AI-Driven Multimodal TMJ Patient Modeling: From Unstructured Notes to Precision Treatment

Alban Gaydamour[1,2], Enzo Tulissi[1,2], Claudia Mattos[3], Rodrigo Teixeira[4], Maxwell Shin[1], Adam Hershey[1], Anabelle Kwon[1], Felicia Miranda[4], Marcela Gurgel[5], Selene Barone[6], Aron Aliaga[1], Marilia Yatabe[1], Paulo Zupelari[1], Marina Zupelari[1], David Hanauer[1], Nina Hsu[1], Steve Pieper[7], Eduardo Caleme[8], Jonas Bianchi[9], Joao Goncalves[10], Daniela Goncalves[10], Lawrence Wolford[11], Antonio Ruellas[12], Juan Prieto[13], Tengfei Li[13], Hongtu Zhu[13], Runpeng Dai[13], Martin Styner[13], Najla Al Turkestani[14], Alexandre F. DaSilva[1], and Lucia Cevidanes[1]

[1] University of Michigan, Ann Arbor MI, USA
[2] CPE Lyon, Lyon, France
[3] Fluminense Federal University, Rio de Janeiro, Brazil
[4] University of São Paulo, São Paulo, Brazil
[5] Federal University of Ceara, Fortaleza, Brazil
[6] Magna Græcia University, Catanzaro, Italy
[7] Isomics Inc., Cambridge MA, USA
[8] Positive University, Curitiba, Brazil
[9] University of the Pacific, San Francisco CA, USA
[10] State University of São Paulo, São Paulo, Brazil
[11] Baylor University Medical Center, Dallas TX, USA
[12] Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
[13] University of North Carolina, Chapel Hill NC, USA
[14] King AbdulAziz University, Jeddah, Saudi Arabia

Abstract. Temporomandibular degenerative joint disease (TM DJD) is a multifactorial condition with complex clinical presentations. This study presents a multimodal framework centered on structured summarization of clinical text, supported by imaging information from automatically registered MRI and CBCT scans. Two large language models, BART and DeepSeek-R1, were fine-tuned on 1,813 annotated text segments from 500 TM DJD patient records to extract 56 clinical indicators, including pain severity, jaw function, imaging findings, and sleep disturbances. The models converted narrative notes into structured data fields for use in clinical dashboards enabling patient-specific and population-level analyses. BART outperformed DeepSeek in clinical field extraction accuracy, precision, and recall, despite DeepSeek achieving slightly higher ROUGE metrics based on word-level overlap. A parallel automated MRI-to-CBCT registration pipeline achieved submillimeter accuracy and a 98.75% success rate. This work extracted clinically meaningful pain comorbidities and radiological findings from unstructured clinical narratives, enabling actionable insights for musculoskeletal precision care. The future integration of structured clinical data and multimodal image analyses may enable holistic, personalized patient models.

Keywords: Temporomandibular degenerative joint disease · Multimodal imaging · Large Language Models.

1 Introduction

Temporomandibular disorders (TMD) are the second most prevalent musculoskeletal condition after chronic low back pain, affecting 5–12% of the population and costing an estimated $4 billion annually [1]. Around 80% of patients exhibit clinical signs such as disc displacement and joint pain often progressing to temporomandibular degenerative joint disease (TM DJD) [1–3]. Comorbidities like headaches and sleep disturbances are common and significantly influence disease progression and treatment outcomes [4–6]. While MRI and CBCT are essential for visualizing joint anatomy, they do not capture a patient's symptom burden or comorbidity profile, factors critical for personalized treatment. Clinical decision-making requires understanding both structural changes and functional impact, such as sleep disruption, jaw locking, or radiating pain. These details, often documented only in free-text notes, were captured at scale using MedEx, a tool developed with two fine-tuned large language models (LLMs): Bidirectional Auto-Regressive Transformer (BART) [10] and DeepSeek-R1. Both models were downloaded and used offline to ensure patient data privacy, with no clinical information transmitted over the internet. Trained on 500 TMD notes processed into 1813 annotated segments, both models extract structured data from unstructured narratives [7–9], mapping each note to 56 predefined clinical fields (e.g., "maximum opening: 38 mm"; "disc displacement: with or without reduction"). This work provides a structured textual layer containing disease-related pain comorbidities and imaging findings, along with MRI-to-CBCT registered images [11]. These datasets support comprehensive TMJ evaluation and are aligned with initiatives such as the TMD IMPACT Consortium [12] and the NIH HEAL Initiative [13].

