Journal of Dental Sciences 21 (2026) 494—500

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.e-jds.com

Original Article

Evaluation of the performance of an artificial intelligence model in recognizing the habitual mixed language in Taiwan for generating periodontal charting text reports

Yi-June Lo [a†], Yung-Chun Chang [b,c†], Chien-Hung Chen [d], Hsuan-Po Wang [a], Hao-Chen Wang [e], Yoichi Ohiro [f], Chih-Yuan Fang [e,g*]

[a] *Division of Periodontics, Department of Dentistry, Wan Fang Hospital, Taipei Medical University, Taipei, Taiwan*
[b] *Graduate Institute of Data Science, Taipei Medical University, New Taipei City, Taiwan*
[c] *Clinical Big Data Research Center, Taipei Medical University Hospital, Taipei, Taiwan*
[d] *Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei, Taiwan*
[e] *Department of Oral Medicine, Wan Fang Hospital, Taipei Medical University, Taipei, Taiwan*
[f] *Oral and Maxillofacial Surgery, Division of Oral Pathobiological Science, Faculty of Dental Medicine and Graduate School of Dental Medicine, Hokkaido University, Sapporo, Japan*
[g] *School of Dentistry, College of Oral Medicine, Taipei Medical University, Taipei, Taiwan*

Received 8 October 2025; Final revision received 28 October 2025
Available online 11 November 2025

KEYWORDS
Speech recognition;
Periodontal examination;
Localized AI;
ChatGPT;
Smart healthcare

Abstract *Background/purpose:* "Periodontal charting" in Taiwan typically involves a two-person collaboration. However, while advances in voice technology has made it feasible to use voice recognition to reduce these manpower demands in English-speaking countries, these artificial intelligent (AI) voice recognition tools have lacked localized features to make them feasible in our context of Taiwan.
Materials and methods: To fill this research gap, we integrated GPT-4o-transcribe and GPT-4.1-mini into a two-stage system for this study, where stage 1 worked on speech-to-text and stage 2 focused on text-to-report. To evaluate this AI model's performance, five representative periodontal charting scenarios were constructed, and 15 voice recordings were collected. The evaluation focused on transcription accuracy and examined factors influencing recognition quality.

* Corresponding author. Department of Oral Medicine, Wan Fang Hospital, Taipei Medical University, No. 111, Sec. 3, Xinglong Road, Wenshan District, Taipei 116081, Taiwan.
E-mail addresses: 100044@w.tmu.edu.tw, d204106002@tmu.edu.tw (C.-Y. Fang).
† Yi-June Lo and Yung-Chun Chang contributed equally to this article.

https://doi.org/10.1016/j.jds.2025.10.034

Results: The system achieved an overall accuracy of 67.40 % in extracting and formatting data fields related to tooth identification, clinical parameters, and periodontal findings. Furcation involvement, mobility assessment, and missing tooth identification exceeded 90 % accuracy, while bleeding on probing reached 85.76 %. The plaque index and keratinized gingiva width had intermediate accuracies of 68.47 % and 70.83 %, respectively. Probing depth and gingival recession were lower at 53.33 % and 52.01 %. The main factors affecting accuracy included unstable speech speed, prolonged pronunciation leading to repeated numbers, and single positional errors causing chain errors.

Conclusion: Our AI model showed strong speech recognition abilities without requiring extensive training in professional knowledge. It effectively identified dental terms and their related examinations and numbers in mixed language contexts. Future optimizations will focus on addressing errors from uneven speech speed and individual examiner habits.

Introduction

According to a survey from the Ministry of Health and Welfare of Taiwan, periodontal disease is highly prevalent and affects 80 % of adults.[1] In order to monitor the disease progression, regular comprehensive periodontal evaluation, also known as periodontal charting, is recommended.[2,3] However, the current two-person workflow, in which a dentist performs the examination and a recorder enters the data, increases costs, creates bottlenecks, and introduces potential communication errors. During the examination, the dentist cannot continually remove gloves to write or input information, making this collaborative workflow difficult to replace. The shortage of medical support staff and frequent turnover after the pandemic have further exacerbated these challenges. A potential solution is voice technologies, which have shown utility in English-speaking countries. Commercial systems have been reported to have very high recognition accuracy, reduced the need for a dedicated recorder, lowered exposure risk, and also to allow clinicians to focus on patient care.[4–6] Products such as VoiceWorks, Open Dental Software, and Dentrix Voice are used in North America and Europe, but they are designed for English-based workflows and terminology, thus limiting direct adoption in Taiwan.

