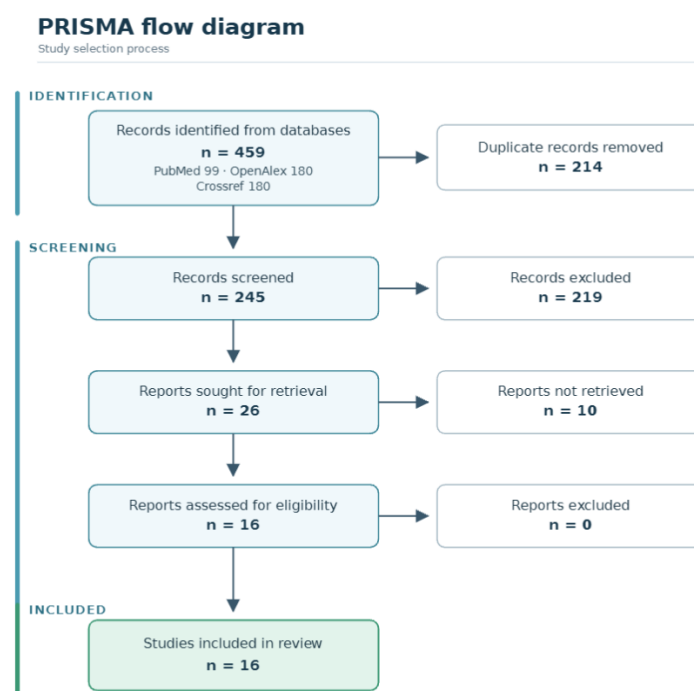
Study selection

The searches identified 459 records, 214 of which were duplicates, leaving 245 for screening (Figure 1). Title and abstract screening excluded 219 records, the large majority of which were reviews, studies of the third molar rather than the canal, or studies using panoramic radiographs rather than cone-beam computed tomography, and the exclusions reflect the breadth of the original search rather than a narrow first pass.

A further 26 reports were sought for retrieval, 10 of which could not be retrieved in open access and were therefore not assessed, an unavoidable constraint of a review without subscription access.

The 16 reports that were assessed in full text all met the eligibility criteria, and 16 studies were included in the review. No study required exclusion at the full-text stage, which suggests that the title-and-abstract screening had been conservative and that the included set is a faithful representation of the retrievable primary literature.

Figure 1. PRISMA 2020 flow diagram of study selection.



Characteristics of the included studies

The sixteen studies were published between 2020 and 2025, which confirms that this is a young field, and they originated from nine countries, with the largest contributions from China and South Korea and a smaller but important European cluster from Finland and Italy, plus single studies from Brazil, Saudi Arabia, and Turkey (Table 1).

The absence of any study before 2020 reflects the time it took for three-dimensional dental

segmentation to move from the whole-jaw problem to the canal itself, and the acceleration in the last two years of the window mirrors the general growth of deep learning in dentistry. The earliest work used two-dimensional SegNet and U-Net architectures before the field moved to three dimensions, and the most recent work spans transformers, active learning, and self-configuring frameworks.(16,17)

The chronological arc is visible in the designs: the first studies established feasibility with modest architectures, the middle period introduced attention and continuity mechanisms, and the latest studies adopted the self-configuring nnU-Net as a standard and turned their attention to data efficiency and generalization.

Dataset size ranged from a few dozen to nearly a thousand scans, with the largest being a multicentre set of 982 training and 150 evaluation volumes drawn from five scanners in two regions, which is the only study in the set that approached the scale needed for robust generalization claims.(18)

Most other studies used between 60 and 240 volumes, and the smallest used only a few dozen, a scale at which the risk of overfitting to the particular annotation style is considerable. Architectures were dominated by the U-Net family, and specifically by 3D U-Net and its nnU-Net variants, with several studies adding attention modules, bidirectional convolutional long short-term memory, or continuity-aware losses aimed at the canal's thin, elongated geometry.(19,20)

The bilateral symmetry of the canal was exploited in a dual-stage design that first localized the volumes of interest and then segmented within them, and the tubular structure was exploited in a Frenet-frame transform that unwraps the canal into a straight volume.(19,21) Three studies released their datasets or source code publicly, and the remainder used private institutional data.(22-24)

Every study used manual expert annotation as the reference standard, but the granularity of that annotation, whether voxel-level, coarse, consensus-based, or single-observer, differed, and this difference matters for the interpretation of the reported accuracy, because a model can only be as good as the labels it is trained to reproduce.

The distinction between coarse and voxel-level annotation is particularly consequential: a model trained on coarse, tube-like labels will learn the centerline of the canal but not its precise walls, and its Dice against a voxel-level reference will therefore be lower than a model trained and tested at the same granularity, which is one reason the figures in this review are not directly comparable across studies.

Table 1. Characteristics of the included studies.