2 Methods

2.1 Study Sample

This study focuses on the structured summarization of clinical notes for patients with TM DJD. A retrospective dataset of de-identified clinical notes was compiled from 500 patients treated at the Baylor University Department of Oral and Maxillofacial Surgery. The dataset includes initial clinical examination notes and radiology reports for CBCTs and MRIs, capturing pain-related comorbidities and diagnostic impressions at baseline diagnosis. Patients were included only if MRI and CBCT imaging had been performed within a one-year interval to ensure temporal alignment between modalities. The imaging pipeline incorporated a previously validated MRI-to-CBCT registration method [11]. This method aligns MRI with CBCT scans using rigid mutual information–based registration, achieving submillimeter translation errors and rotation differences under 3 degrees. Follow-up visits and patients under 12 years of age were excluded. This study was approved as exempt by the University of Michigan Institutional Review Board (HUM00239207).

2.2 Data Preprocessing and Annotation

Clinical notes (PDF/DOCX) were converted to plain text via a custom Python script. To meet large language model input constraints, each document was segmented into discrete, non-overlapping text segments, with a maximum length of 1024 tokens. Text segmentation to 1024 tokens was required due to input size constraints of the BART model. While DeepSeek supports larger context windows, identical chunk sizes were used across models to ensure fair comparison and maintain consistency. Segmentation was designed to preserve semantic integrity, aggregating full paragraphs and breaking only at sentence boundaries when necessary. Annotation guidelines were defined through a clinician-led calibration process to ensure consistent labeling. These guidelines were then applied to annotate all 1813 segments using a predefined set of 56 descriptors related to TM DJD, including pain, sleep disturbances, hearing loss, and jaw function. Only terms with explicit textual evidence were included in the final dataset, which was then used to train and evaluate multiple summarization models. This textual layer complements the anatomical context captured through MRI-CBCT registration, without requiring additional image-based annotation.

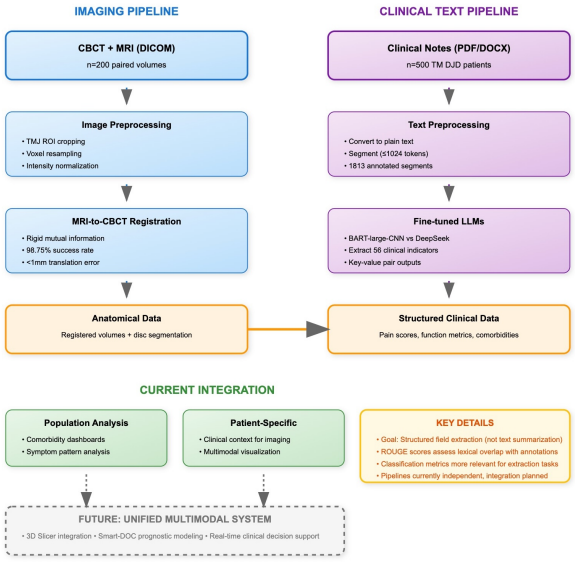

Fig. 1. Overview of the multimodal TMJ analysis workflow. The imaging pipeline (left) processes CBCT and MRI volumes through preprocessing and registration, enabling disc segmentation; over 200 patients have been processed to date. The clinical text pipeline (right) extracts 56 structured diagnostic fields from unstructured notes using fine-tuned LLMs, applied to 500 patients. Unlike traditional summarization, the models perform structured field extraction. Current outputs support both population-level comorbidity analysis and patient-specific interpretation. ROUGE scores assess lexical similarity, while classification metrics reflect clinical extraction accuracy. Future work will unify these tools within 3D Slicer for real-time diagnostic support.