Taiwanese dentists frequently code switch between Chinese and English, producing equivalent statements mixed forms. For example, Taiwanese dentists might describe the absence of four third molars as "four eight all missing" in Chinese, while English native speakers might say "All four third molars are missing" or "All wisdom teeth are missing". This makes it more difficult for recognition and normalization.[7] Research on speech recognition in Chinese medical terminology, particularly in dental contexts, remains limited, not to mention the challenges of mixed Chinese and English speech recognition.[7,8] Tonal language features, speech speed, individual habits, accents, noisy operatories, and hands-free requirements further complicate performance. To minimize the need for glove removal during examinations, an ideal artificial intelligent (AI) model should not only be able to recognize speech and convert it into text but also understand transitions between examination items, such as pocket depth, bleeding, or the degree of furcation involvement. These challenges highlight the need for Taiwan dental researchers and AI engineers to focus more on developing localized AI speech recognition. Previous studies have shown that the accuracy of speech recognition remains lower than that of humans.[9] Although such systems offer faster reporting and greater cost efficiency, accuracy represents a safety issue rather than a convenience feature in medical field. Therefore, any technological advance must undergo rigorous evaluation before clinical use. To address this gap, we designed a study to analyse the performance of an AI model for mixed Chinese and English recognition. We integrated GPT-4o-transcribe model for multilingual speech-to-text processing with GPT-4.1-mini for text report generation, guided by Taiwan-specific prompt templates. Five designed periodontal charts with 15 voice recordings were used for validation to access the model's accuracy. Through this study, we aim to (1) understand the performance and challenges of AI models in mixed-language recognition in Taiwan, and (2) provide a foundation for future research.

Materials and methods

Overall architecture

We integrated a mature automatic speech recognition engine with a large language model (LLM) to build a voice-driven periodontal charting assistant suitable for Taiwanese dentists. The system was co-designed with dental specialists from needs analysis through prompt engineering and verification, and it followed a two-stage pipeline. Stage 1 converted speech to text, and stage 2 converted text to a standardized periodontal record. Dedicated prompt templates handled the mixed Chinese and English throughout.

Stage 1. Speech-to-text. We employed the GPT-4o-transcribe module, which incorporates Whisper technology, to process multilingual and code-switching input recorded under typical clinical conditions without specialized soundproofing. The system applied signal preprocessing to account for common operatory noise, recognized mixed Chinese—English expressions, and performed light post-processing to reduce errors in dental terminology. Examples of recognized input include "one-seven (Chinese: tooth 17) mesial probing depth (English) four (Chinese) millimeter" and "upper premolar (Chinese) bleeding on probing (English)." These mixed expressions were handled without the need for task-specific ASR retraining.

Stage 2. Text-to-report. For the second stage, we used GPT-4.1-mini to convert transcripts into structured periodontal charting reports. The process relied on prompt-driven templates to ensure consistent terminology mapping, item categorization, numeric extraction, and formatted output. Integrated validation checks monitored semantic coherence, verified numeric values against expected ranges, and enforced logical consistency within the charting structure. These mechanisms collectively improved the accuracy and reliability of the generated reports.