Study (year)	Country	Dataset	Architecture	Reference standard
Kwak et al. (2020)	South Korea	CBCT scans	2D SegNet, 2D and 3D U-Net	Manual segmentation
Jaskari et al. (2020)	Finland	637 + 15 CBCT	Fully convolutional network	Radiologist annotation
Lim et al. (2021)	South Korea	138 CBCT (3 centres)	3D nnU-Net, active learning	Manual segmentation
Usman et al. (2022)	Saudi Arabia	500 + 15 CBCT	Dual-stage attention U-Net	Manual annotation
Järnstedt et al. (2022)	Finland	982 + 150 CBCT (5 scanners)	Deep learning system	Four radiologists
Jeoun et al. (2022)	South Korea	CBCT volumes	Canal-Net (3D U-Net + ConvLSTM)	Manual segmentation
Cipriano et al. (2022)	Italy	Public CBCT dataset	Deep learning (Pulpy3D)	Expert clinicians
Lin et al. (2023)	China	220 CBCT	Two-stage 3D-UNet	Two senior dentists
Zhao et al. (2023)	China	CBCT volumes	Frenet-frame with cIDice	Manual segmentation
Oliveira-Santos et al. (2023)	Brazil	CBCT scans	AI segmentation tool	Manual delineation
Lv et al. (2023)	China	Public MC dataset	Transformer + cIDice	Manual segmentation
Järnstedt et al. (2023)	Finland	165 CBCT (72 patients)	Deep learning system	Radiologist
Huang et al. (2025)	China	236 CBCT	ManCan_ResU-Net (3D U-Net)	Fine-grained annotation
Yang et al. (2025)	China	Public and external	2D/2.5D/3D and transformer	Manual segmentation
Gümüşsoy et al. (2025)	Turkey	69 CBCT	nnU-Net v2	Manual annotation
On et al. (2025)	South Korea	60 CBCT + external	3D nnU-Net, multi-label	Manual segmentation

Note: CNN = convolutional neural network; MC = mandibular canal; nnU-Net = self-configuring segmentation framework.

Segmentation accuracy

Eleven of the sixteen studies reported the Dice similarity coefficient, and the values span a wide range from 0.55 to 0.95, a spread that is itself the most informative result of the review (Figure 2, Table 2). The highest values came from architectures and training strategies that addressed the canal's specific geometry: a two-stage 3D-UNet reached a Dice of 0.875, a continuity-aware Canal-Net reached 0.87, and a Frenet-frame transform with cIDice loss reached 0.865.(20,21,25)

A custom 3D U-Net with connected-component post-processing reached a Dice of 0.85 with a 95th percentile Hausdorff distance of 1.8 mm, and a transformer with a connectivity loss reached a Dice of 0.844.(23,24) The lowest values came from studies that used coarser or external annotations or the more demanding task of segmenting bifid canals, where one model achieved a Dice of 0.82 for the main canal but only 0.46 for the bifid branch.(26)

Studies that reported distance metrics instead of overlap told a consistent story of sub-millimetre localization, with a mean curve distance of 0.56 mm in the earliest fully convolutional approach and a mean surface distance of 0.74 mm in the multicentre validation.(18,27) Two studies reported global accuracy rather than overlap, with values of 0.82 and 0.997, the latter from a study that emphasized the anterior loop, and one reported a mean intersection over union of 0.795, a figure that, like global accuracy, is inflated by the dominance of the easy background voxels and therefore should not be read as equivalent to a Dice value.(16,19,28)

The wide dispersion warns against quoting any single figure as the accuracy of the field. Part of that dispersion is also a difference in what was counted: some studies evaluated the full three-dimensional canal, others the two-dimensional projections, and one evaluated a multi-label segmentation of several structures together, so a Dice of 0.95 from the multi-label study is not comparable to a Dice of 0.548 from a single-structure trace of a temporal dataset.(17,29)

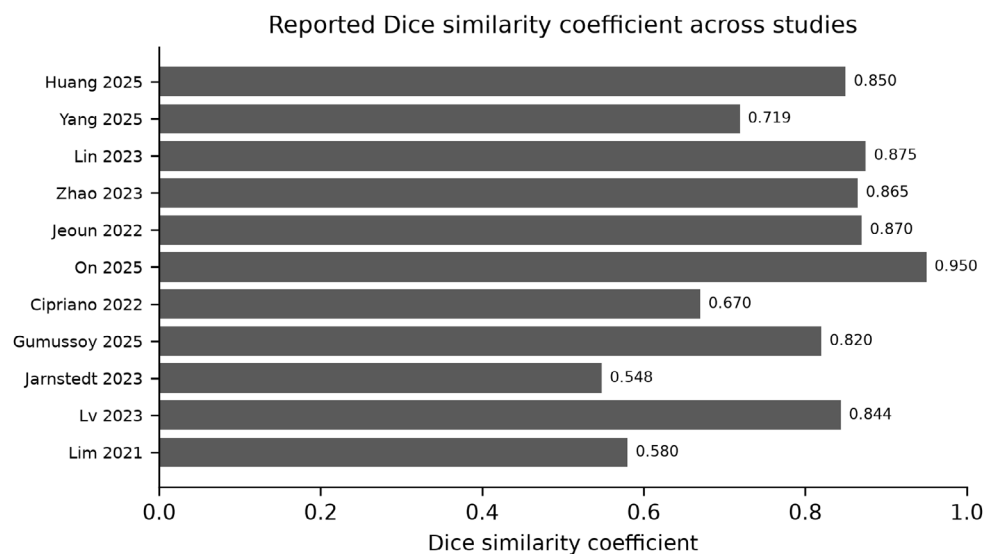
The distance metrics, where reported, show that the models localize the canal to within about half a millimetre of the reference on average, which is the more clinically meaningful number for a nerve that must be avoided by a margin, and these values are more stable across studies than the overlap measures.(18,27)