2.3 Model Architectures and Training Details

Two large language models were selected to evaluate clinical summarization performance: BART-large-CNN [14], a summarization-specific encoder–decoder model, and DeepSeek-R1-Distill-Qwen-1.5B, a distilled general-purpose language model. It is important to note that although BART is a summarization-specific model, the training objective was not to generate shorter narrative summaries, but rather to produce structured outputs (key–value pairs). These models were chosen for their wide adoption and architectural diversity, enabling comparison between task-specific and general-purpose modeling approaches. BART was obtained from the Hugging Face Transformers library and fine-tuned using its default summarization configuration. DeepSeek was fine-tuned using Low-Rank Adaptation (LoRA) and 4-bit quantization. LoRA adapters were applied to key attention and MLP components to enable adaptation of structural and content representations for summarization. Both models were trained on the same dataset of annotated clinical text segments. Training was conducted using cosine learning rate scheduling, gradient accumulation over 4 steps, and without any prompt engineering or instruction tuning. Model outputs were evaluated using Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metrics to assess summarization quality [15].

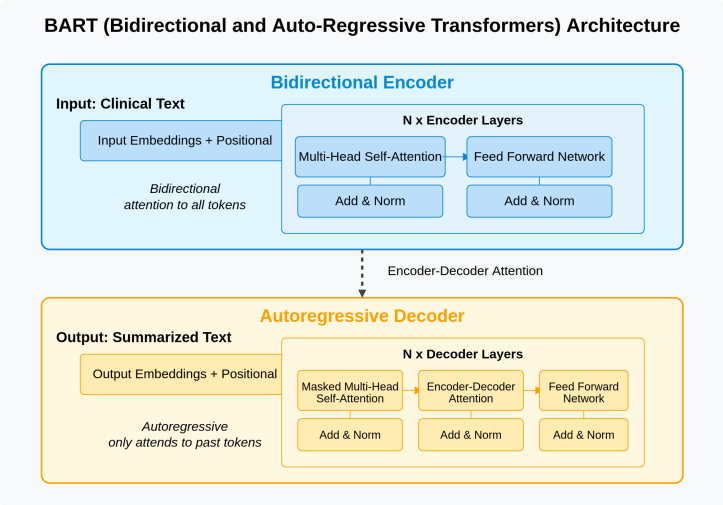

Fig. 2. Architecture of the BART LLM used for fine-tuning.

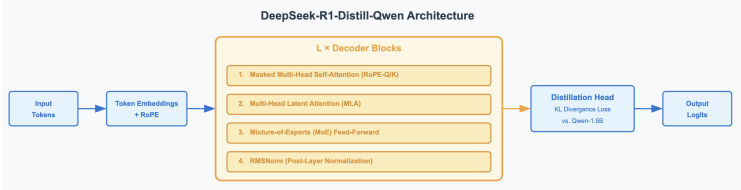

Fig. 3. Architecture of the DeepSeek LLM used for fine-tuning.

2.4 Training and Validation Protocol

The dataset was divided using an 80:10:10 split (80% of samples for training, 10% for validation, and 10% for testing). A 5-fold cross-validation protocol was employed to mitigate overfitting and assess generalizability. Both BART and DeepSeek models were evaluated across all folds under identical conditions. Training was monitored via ROUGE scores, and the model with the best ROUGE-1 score across folds was designated as the final summarization model.

2.5 Performance Evaluation

Model performance was evaluated using two categories of metrics: ROUGE-based lexical similarity and classification metrics for clinical field extraction. ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum were used to assess overlap between predicted and reference summaries. Additionally, accuracy, precision, recall, and F1 scores were computed for ten representative clinical categories, including headache intensity, jaw function, and disability rating. These values were derived using structured post-processed outputs and corresponding clinician annotations. Particular attention was given to comorbidity indicators with variable expression patterns, which tend to challenge LLMs in medical documentation. Both the BART and DeepSeek models were evaluated under this framework. The extracted summaries were formatted as key–value pairs (e.g., `maximum opening: 48mm`, `daily pain: 6`), enabling alignment with image-based dashboards and structured visualization tools.