System modules. The system was organized into three modular components connected through APIs: (1) speech recognition, (2) language understanding and format conversion, and (3) data management. The front—end interface adhered to established dental charting conventions, provided progress visualization, enabled real-time transcription, and supported interactive correction. Voice commands such as "start recording the probing depth" and "correct the previous values" allowed hands-free operation during procedures. The data management module ensured record integrity, facilitated export in both JSON and Excel formats, and included a validator that compared the model's Excel output with a reference sheet field by field, thereby incorporating schema and range checks to generate discrepancy reports. Additional functions included conversion of JSON to Excel for clinician review, integration with hospital information systems, PDF export, data backup and restoration, all secured by role-based access control. Finally, the monitoring module recorded key performance metrics, including ASR latency, transcription accuracy, and system response time, to guide iterative optimization and system reliability.

Prompt design

Prompt engineering is central to the performance of this system. In collaboration with senior dentists at Wanfang Hospital and guided by domain-specific materials, we designed task-specific templates tailored to Taiwan's periodontal examination context. The design followed a hierarchical structure consisting of (1) base context settings, (2) professional guidance, (3) task instructions, and (4) output specifications.[10] Base context settings established the workflow and captured common clinical expression patterns to enable the model to align with actual dental practice. Professional guidance encoded periodontal concepts and standards, including the Federation Dentaire Internationale (FDI) tooth numbering system, anatomical terminology, and diagnostic categories to ensure medical accuracy. Task instructions were defined in a modular fashion, each addressing a specific clinical task such as identifying missing teeth, detecting mobility, recording probing depth, recession, bleeding on probing, plaque index, or furcation involvement. This modularization reduced ambiguity and improved task-specific accuracy. Output specifications enforced structured JSON outputs with consistent field names, numeric formats, and abnormality flags, to seamlessly integrate with downstream clinical databases and electronic health record (EHR) systems.

A key principle of the design was localization to clinical context. The prompts were adapted to handle the unique characteristics of Taiwan's bilingual clinical environment, where Chinese—English code-switching, colloquial expressions, and filler words were common. For instance, synonyms and variants were normalized through a post-processing layer that maps Chinese dental terms to their English equivalents. Numerical variations were also addressed, such as colloquial shorthand (e.g., 'upper jaw' referring to the maxilla). Furthermore, a dynamic context mechanism was introduced to infer intent based on the progression of the conversation. For example, when the dentist states 'next' in Chinese, the system correctly infers the subsequent target tooth from the current sequence. From an engineering standpoint, this design emphasized structured, rule-guided interactions with the language model. By encoding examination order (e.g., upper buccal probing from 18 to 28, then palatal probing in reverse order) and enforcing error-handling rules (e.g., prioritizing corrected tooth numbers over initial misstatements), the system mitigated recognition noise inherent in clinical speech. Clinically, this made sure that the extracted data preserved fidelity to the dentist's diagnostic intent while minimizing errors caused by speech rate, filler words, or numeric ambiguities.

Experiment design

A senior periodontist designed five representative periodontal charting scenarios commonly encountered in clinical practice in Taiwan. Fig. 1 illustrates these scenarios and summarizes their underlying concepts and objectives. Recordings were produced by three dentists: one senior periodontist and two junior dentists with differing levels of familiarity with periodontal charting. Junior A was a general dentist with prior exposure to standardized periodontal recording and terminology, while Junior B was less experienced in charting protocols and had limited familiarity with structured periodontal documentation. Each session was conducted in a typical clinical room, with the order of scenarios randomized for each participant. Audio inputs were transcribed using the GPT-4o-transcribe module, and the resulting transcripts were processed by GPT-4.1-mini with a single master prompt to produce structured periodontal charts and brief textual summaries. Ground truth was established directly from the recordings: the senior periodontist prepared a reference chart from the audio and adjudicated all model outputs against this reference standard. In total, 15 reports were generated and included for evaluation.