A reader looking for a single summary should therefore hold two numbers in mind rather than one: the overlap is highly variable and dataset-dependent, while the localization error is consistently sub-millimetre.

Table 2. Reported segmentation accuracy of the included studies.

Study (year)	Metric	Reported value
Kwak et al. (2020)	Global accuracy	0.82
Jaskari et al. (2020)	Mean curve distance / ASSD	0.56 mm / 0.45 mm
Lim et al. (2021)	DSC (internal / external)	0.58 / 0.49
Usman et al. (2022)	mIoU	0.795
Järnstedt et al. (2022)	Mean surface distance	0.74 mm
Jeoun et al. (2022)	DSC / IoU	0.87 / 0.80
Cipriano et al. (2022)	Dice (dense labels)	0.67
Lin et al. (2023)	DSC / 95% HD	0.875 / 0.442 mm
Zhao et al. (2023)	Dice / cIDice	0.865 / 0.971
Oliveira-Santos et al. (2023)	Global accuracy	0.997
Lv et al. (2023)	Dice / IoU	0.844 / 0.731
Järnstedt et al. (2023)	DSC / SMCD	0.548 / 0.643 mm
Huang et al. (2025)	DSC / HD95	0.85 / 1.8 mm
Yang et al. (2025)	DSC (public / external)	0.719 / 0.716
Gümüşsoy et al. (2025)	DSC (MC / bifid)	0.82 / 0.46
On et al. (2025)	DSC (multi-label / single)	0.95 / 0.91

Note: DSC = Dice similarity coefficient; IoU = intersection over union; HD = Hausdorff distance; ASSD = average symmetric surface distance; SMCD = symmetric mean curve distance; — = not reported in the extractable text.

Figure 2. Reported Dice similarity coefficient across the included studies.

Reporting and validation

The reporting and validation practices reveal the gap between the headline numbers and the evidence that would support deployment (Table 3). Only 4 studies validated their model on data from outside the training institution, and of the remaining twelve, most relied on a random internal split, which tests the model on scans from the same scanner, the same acquisition protocol, and the same annotators as the training data.

An internal split is not valueless, it does establish that the model can segment scans it has not memorized, but it says nothing about the scanner, population, and annotation drift that a deployed tool would encounter, and the difference between the two designs is the difference between internal consistency and generalization. One of those, an active-learning study across three hospitals, saw its Dice fall from 0.58 internally to 0.49 externally, a concrete demonstration of how internal performance overstates generalization.(30)

Another external comparison, using a multi-label nnU-Net, reported a Dice of 0.95 internally but did not sustain that figure on the external set, and a systematic comparison of two-dimensional, two-and-a-half-dimensional, and three-dimensional networks found that the three-dimensional network generalized better to an external dataset.(17,31)

The single study that compared a model directly against multiple radiologists on a multicentre set found that the model was slightly more consistent than the observers themselves, with a mean surface distance of 0.74 mm against an interobserver variability of 0.77 mm, and that it generalized to scanners unseen in training, which is the strongest validation evidence in the included set.(18)

A follow-up reproducibility study on a temporal dataset reported a Dice of 0.548 and identified postoperative changes as the main source of error, a reminder that even a validated model will degrade on the surgically altered, artifact-laden scans that are common in the very patients who need a second assessment.(29)

Public datasets or code were released by only 3 studies, and the rest trained on private institutional data that cannot be reproduced, which limits both comparison and independent verification.(22-24)

The public releases are nonetheless consequential, because the Pulpy3D dataset and the ToothFairy challenge together provide the shared benchmark that the rest of the field lacked, and the study by Cipriano et al. showed concretely how much a well-annotated public dataset improves a model, with its Dice rising as the annotation density increased.(22)

Reporting was heterogeneous: some studies reported Dice and Hausdorff distance, others accuracy or intersection over union, and few reported all metrics, confidence intervals, or the full annotation protocol, which makes the individual figures harder to compare than they should be. The overall picture is of a field that has produced strong single-centre results and has begun, but only begun, to build the shared data and the external validation that would turn those results into

clinical evidence.

The trajectory is nonetheless positive: the earliest studies in this review did not release data or validate externally at all, whereas the most recent studies increasingly do, and the appearance of public datasets and a challenge during the review window marks a genuine shift in the field's norms, even if the shift has not yet reached the majority of published work.