3 Results

The MRI-to-CBCT registration framework previously validated on 70 paired volumes achieved a 98.75% success rate, with mean translation error below 1 mm and rotational deviation under 3° [11]. Figure 4 and Table 1 demonstrate this high-accuracy pipeline that provides a reliable anatomical foundation for multimodal TMJ analysis.

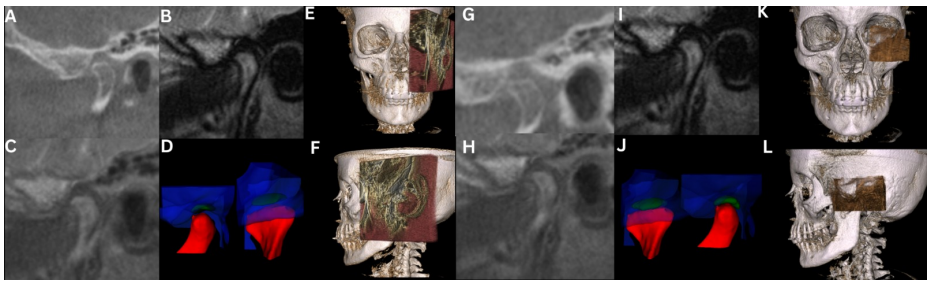

Fig. 4. MRI-CBCT registration comparison: A,G) Fixed CBCT, B,I) MRI manual/automated registration, C,H) MRI-CBCT overlays, D,J) TMJ segmentation (mandible: red, cranial base: blue, disc: green), E,K) Frontal views, F,L) Lateral views.

Table 1. Differences in six Degrees of Freedom between clinician and Elastix Image Registration.

ROTATION (°)			
	Pitch	**Roll**	**Yaw**
Mean difference (SD)	0.49 (2.28)	-0.62 (2.21)	-0.99 (3.84)
Mean absolute difference (SD)	1.53 (1.75)	1.69 (1.54)	2.70 (2.89)
Minimum absolute difference	0.00	0.02	0.03
Maximum absolute difference	8.72	6.79	16.60
75th percentile of absolute difference	1.93	2.56	3.04
90th percentile of absolute difference	3.51	3.32	5.10
TRANSLATION (mm)			
	LR	**AP**	**SI**
Mean difference (SD)	0.3 (1.86)	0.89 (0.94)	-0.20 (0.63)
Mean absolute difference (SD)	0.92 (1.64)	0.98 (0.85)	0.50 (0.43)
Minimum absolute difference	0.01	0.01	0.00
Maximum absolute difference	13.10	3.52	2.57
75th percentile of absolute difference	1.16	1.48	0.72
90th percentile of absolute difference	1.55	2.12	0.95

Tables 2 and 3 report ROUGE summarization scores under 5-fold cross-validation for the BART and DeepSeek models, respectively. DeepSeek slightly outperformed BART in ROUGE scores (e.g., ROUGE-L +1.5%), indicating a minor advantage in surface-level fluency and lexical overlap. However, ROUGE metrics primarily assess general summarization quality and do not directly reflect field-level extraction accuracy required for clinical deployment.

Table 2. Summarization performance (ROUGE scores) of fine-tuned BART across 5-fold cross-validation.

	ROUGE-1	**ROUGE-2**	**ROUGE-L**	**ROUGE-Lsum**
Fold 1	83.68	71.99	83.50	83.49
Fold 2	83.48	73.40	83.11	83.14
Fold 3	84.93	74.23	84.38	84.57
Fold 4	85.50	74.73	85.11	85.21
Fold 5	85.47	74.64	84.98	85.01
Average	84.61	73.80	84.22	84.29

Table 3. Summarization performance (ROUGE scores) of fine-tuned DeepSeek across 5-fold cross-validation.