Validation and performance evaluation

All items generated from the 15 voice records were studied, and the accuracy of each identified examination result was

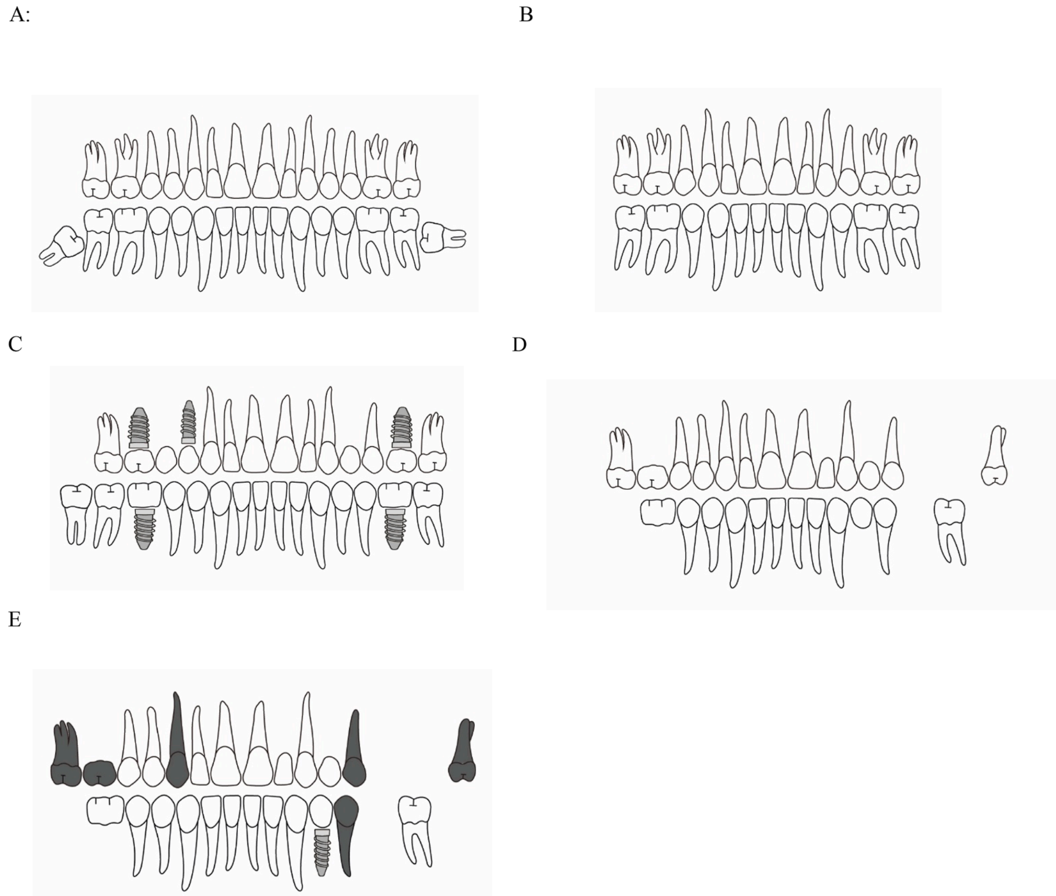

Figure 1 Features and illustrations of the designed periodontal charting. (A) Case 1: Nearly all teeth present, including horizontally impacted teeth 38 and 48, with missing teeth 18 and 28. Probing depths may exceed 10 mm, and we assessed if values of 12 would be misrecognized as 10 and 2 in Chinese. (B) Case 2: Patient had orthodontic treatment with four premolars extracted. All remaining teeth appeared continuous. We evaluated if the AI model could recognize correction commands after initial misidentification of missing teeth. (C) Case 3: Several missing teeth restored with implants or implant-supported bridges. Complexity arose from varied terminology for implants and bridges. (D) Case 4: Multiple missing teeth and aggressive pocket depths (>10 mm). The goal was to see if the AI model could accurately map results to tooth positions amid significant tooth loss. (E) Case 5: Follow-up to Case 4 with additional extractions and new implants. The objective was to evaluate if the AI model could generate error alerts for anomalies like previously missing teeth restored with implants.

assessed by two senior dental experts. The duration of each voice record was also recorded as a reference factor.

Results

The AI voice assistant processed 15 periodontal reports from five clinical scenarios recorded by three dentists of varying experience levels. The system achieved an overall field-level accuracy of 67.40 %, representing correctly extracted and formatted data fields across tooth identification, clinical parameters, and periodontal findings. Table 1 summarizes performance across speakers, clinical scenarios, measurement types, and individual teeth.