Table 3. Reporting and validation practices of the included studies.

Study (year)	External validation	Public data/code	RoB
Kwak et al. (2020)	No	No	U
Jaskari et al. (2020)	No	No	U
Lim et al. (2021)	Yes (40 external)	No	U
Usman et al. (2022)	No	No	U
Järnstedt et al. (2022)	Multicentre, new scanners	No	L
Jeoun et al. (2022)	No	No	U
Cipriano et al. (2022)	No	Yes (dataset)	U
Lin et al. (2023)	No	No	U
Zhao et al. (2023)	No	No	U
Oliveira-Santos et al. (2023)	No	No	U
Lv et al. (2023)	Public dataset	Yes (dataset)	U
Järnstedt et al. (2023)	Temporal dataset	No	U
Huang et al. (2025)	No	Yes (dataset and code)	U
Yang et al. (2025)	Yes (external)	Public dataset	U
Gümüşsoy et al. (2025)	No	No	U
On et al. (2025)	Yes (external)	No	U

Note: RoB = risk of bias; H = high; U = unclear; L = low, judged per study across the four adapted QUADAS-2 domains.

DISCUSSION

The characteristics table shows a field that is young, geographically concentrated in East Asia and Northern Europe, and built almost entirely on the U-Net family. The dominance of 3D U-Net and nnU-Net is unsurprising, because the self-configuring framework set the standard for biomedical segmentation and removed the need for bespoke architecture design, and the studies that departed from it did so to address the canal's specific geometry rather than to chase novelty.(6,32)

The layered approach to whole-mandible segmentation that predates the canal-specific work also shaped the field by showing that a cascade of coarse-to-fine stages could handle the very different scales of the jaw and its internal structures, and the two-stage and dual-stage designs that followed in the canal literature are direct descendants of that idea.(19,32)

The thin, elongated, low-contrast nature of the canal is the recurrent technical challenge, and the most accurate methods in this review, the two-stage, continuity-aware, and Frenet-frame designs, are all responses to it, a pattern consistent with broader reviews of maxillofacial segmentation that identify the small size and indistinct boundary of the canal as the limiting factors.(12,33)

The move from two-dimensional to three-dimensional networks over the review window mirrors the wider trajectory of medical image segmentation, and the direct comparison by Yang et al. showing that a 3D network beat its two-dimensional counterparts on both internal and external data provides the cleanest evidence for that move within this specific task.(31)

The architectural innovations that followed, attention modules, continuity-aware losses, and the Frenet-frame transform, are all attempts to exploit the fact that the canal is not an arbitrary blob but a connected tube, and the consistency with which these structure-aware methods cluster at the top of the accuracy table suggests that the prior information embedded in them is doing real work.(19-21)The geographic distribution also mirrors the wider dental artificial-intelligence literature, and the near absence of African and Latin American data, with a single Brazilian study, is a gap worth noting for generalizability, because canal morphology and the surrounding bone quality vary across populations and scanners, and a model trained on East Asian and European scans cannot be assumed to perform identically on other populations.(28)

The concentration of the field in a few countries also reflects where the annotated data and the engineering capacity reside, and the public datasets released by the included studies are a partial remedy, because they allow groups without their own annotation pipeline to train and to benchmark, which is the mechanism by which the geographic center of such a field can broaden.

The accuracy table delivers a caution that is more important than any single number: the Dice similarity coefficient across the included studies ranges over more than 0.4, from 0.55 to 0.95, and that range is explained by the dataset, the annotation, and the task, not by the model. The highest figures, near 0.87, come from clean, carefully annotated, single-centre datasets and from models tuned to the canal's continuity, while the lowest figures come from external validation, coarse annotations, or the harder problem of bifid canals.(20,25,26)

This mirrors the wider observation that segmentation performance on small, thin structures is dominated by annotation quality and dataset composition, and it is consistent with the Tooth-Fairy challenge, which found a large spread across submitted algorithms on a common, difficult benchmark, with the best entries separating themselves from the rest largely through careful preprocessing and postprocessing rather than through radically different architectures.(34)

The distance metrics, where reported, tell a reassuring story of sub-millimetre localization, with mean curve and surface distances around half a millimetre, which is clinically relevant because it is the distance from the planned drill path to the nerve that determines safety, and a localization error of half a millimetre is within the safety margin that most implant protocols already apply.(18,27)

The honest reading is that a well-annotated internal dataset yields excellent numbers that do not yet transfer to the messy, heterogeneous scans of routine practice, and the gap between the two is the single most important caveat for anyone reading this literature with an eye to adoption.

A clinician evaluating a vendor's claim of a particular Dice value should therefore ask not what the number is but where it came from: on which dataset, against which annotation, and on scans from the same or a different institution, because the answer to those questions, more than the number itself, determines whether the model is likely to work in their own clinic.

The same principle applies to any reader of this review: a Dice value detached from its provenance is not evidence, and the tables here deliberately attach every figure to its study, dataset, and validation design so that the provenance can be checked.