	ROUGE-1	**ROUGE-2**	**ROUGE-L**	**ROUGE-Lsum**
Fold 1	86.55	86.49	86.53	86.54
Fold 2	84.90	84.79	84.86	84.82
Fold 3	86.08	86.09	86.10	86.11
Fold 4	85.96	85.91	85.95	85.92
Fold 5	85.21	85.19	85.17	85.21
Average	85.74	85.70	85.72	85.72

Tables 4 and 5 present the classification performance of both models on 10 comorbidity-related clinical fields. Despite lower ROUGE scores, BART achieved substantially higher accuracy, precision, recall, and F1 scores across most categories. BART's F1 scores exceeded 70% in structured fields such as patient age, airway obstruction, and arthritis location. In contrast, DeepSeek performance

varied more widely and was notably lower on low-prevalence items such as tinnitus and disability rating.

Table 4. Fine-tuned BART Model Performance for Best Fold

	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
Patient Age	78.6	88.0	88.0	88.0
Airway Obstruction	65.3	78.1	80.0	79.0
Arthritis Location	62.5	72.3	81.4	76.9
Headache Intensity	54.4	77.5	64.6	70.4
Fibromyalgia Present	50.0	100.0	50.0	66.7

Table 5. Fine-tuned DeepSeek Model Performance for Best Fold

	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
Jaw Function	51.1	84.4	61.4	71.1
Arthritis Location	49.2	64.0	68.1	66.0
Earache	35.7	72.3	60.6	52.6
Airway Obstruction	29.9	40.0	54.1	46.0
Muscle Tenderness	25.4	32.6	53.6	40.6

To illustrate the practical value of the extracted information, the manually curated clinical summaries for the full 500-patient cohort were compiled and analyzed. Structured key–value pairs—such as patient age, headache intensity, and headache location—were aggregated across the dataset. Descriptive statistics including prevalence, mean scores, and standard deviations were computed for relevant fields. This analysis revealed consistent TM DJD comorbidity patterns, including high prevalence of sleep disturbances, lateralized jaw pain, and variation in headache intensity. Figure 5 presents a visual dashboard generated from the manually extracted summaries, offering insight into population-level symptom trends and TMJ function characteristics.

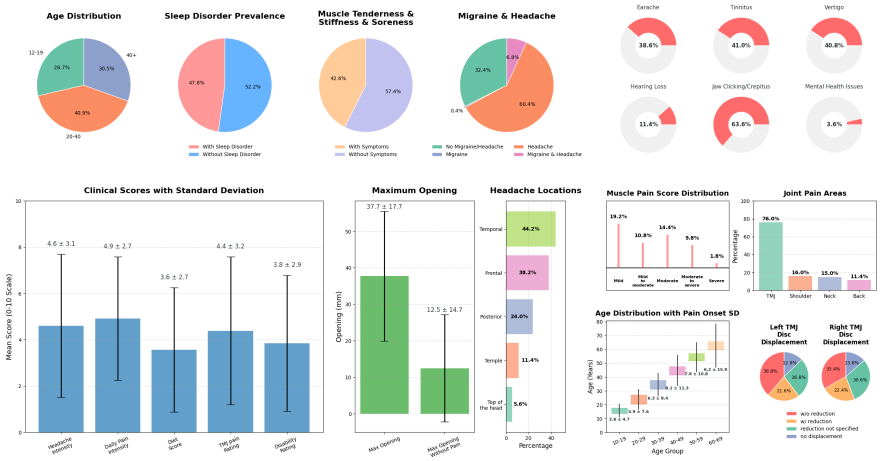

Fig. 5. Population-level dashboard of TMJ comorbidities (e.g., pain severity, jaw function) extracted from 500 clinical notes.

4 Discussion

This study presents a framework that combines an automated MRI-to-CBCT registration pipeline with structured clinical note summarization to support comprehensive assessment of TM DJD. Building on prior work [11], the imaging pipeline now automates previously manual steps such as initial approximation and TMJ cropping, thereby reducing inter-observer variability and enhancing reproducibility. Segmentations were first generated using an automated AI-based approach, and then manually reviewed and corrected by trained clinicians to ensure high-quality anatomical accuracy prior to registration. The workflow is implemented as an open-source 3D Slicer module, facilitating both clinical and research use. Validated on paired MRI-CBCT volumes, the registration achieved a 98.75% success rate with submillimeter translation and sub–3° rotational error, yielding anatomically coherent 3D representations of the TMJ [11].