Speaker performance showed moderate variation, ranging from 62.60 % to 71.83 %. The senior periodontist achieved the highest accuracy (71.83 %), followed by Junior A (67.73 %) and Junior B (62.60 %). Despite these differences, all three speakers were within 5 % of the mean, indicating relatively stable performance, though experience and systematic dictation offered a modest advantage.

Clinical scenarios exhibited greater variation than speakers. Case-level accuracy ranged from 60.04 % to 75.04 %, with Cases 2 (74.77 %) and 4 (75.04 %) performing best, and Cases 1 (60.04 %) and 3 (61.54 %) performing worst, below the average. Case 5 recorded intermediate accuracy at 65.59 %. This wider spread across scenarios compared to speakers indicated that clinical presentation characteristics such as disease complexity, distribution of affected sites, number of abnormal findings, and frequency of code-switching substantially influenced extraction accuracy, with cases involving extensive periodontal involvement or complex mixed-language expressions posing greater challenges to the recognition and parsing algorithms than simpler clinical presentations.

Measurement type produced the most distinct performance differences. Binary and categorical parameters far outperformed continuous numeric ones. Furcation

Table 1 System accuracy (in %) across speakers, scenarios, measurements, and individual teeth.

Speaker

Speaker identity	Senior	Junior A	Junior B
Accuracy	**71.83**	67.73	*62.60*

Clinical scenario

Case no.	1	2	3	4	5
Accuracy	*60.04*	74.77	61.54	**75.04**	65.59

Measurement

Parameters	Bleeding on probing	Furcation	Keratinized gingiva	Missing	Mobility	Plaque	Probing depth	Recession
Accuracy	85.76	**95.59**	70.83	90.00	95.31	68.47	53.33	*52.01*

Tooth-level

FDI	**18–11**								**21–28**							
	18	17	16	15	14	13	12	11	21	22	23	24	25	26	27	28
Acc.	**88.94**	78.37	69.95	*64.60*	67.23	75.25	73.76	66.09	65.84	67.57	*63.12*	72.82	68.56	66.83	67.07	**88.22**
FDI	**48–41**								**31–38**							
	48	47	46	45	44	43	42	41	31	32	33	34	35	36	37	38
Acc.	**83.01**	61.17	58.25	*51.49*	59.16	57.67	57.67	59.41	61.63	62.13	*51.73*	63.37	59.16	66.50	66.99	**91.02**
Overall accuracy: 67.40																

Note 1. FDI denotes to the abbreviation of Fédération Dentaire Internationale numbering system. Acc. denotes to the abbreviation of accuracy.
Note 2. The highest value in each field is written by bold letter, while the lowest value is in italics.

involvement achieved 95.59 %, mobility 95.31 %, missing tooth identification 90.00 %, and bleeding on probing 85.76 %, all well above the system mean. Plaque index and keratinized gingiva width performed near the mean at 68.47 % and 70.83 %. In contrast, probing depth (53.33 %) and gingival recession (52.01 %) were two fundamental measurements that underpinned periodontal diagnosis and treatment planning, yet both failed to exceed 55 % accuracy. This deficit stemmed from the sequential nature of numeric dictation. Each examination requires the continuous verbal reporting of long strings of numbers across multiple tooth sites. Once a single recognition error occurred, it propagated through subsequent values, generating cumulative drift and distorting entire tooth-level data entries. Rapid or unclear speech often compounded these errors and caused merged or repeated digits. In contrast, categorical parameters depend on discrete, contextually constrained terms that are less vulnerable to cascading recognition failures. This pronounced categorical and numeric disparity underscores an inherent performance limitation tied to the sequential dependency of numeric transcription.