A word on the metric itself is warranted, because the Dice similarity coefficient is unusually harsh for a thin tubular structure. The mandibular canal is only a few voxels wide, so a single-voxel displacement at the boundary penalizes the Dice far more than the same displacement would for a large organ, and a small absolute error therefore appears as a large apparent drop in overlap.

This is why the distance metrics in this review, which are stable around half a millimetre, are arguably a fairer summary of clinical accuracy than the Dice values, which swing from 0.548 to 0.95 across studies that differ less in true performance than in annotation and dataset.(17,29)

Future studies should report both, because a Dice without a distance measure hides whether a low overlap reflects a truly wrong trace or merely a boundary that is a voxel out of place, and the distinction is exactly what a clinician cares about when deciding whether a segmentation is safe to plan against.

The reporting and validation table explains why the field is not yet clinically deployable despite good headline accuracy. External validation is the exception rather than the rule, and where it was performed, performance fell, as in the active-learning study that dropped from 0.58 to 0.49 and the multi-label study whose external figures did not sustain its internal ones.(17,30)

The one multicentre study with a human benchmark is the most credible evidence in the review, because it showed both that the model matched expert consistency and that it generalized to unseen scanners, which is the property a clinical tool actually needs.(18)

The scarcity of public data, with only three studies releasing datasets or code, reproduces a known weakness of the field that the ToothFairy and Pulpy3D releases were explicitly designed to fix, and the absence of shared benchmarks makes the published numbers difficult to compare or to reproduce.(22,34)

The consequences of private data go beyond comparability: without a shared reference, a reader cannot tell whether a low Dice in one study reflects a worse model or simply a harder dataset, and the field cannot accumulate knowledge in the way that a common benchmark permits. The prevalence of unclear risk of bias across the adapted QUADAS-2 domains is a direct consequence of incomplete reporting, and it is the reporting, more than the modeling, that most needs to improve.

It is also instructive to place the segmentation task beside the adjacent literature that detects or classifies the relationship between the third molar and the canal, because the two problems share the same anatomical target but ask different questions.

A substantial parallel literature has used deep learning to determine whether a third molar

contacts, overlaps, or merely approaches the canal on CBCT or panoramic radiographs, with fused, distance-aware, and two-stage networks reporting high classification accuracy, and these methods are closer to clinical deployment precisely because a categorical answer is easier to validate than a voxel-wise trace.(35-38)

The segmentation task is more demanding and more useful, because a complete trace of the canal supports measurement of the true safety margin in three dimensions, not just a yes-or-no judgment of proximity, and the segmentation literature reviewed here can be read as the natural successor to that detection work.(39,40) The detection literature has its own value, because a proximity classifier can be validated against a binary outcome more easily than a trace against a full annotation, and a systematic review of automated proximity quantification on CBCT has catalogued accuracy figures that are, on their face, higher than the segmentation figures reviewed here, but the two are not measuring the same thing and should not be directly compared.(40)

A clinician who only needs to know whether the nerve is at risk may be served today by a classifier, while a clinician who needs to plan an implant trajectory needs the segmentation, and the higher bar of the segmentation task explains why its validation is harder and its reported accuracy more variable.

The review has limitations. The search was restricted to three databases and to English, so work published in Korean or Chinese journals, of which there is likely a substantial amount given the geographic center of the field, was excluded, and the result may understate East Asian research.

Full-text retrieval was limited to open access, which left 10 otherwise relevant reports unassessed and may have biased the included set toward open-access venues, a selection that could distort both the accuracy range and the reporting-practice summary.

Screening was performed by a single reviewer assisted by artificial intelligence rather than by two independent reviewers, and single-reviewer screening is less reliable than dual screening. No meta-analysis was possible because of the heterogeneity of metrics, annotations, and datasets, so the synthesis is qualitative and the accuracy figures are not pooled. Each of these is a target for a future update rather than a reason to discount the main findings, but they mean the review should be read as a map of the field rather than a precise audit of its central tendency.

The limitation of most consequence is the language restriction, because the two countries that contribute the most studies in this review, China and South Korea, also maintain substantial national-language literatures, and an update that searched Chinese and Korean databases would likely enlarge the evidence base and could shift the accuracy range.

A second consequential limitation is the restriction to open-access full text, which may have excluded larger and more carefully validated studies published in subscription journals, and if those studies are systematically better validated than the open-access ones, the present synthesis would understate the field's maturity.

A third limitation is the absence of a formal certainty-of-evidence assessment; the risk-of-bias judgments were made study by study, but no GRADE-style summary was produced, so the overall confidence in the findings is stated qualitatively rather than graded. These limitations together mean that the review's conclusions are robust for the direction of the evidence, high accuracy on good data with poor generalization evidence, but not for the exact magnitude of the accuracy, which the heterogeneity of the field precludes stating precisely.