This work focuses on the automated extraction of structured information from clinical and imaging notes. A dataset of 500 de-identified notes from TM DJD patients was used to fine-tune two LLMs: BART and DeepSeek-R1. While DeepSeek achieved marginally higher ROUGE scores, BART consistently outperformed on field-level metrics including accuracy, precision, and recall. These findings are consistent with prior research showing that BART can outperform general-purpose models by over 15% in ROUGE-L for EHR summarization [10], and that smaller, domain-adapted models provide more reliable performance in precision-sensitive clinical NLP tasks [8].

A key strength of this framework lies in MedEx's ability to navigate diverse documentation styles. Given the absence of standardized questionnaires across TMD centers, clinical heterogeneity often hinders comparability. MedEx's structured outputs help normalize unstructured documentation and enable aggregation of pain, function, and sleep metrics across a 500-patient dataset. This structured approach enables direct correlation between functional limitations and anatomical findings. Dashboards constructed from these summaries highlighted trends in headache intensity, lateralized joint pain, and functional limitation [16], reflecting the model's robustness in capturing fragmented or variable text. The structured summaries also support patient-specific visualization and analysis. Each individual's comorbidities, such as joint arthritis, headache location, and airway obstruction, are aligned with their corresponding MRI-CBCT imaging findings, enabling a unified, subject-level diagnostic view.

Several areas of improvement remain. Comparison with a naive baseline using EMERSE, a text-mining tool for clinical notes, is ongoing to quantify MedEx's added value. Additionally, although BART and DeepSeek were selected for task-specific and general-purpose comparison, future work will assess zero-shot LLMs like GPT-4 and Gemini using PHI-safe platforms. These models may enable few-shot prompting or direct deployment in resource-constrained environments. Planned development will align the structured outputs from LLMs with 3D Slicer dashboards that integrate MRI-CBCT visualizations, where each patient's extracted comorbidities will be displayed alongside their anatomical imaging data, enabling personalized, multimodal planning.

Another challenge lies in the model's sensitivity to variations in note structure and terminology, which can limit generalizability. Errors may stem from underrepresented terms or occasional hallucinations [17]. Planned data augmentation includes structural and lexical modifications—such as synonym replacement, noise injection, and domain-specific paraphrasing—drawing from denoising pretraining strategies [18]. These techniques can enhance robustness while preserving clinical validity. Future work will explore longer-context models (e.g., DeepSeek's full context window, GPT-4) to evaluate whether document-level coherence improves field extraction, especially for cross-sentence inferences.

Although ROUGE and classification scores offer valuable benchmarks for performance, clinical deployment will require human-in-the-loop evaluation and broader generalization across patient populations. By aligning imaging data with structured summaries extracted from clinical notes, this work establishes a scalable foundation for TMJ analysis that links radiologic context with diagnostic information captured in text, enhancing the utility of both data sources for clinical decision-making. Future directions include automated segmentation of the articular disc from MRI, comparisons with rule-based systems such as EMERSE, and evaluation of zero-shot LLMs for clinical deployment. Outputs such as CSV files and dashboards facilitate compatibility with clinical workflows and integration into open-source visualization tools [19].

5 Conclusion

This study introduces a multimodal framework for TMJ assessment that combines automated MRI-to-CBCT registration with structured clinical summarization using fine-tuned language models. Summarization models extract comorbidity and imaging related indicators from clinical notes, with BART outperforming DeepSeek in structured output accuracy. The resulting structured datasets and visual dashboards reveal clinically relevant patterns in pain, function, and sleep disturbances, supporting population-level analysis and providing integrated visualization of patient-specific diagnoses.