Individual tooth-level accuracy showed substantial variation and clear anatomical patterns. Performance ranged from 51.49 % (tooth 45) to 91.02 % (tooth 38), reflecting considerable heterogeneity across the dentition. Third molars in all quadrants achieved the highest accuracies, with teeth 18 (88.94 %), 28 (88.22 %), 38 (91.02 %), and 48 (83.01 %), all exceeding the overall mean, and tooth 38 ranking highest overall. Maxillary anterior teeth performed moderately well, with teeth 11 through 13 recording 66.09 %, 73.76 %, and 75.25 %, while maxillary premolars and first molars ranged from 64.60 % to 78.37 %, generally near or above the system average. In contrast, mandibular anterior teeth and premolars consistently underperformed: teeth 42 (57.67 %), 43 (57.67 %), 44 (59.16 %), and 45 (51.49 %) all fell below the mean, with tooth 33 (51.73 %) showing similarly low accuracy. Mandibular molars showed mixed performance, with teeth 36 (66.50 %) and 37 (66.99 %) outperforming 31 (61.63 %) and 32 (62.13 %). The clustering of high-performing third molars and low-performing mandibular anterior-premolar regions highlights systematic anatomical influences on recognition accuracy.

Overall, measurement complexity was the strongest predictor of accuracy, with categorical parameters outperforming continuous numeric inputs. Clinical scenario complexity exerted more influence than speaker experience, and anatomical location followed consistent trends favoring posterior teeth and third molars. Although the 67.40 % overall accuracy remained a stable central tendency, substantial variability persisted across data types and clinical contexts.

Discussion

Advancements in AI and technology have significantly impacted both daily life and medical applications.[11] In dentistry, these technologies have reshaped how professionals record and process information. With the progress of speech recognition, users increasingly prefer speaking to devices for text input as it improves efficiency and reduces the inconvenience of small screens. ChatGPT, a widely recognized AI model, has demonstrated high performance in medical contexts despite lacking domain-specific training. For instance, its accuracy in answering medical questions comes close to that of clinical professionals,[12—14] and its diagnostic accuracy in dentistry can reach up to 80 %.[15] These findings underscore AI's potential to reduce manpower and improve efficiency in healthcare.

To support localized voice-recognition development for periodontal examinations in Taiwan, we proposed and evaluated a ChatGPT-based model combining GPT-4o-transcribe and GPT-4.1-mini. The former is optimized for precise speech-to-text transcription, while the latter balances performance and computational efficiency. However, despite these advantages, our experimental accuracy remains insufficient for clinical adoption. Analysis of 15 simulated reports revealed three primary challenges that limit performance. First, linguistic and numeric variability caused significant recognition errors. While simple numeric entries were handled well, performance dropped sharply in mixed Chinese—English phrases or shorthand expressions. Ambiguities such as "0000 ..." or "all are zeros" led to inconsistent outputs, suggesting the need for better disambiguation and normalization mechanisms. Second, speech speed and sequence dependency affected recognition stability. Both rapid and prolonged speech increased error rates, often triggering cascading misrecognitions, such as interpreting a pocket depth of 323 as 33333323. This highlights the fragility of sequence-based recognition and the need for error-tolerant strategies like dynamic segmentation or confidence-based re-prompting. Third, semantic ambiguity in short phrases or noun expressions also reduced accuracy, indicating that contextual awareness is essential for reliable transcription.

Since commercially available single-language speech recognition software claims an accuracy rate of 90 %—99 %, we believe that the multilingual AI voice recognition model must achieve an accuracy rate of at least 90 % before applied in real world. However, a more ideal evaluation metric would involve multiple groups of dentists working alongside experienced dental assistants to measure the accuracy of human collaboration in identical simulated scenarios. This data can then serve as a benchmark for assessing the accuracy of the AI model. Furthermore, the accuracy of human collaboration should be provided as a reference point, and a larger-scale clinical survey of dentists should be conducted to better understand their expectations regarding the accuracy of AI voice assistants. Except for the accuracy not meeting clinical ideals, this study had several limitations that should be considered when interpreting the findings. From a technical perspective, the proposed system was built by integrating general-purpose models (Whisper, GPT-4o-transcribe, and GPT-4.1-mini) rather than models fine-tuned specifically on dental corpora. While this approach demonstrates the feasibility of leveraging widely available tools, recognition accuracy was more variable in complex contexts, particularly for code-switching expressions and continuous numeric measurements such as probing depth and gingival recession. In addition, the current system has not yet been fully adapted to dialectal variations or broader clinical speech patterns, which may influence its performance in more diverse linguistic settings. From a clinical validation perspective, the evaluation was based on five designed periodontal charting scenarios and 15 voice recordings in a controlled environment. Although this design allowed us to systematically examine error patterns, it may not fully capture the complexity, background noise, or time constraints of routine dental practice. Moreover, the recordings were provided by three dentists from a single institution, which limits the representativeness of the dataset. Finally, our study didn't assess workflow integration or user acceptance in clinical settings.