The implications for practice and research follow directly from the three tables. For practice, the models are not yet ready to replace the clinician's tracing of the canal, but they are close enough on well-annotated data to serve as a second reader, and the multicentre study suggests they can already match human consistency on suitable scans, which is the first thing a clinician would want to know before trusting an automated trace.(18)

A realistic near-term role is therefore triage and editing rather than full autonomy: the model produces a draft trace in seconds, and the clinician reviews and corrects it, which is faster than tracing from scratch but leaves the responsibility where it belongs.

This editing workflow has the further advantage that it does not require the model to be perfect, only to be close enough that correction is quicker than de novo tracing, and the sub-millimetre localization errors reported in this review suggest that many of the included models already clear that lower bar.

For research, the priorities are external validation on consecutive, unselected patients, the release of annotated datasets and code, the adoption of a common metric set that includes Dice, a distance measure, and confidence intervals, and head-to-head comparison against human observers. The canal is a small target, but it is a well-defined one, and a field that has already reached Dice values above 0.85 on good data has a clear path to the external, prospective evidence that would justify clinical use.(20,25)

The immediate, achievable steps are not glamorous but they are decisive: agree on a minimum reporting set, release the data that make comparison possible, and validate on patients rather than on the same scans the model has already seen, and none of these requires a technical advance, only a collective decision by the groups working in the field. If the field does those three things, the next systematic review of this topic will be able to do what this one could not, which is to state, with a pooled estimate and a confidence interval, how accurately the mandibular canal can be segmented, and the answer will then be usable by clinicians rather than merely interesting to them.

CONCLUSIONS

Deep-learning models segment the mandibular canal on cone-beam computed tomography with Dice similarity coefficients that reach expert-level consistency on well-annotated, single-centre data, but the reported values vary widely with the dataset, the annotation protocol, and the inclusion of difficult variants such as bifid canals.

The field is dominated by the U-Net family, and the most accurate methods are those that address the canal's thin, elongated geometry, while the localization error, where reported, is consistently sub-millimetre, which is the more clinically reassuring of the two findings.

The evidence is limited by small and mostly private datasets, heterogeneous reporting, and scarce external validation, and where external validation was performed, performance fell, so the published figures should be read as upper bounds on clinical performance rather than as predictions of it.

The next step is prospective multicentre validation on unselected patients with shared datasets and a common metric set that reports both overlap and distance, and until that exists the models should be regarded as promising research tools rather than as replacements for expert tracing.

The question the field must now answer is not whether the canal can be segmented, which the evidence shows it can, but whether it can be segmented reliably enough, in the scans of the patients who will actually need it, to be trusted with a nerve, and the honest answer, on the evidence assembled here, is not yet.