Acknowledgments. This work was funded by NIH, grant number R01-DE024450.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Gatchel, R.J., Stowell, A.W., Wildenstein, L., Riggs, R., Ellis, E.: Efficacy of an early intervention for patients with acute temporomandibular disorder-related pain: a one-year outcome study. J. Am. Dent. Assoc. **137**(3), 339–347 (2006)
2. Plesh, O., Sinisi, S.E., Crawford, P.B., Gansky, S.A.: Diagnoses based on the research diagnostic criteria for temporomandibular disorders in a biracial population of young women. J. Orofac. Pain **19**, 65–75 (2005)

3. Manfredini, D., Segu, M., Bertacci, A., Binotti, G., Bosco, M.: Diagnosis of temporomandibular disorders according to RDC/TMD axis I findings, a multicenter Italian study. Minerva Stomatol. **53**, 429–438 (2004)
4. Pihl, K., Roos, E.M., Taylor, R.S., Grønne, D.T., Skou, S.T.: Associations between comorbidities and immediate and one-year outcomes following supervised exercise therapy and patient education - A cohort study of 24,513 individuals with knee or hip osteoarthritis. Osteoarthritis Cartilage **29**(1), 39–49 (2021)
5. Al Turkestani, N., Li, T., Bianchi, J., et al.: A comprehensive patient-specific prediction model for temporomandibular joint osteoarthritis progression. Proc. Natl. Acad. Sci. U.S.A. **121**(8), e2306132121 (2024)
6. Scattergood, S.D., Cheng, V., Wylde, V., Blom, A.W., Whitehouse, M.R., Lenguerrand, E.: Influence of pre-operative co-morbidities on pain and function outcomes at 1 year after primary total knee arthroplasty. Knee **54**, 263–274 (2025)
7. Schiffman, E., Ohrbach, R., Truelove, E., et al.: Diagnostic criteria for temporomandibular disorders (DC/TMD) for clinical and research applications: recommendations of the International RDC/TMD Consortium Network and Orofacial Pain Special Interest Group. J. Oral Facial Pain Headache **28**(1), 6–27 (2014)
8. Ge, J., Li, M., Delk, M.B., Lai, J.C.: A comparison of a large language model vs manual chart review for the extraction of data elements from the electronic health record. Gastroenterology **166**(4), 707–709 (2024)
9. Hsu, E., Roberts, K.: LLM-IE: a python package for biomedical generative information extraction with large language models. JAMIA Open **8**(2), ooaf012 (2025)
10. Zhou, F., Qin, B., Lan, G., Ye, Z.: News text generation method integrating pointer-generator network with bidirectional auto-regressive transformer. In: 2023 2nd International Conference on Artificial Intelligence and Intelligent Information Processing (AIIIP). IEEE, 114–118 (2023)
11. Leroux, G., Mattos, C., Claret, J., et Al.: Novel CBCT-MRI registration approach for enhanced analysis of temporomandibular degenerative joint disease. In: Clinical Image-Based Procedures. CLIP 2024. Lecture Notes in Computer Science, vol. **15196**, pp. 63–72. Springer, Cham (2024).
12. National Institute of Dental and Craniofacial Research: TMD Collaborative for Improving Patient-Centered Translational Research (TMD IMPACT), https://www.nidcr.nih.gov/grants-funding/research-funded-by-nidcr-extramural/tmd-impact last accessed 2025/03/18
13. NIH Heal Initiative, https://heal.nih.gov/ last accessed 2025/03/18
14. Lewis, M., Liu, Y., Goyal, N., et al.: BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 7871–7880 (2020)
15. Lin, C.Y.: ROUGE: a package for automatic evaluation of summaries. In: Text Summarization Branches Out, 74–81 (2004)
16. Dowding, D., Randell, R., Gardner, P., et al.: Dashboards for improving patient care: review of the literature. Int. J. Med. Inform. **84**(2), 87–100 (2015)
17. Huang, L., Yu, W., Ma, W., et al.: A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. ACM Trans. Inf. Syst. **43**(2), 1–55 (2025)
18. Chen, X., Long, G., et al.: Improving the robustness of summarization systems with dual augmentation. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 6846–6857 (2023)
19. Beam, A.L., Kohane, I.S.: Big Data and Machine Learning in Health Care. JAMA **319**(13), 1317–1318 (2018)