Our findings emphasize the importance of localization in clinical AI applications. Dental speech is characterized by bilingual mixing, rapid tempo, and shorthand communication adapted to local workflows. Without explicit adaptation, AI systems risk misinterpreting diagnostic content. Although our dataset reflected the speech patterns of Taiwanese dentists, similar issues are expected in other bilingual environments, thus making these results broadly applicable. Future work should explore hybrid recognition frameworks combining statistical and rule-based correction, expand bilingual training corpora, and incorporate contextual knowledge of dental workflows. By addressing these identified challenges, this study provides practical insights toward achieving reliable LLM-assisted periodontal charting in clinical settings.

Declaration of competing interests

The authors have no conflicts of interest relevant to this article.

Acknowledgements

This research was supported by Taipei Medical University and Wan Fang Hospital, under the grant number 114TMU-WFH-21.

References

1. Cheng FC, Wang YL, Chiang CP. The dental use for periodontal diseases under the national health insurance system in Taiwan in 2021. *J Dent Sci* 2023;18:1251—7.
2. Greenwell H. Committee on research, science and therapy. American academy of periodontology. Position paper: guidelines for periodontal therapy. *J Periodontol* 2001;72:1624—8.
3. Huang KC, Lai CH, Huang CF, Lu HK. A comprehensive periodontal treatment project: the periodontal status, compliance rates, and risk factors. *J Dent Sci* 2016;11:182—8.
4. Lister P, Sudharson NA, Kaur P, Joseph M. Voice-activated charting systems. *Br Dent J* 2024;236:734.
5. Drevenstedt GL, McDonald JC, Drevenstedt LW. The role of voice-activated technology in today's dental practice. *J Am Dent Assoc* 2005;136:157—61.
6. Holmes RD, Burford B, Vance G. Development and retention of the dental workforce: findings from a regional workforce survey and symposium in England. *BMC Health Serv Res* 2020;20:255.
7. Li Q, Mai Q, Wang M, Ma M. Chinese dialect speech recognition: a comprehensive survey. *Artif Intell Rev* 2024;57:25.
8. Lee MC, Yeh SC, Chang JW, Chen ZY. Research on Chinese speech emotion recognition based on deep neural network and acoustic features. *Sensors* 2022;22:4744.
9. Johnson M, Lapkin S, Long V, et al. A systematic review of speech recognition technology in health care. *BMC Med Inf Decis Making* 2014;14:94.

10. Giray L. Prompt engineering with ChatGPT: a guide for academic writers. *Ann Biomed Eng* 2023;51:2629—33.
11. Fang EP, Liew DJ, Chang YC, Fang CY. Enhancing wisdom teeth detection in panoramic radiographs using multi-channel convolutional neural network with clinical knowledge. *Comput Biol Med* 2025;192:110368.
12. Sabry AMM, Kamel BMN. ChatGPT in clinical toxicology. *JMIR Med Educ* 2023;9:e46876.
13. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. *J Med Internet Res* 2023;25:e48568.
14. Wu YH, Tso KY, Chiang CP. Impact of language and question types on ChatGPT-4o's performance in answering oral pathology questions from Taiwan national dental licensing examinations. *J Dent Sci* 2025;20:2176—80.
15. Danesh A, Danesh A, Danesh F. Innovating dental diagnostics: ChatGPT's accuracy on diagnostic challenges. *Oral Dis* 2025;31:911—7.