REFERENCES

1. Balel Y, Coşkun B, Kaplan Ç. Three-dimensional analysis of the inferior alveolar nerve in relation to mandibular morphology. *BMC Oral Health*. 2025;25(1):1813. doi: 10.1186/s12903-025-07244-z
2. Meyyappan S, Rangaswamy S, Singh M. Determination of anatomical position of inferior alveolar nerve canal in relation to buccal cortical bone in patients with different mandibular skeletal relationships. *Natl J Maxillofac Surg*. 2025;16(2):328-337. doi: 10.4103/njms.njms_7_24
3. Urechescu H, Banu A, Pricop M, Streian F, Pricop A, Cuzic C. The Positional Relationship Between the Mandibular Canal and the Lower Third Molar Determined on Cone-Beam Computed Tomography. *Medicina*. 2025;61(7):1291. doi: 10.3390/medicina61071291
4. Sehgal A, Saxena V, Rangarajan H, Gopalakrishnan V, Dewan K. A Clinico-Radiological Study to Evaluate the Risk of Inferior Alveolar Nerve Damage in Impacted Mandibular Third Molar Removal. *J Maxillofac Oral Surg*. 2025;24(3):775-784. doi: 10.1007/s12663-025-02498-0
5. Couto DAF, Tanaka JLO, Souza PHC, Ignacio SA, Carneiro E, Alanis LRA. Comparison of Panoramic Radiography and Cone Beam Computed Tomography for Lower Third Molar Assessment Among Different Dental Specialists. *Journal of Oral and Maxillofacial Surgery*. 2025;83(8):1000-1011. doi: 10.1016/j.joms.2025.05.003
6. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. 2021;18(2):203-211. doi: 10.1038/s41592-020-01008-z
7. Dot G, Schouman T, Dubois G, Rouch P, Gajny L. Fully automatic segmentation of craniomaxillofacial CT scans for computer-assisted orthognathic surgery planning using the nnU-Net framework. *Eur Radiol*. 2022;32(6):3639-3648. doi: 10.1007/s00330-021-08455-y
8. Vinayahalingam S, Xi T, Bergé S, Maal T, de Jong G. Automated detection of third molars and mandibular nerve by deep learning. *Sci Rep*. 2019;9(1):9007. doi: 10.1038/s41598-019-45487-3
9. Fan W, Zhang J, Wang N, Li J, Hu L. The Application of Deep Learning on CBCT in Dentistry. *Diagnostics*. 2023;13(12):2056. doi: 10.3390/diagnostics13122056
10. Zhang L, Li W, Lv J, Xu J, Zhou H, Li G, et al. Advancements in oral and maxillofacial surgery medical images segmentation techniques: An overview. *Journal of Dentistry*. 2023;138:104727. doi: 10.1016/j.jdent.2023.104727
11. Issa J, Olszewski R, Dyszkiewicz-Konwińska M. The Effectiveness of Semi-Automated and Fully Automatic Segmentation for Inferior Alveolar Canal Localization on CBCT Scans: A Systematic Review. *IJERPH*. 2022;19(1):560. doi: 10.3390/ijerph19010560
12. Alahmari M, Alahmari M, Almuaddi A, Abdelmagyd H, Rao K, Hamdoon Z, et al. Accuracy of artificial intelligence-based segmentation in maxillofacial structures: a systematic review. *BMC Oral Health*. 2025;25(1):350. doi: 10.1186/s12903-025-05730-y
13. Kot WY, Au Yeung SY, Leung YY, Leung PH, Yang WF. Evolution of deep learning tooth segmentation from CT/CBCT images: a systematic review and meta-analysis. *BMC Oral Health*. 2025;25(1):800. doi: 10.1186/s12903-025-05984-6
14. Farid Naufal M, Fatichah C, Renwi Astuti E, Hardani Putra R. Deep Learning for Mandibular Canal Segmentation in Digital Dental Radiographs: A Systematic Literature Review. *IEEE Access*. 2024;12:76794-76815. doi: 10.1109/access.2024.3406342
15. Dashti M, Khosraviani F, Amirzade-Iranaq MH, Tajbakhsh N, Rezaei Taleshi SF, Sofia Ani, et al. Application of Deep Learning Algorithms in Segmentation of Mandibular Nerve Canal in Orthopantomogram (Panoramic) Radiographs: A State-Of-Art Systematic Review. *AAIML*. 2024;04(04):3173-3185. doi: 10.54364/aauml.2024.44182
16. Kwak GH, Kwak EJ, Song JM, Park HR, Jung YH, Cho BH, et al. Automatic mandibular canal detection using a deep convolutional neural network. *Sci Rep*. 2020;10(1):5711. doi: 10.1038/s41598-020-62586-8
17. On S, Ock J, Bae M, Park JW, Baek SH, Ham S, et al. Improving accuracy for inferior alveolar nerve segmentation with multi-label of anatomical adjacent structures using active learning in cone-beam computed tomography. *Sci Rep*. 2025;15(1):7441. doi: 10.1038/s41598-025-91725-2
18. Järnstedt J, Sahlsten J, Jaskari J, Kaski K, Mehtonen H, Lin Z, et al. Comparison of deep learning segmentation and multigrader-annotated mandibular canals of multicenter CBCT scans. *Sci Rep*. 2022;12(1):18598. doi: 10.1038/s41598-022-20605-w
19. Usman M, Rehman A, Saleem AM, Jawaid R, Byon SS, Kim SH, et al. Dual-Stage Deeply Supervised Attention-Based Convolutional Neural Networks for Mandibular Canal Segmentation in CBCT Scans. *Sensors*. 2022;22(24):9877. doi: 10.3390/s22249877
20. Jeoun BS, Yang S, Lee SJ, Kim TI, Kim JM, Kim JE, et al. Canal-Net for automatic and robust 3D segmentation of mandibular canals in CBCT images using a continuity-aware contextual network. *Sci Rep*. 2022;12(1):13460. doi: 10.1038/s41598-022-17341-6
21. Zhao H, Chen J, Yun Z, Feng Q, Zhong L, Yang W. Whole mandibular canal segmentation using transformed dental CBCT volume in Frenet frame. *Heliyon*. 2023;9(7):e17651. doi: 10.1016/j.heliyon.2023.e17651
22. Cipriano M, Allegretti S, Bolelli F, Di Bartolomeo M, Pollastri F, Pellacani A, et al. Deep Segmentation of the Mandibular Canal: A New 3D Annotated Dataset of CBCT Volumes. *IEEE Access*. 2022;10:11500-11510. doi: 10.1109/access.2022.3144840
23. Huang J, Jie J, Ma H, Xie S, Liao H, Ouyang K, et al. Deep learning for automated mandibular canal segmentation in CBCT scans. *BMC Oral Health*. 2025;25(1):1699. doi: 10.1186/s12903-025-07098-5
24. Lv J, Zhang L, Xu J, Li W, Li G, Zhou H. Automatic segmentation of mandibular canal using transformer based neural networks. *Front Bioeng Biotechnol*. 2023;11:1302524. doi: 10.3389/fbioe.2023.1302524

25. Lin X, Xin W, Huang J, Jing Y, Liu P, Han J, et al. Accurate mandibular canal segmentation of dental CBCT using a two-stage 3D-UNet based segmentation framework. *BMC Oral Health*. 2023;23(1):551. doi: 10.1186/s12903-023-03279-2
26. Gumussoy I, Demirezer K, Duman SB, Haylaz E, Bayrakdar IS, Celik O, et al. AI-powered segmentation of bifid mandibular canals using CBCT. *BMC Oral Health*. 2025;25(1):907. doi: 10.1186/s12903-025-06311-9
27. Jaskari J, Sahlsten J, Järnstedt J, Mehtonen H, Karhu K, Sundqvist O, et al. Deep Learning Method for Mandibular Canal Segmentation in Dental Cone Beam Computed Tomography Volumes. *Sci Rep*. 2020;10(1):5842. doi: 10.1038/s41598-020-62321-3
28. Oliveira-Santos N, Jacobs R, Picoli FF, Lahoud P, Niclaes L, Groppo FC. Automated segmentation of the mandibular canal and its anterior loop by deep learning. *Sci Rep*. 2023;13(1):10819. doi: 10.1038/s41598-023-37798-3
29. Järnstedt J, Sahlsten J, Jaskari J, Kaski K, Mehtonen H, Hietanen A, et al. Reproducibility analysis of automated deep learning based localisation of mandibular canals on a temporal CBCT dataset. *Sci Rep*. 2023;13(1):14159. doi: 10.1038/s41598-023-40516-8
30. Lim HK, Jung SK, Kim SH, Cho Y, Song IS. Deep semi-supervised learning for automatic segmentation of inferior alveolar nerve using a convolutional neural network. *BMC Oral Health*. 2021;21(1):630. doi: 10.1186/s12903-021-01983-5
31. Yang S, Jeong JS, Song D, Han JY, Lim SH, Kim S, et al. Comparison of 2D, 2.5D, and 3D segmentation networks for mandibular canals in CBCT images: a study on public and external datasets. *BMC Oral Health*. 2025;25(1):1126. doi: 10.1186/s12903-025-06483-4
32. Verhelst PJ, Smolders A, Beznik T, Meewis J, Vandemeulebroucke A, Shaheen E, et al. Layered deep learning for automatic mandibular segmentation in cone-beam computed tomography. *Journal of Dentistry*. 2021;114:103786. doi: 10.1016/j.jdent.2021.103786
33. Hung KF, Ai QYH, Wong LM, Yeung AWK, Li DTS, Leung YY. Current Applications of Deep Learning and Radiomics on CT and CBCT for Maxillofacial Diseases. *Diagnostics*. 2022;13(1):110. doi: 10.3390/diagnostics13010110
34. Bolelli F, Lumetti L, Vinayahalingam S, Di Bartolomeo M, Pellacani A, Marchesini K, et al. Segmenting the Inferior Alveolar Canal in CBCT Volumes: The ToothFairy Challenge. *IEEE Trans Med Imaging*. 2025;44(4):1890-1906. doi: 10.1109/tmi.2024.3523096
35. Liu MQ, Xu ZN, Mao WY, Li Y, Zhang XH, Bai HL, et al. Deep learning-based evaluation of the relationship between mandibular third molar and mandibular canal on CBCT. *Clin Oral Invest*. 2022;26(1):981-991. doi: 10.1007/s00784-021-04082-5
36. Buyuk C, Akkaya N, Arsan B, Unsal G, Aksoy S, Orhan K. A Fused Deep Learning Architecture for the Detection of the Relationship between the Mandibular Third Molar and the Mandibular Canal. *Diagnostics*. 2022;12(8):2018. doi: 10.3390/diagnostics12082018
37. Chun SY, Kang YH, Yang S, Kang SR, Lee SJ, Kim JM, et al. Automatic classification of 3D positional relationship between mandibular third molar and inferior alveolar canal using a distance-aware network. *BMC Oral Health*. 2023;23(1):794. doi: 10.1186/s12903-023-03496-9
38. Huang C, Wang Y, Wang Y, Zhao Z. Automatic detection and proximity quantification of inferior alveolar nerve and mandibular third molar on cone-beam computed tomography. *Clin Oral Invest*. 2024;28(12):648. doi: 10.1007/s00784-024-05967-x
39. Arijji Y, Mori M, Fukuda M, Katsumata A, Arijji E. Automatic visualization of the mandibular canal in relation to an impacted mandibular third molar on panoramic radiographs using deep learning segmentation and transfer learning techniques. *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*. 2022;134(6):749-757. doi: 10.1016/j.oooo.2022.05.014
40. Alanazi ARF, Alshammari SAT, Aljudaya SAS, Alshammari DAS, Alazmi FMS, Prabhu N. Automatic Detection and Proximity Quantification of Inferior Alveolar Nerve and Mandibular Third Molar on CBCT: A Systematic Review. *J Pharm Bioallied Sci*. 2025;17(Suppl 3):S2071-S2073. doi: 10.4103/jpbs.jpbs_1109_25

FUNDING

No financing

CONFLICT OF INTEREST

None

AUTHORSHIP CONTRIBUTIONS

Drafting – original draft: Greta Cordovi Vallongo, Youris Samón Mayan.

Writing–review and editing: Greta Cordovi Vallongo, Youris Samón Mayan.