Imputation of Missing ECG Data with Implicit Neural Representations

Degian Ghebermicael, Praneeth Sadda, Uzodinma Emerenini

Background:

Electrocardiograms (ECGs) are a frequently ordered diagnostic test. In recent years, there has been increased interest in digitizing old ECGs for archival purposes and for training AI algorithms. However, the digitization process can be challenging due to issues such as faded ink, handwritten annotations, and stamped labels that obscure parts of the printed ECG waveform, leading to gaps in the digitized signals.

Neural networks have been used to impute these missing ECG fragments [1]. However, this generally requires a training dataset of tens to hundreds of thousands of ECGs [1]. In contrast, Implicit Neural Representations (INRs) can be trained on a single example by treating each location within the example as a separate datapoint [2]. In this study, we show that INRs can be used to impute missing fragments in an ECG and show that they are competitive with a traditional model trained on over 10,000 ECGs.

Methods:

This study was conducted on the PTB-XL database [3], a publicly available collection of 21,837 twelve lead ECGs.

Each ECG in the database was corrupted by randomly removing up to 30% of the signal. Two neural network models were tasked with reconstructing the original signal from the corrupted signal:

1. A U-Net was trained on 10,918 (50%) of the ECGs in the database and was evaluated on a separate subset of 1,000 ECGs.
2. An INR was evaluated on a subset of 1,000 ECGs from the database. Since an INR is designed to learn a representation of a single item, a fresh INR was trained for each ECG. The INR was trained on the known values of the ECG and evaluated on the unknown values.

The models were evaluated in terms of mean absolute error (MAE) between the original ECGs and their reconstructions.

Results:

The U-Net reconstructed ECGs with a MAE of 0.0610 mV (standard deviation 0.5807). The INR reconstructed ECGs with a MAE of 0.0535 mV (standard deviation 0.0233). This represented a relative error reduction of 12.3%. The mean difference in error was statistically significant by the Mann-Whitney U Test ($p < 0.01$).

Conclusion:

INRs have can impute missing fragments of ECG signals with greater accuracy than a U-Net trained on over 10,000 ECGs. The main limitation is slower speed of imputation, since the INR must be retrained for each new ECG encountered. However, in situations where training data is scarce, INRs can be a useful tool for ECG analysis.

References:

[1] Kazijevs, M., & Samad, M. D. (2023). Deep Imputation of Missing Values in Time Series Health Data: A review with benchmarking. *Journal of Biomedical Informatics*, *144*, 104440.

[2] Sitzmann, V., Martel, J., Bergman, A., Lindell, D., & Wetzstein, G. (2020). Implicit Neural Representations with Periodic Activation Functions. *Advances in Neural Information Processing Systems*, *33*, 7462-7473.

[3] Śmigiel, S., Pałczyński, K., & Ledziński, D. (2021). ECG Signal Classification using Deep Learning Techniques Based on the PTB-XL Dataset. *Entropy*, *23*(9), 1121.

Automated Analysis of Sphericity Deviation and Articulotrochanteric Distance in Legg-Calve-Perthes Disease

Salar Sanati, BS*[1], Michael Wilde, BS*[1], Vishal Gokani, MD[1], Joonoh Lee, BS[1], John Carter, BS[1], Ahan Mistry, BS[1], Adeeb Alomar, BS[1], Adan de Hoyos, BS[1], Kevin Li, BS[1], John A Herring, MD[12], Harry KW Kim, MD[12]

1 Center for Excellence in Hip, Scottish Rite for Children, Dallas, Texas, USA
2 Department of Orthopedic Surgery, University of Texas Southwestern Medical Center, Dallas, Texas, USA
*These authors contributed equally.

Introduction: Legg-Calve-Perthes disease is caused by avascular necrosis of the femoral head in pediatric patients. Radiographic evaluation is essential for assessing femoral remodeling and prognosis. Although the disease presents during skeletal growth, deformity is typically quantified at skeletal maturity. The Sphericity Deviation Score (SDS) on AP and lateral radiographs quantifies femoral head asphericity, providing an objective measure for epiphyseal collapse. The Articulotrochanteric Distance (ATD) on AP radiographs measures the distance between the femoral head apex and the tip of the greater trochanter, serving as a surrogate for femoral neck shortening and abductor mechanics. Abnormal SDS and ATD have been associated with reduced hip motion, early osteoarthritis, and progression to hip arthroplasty. However, these measurements require manual labor and are prone to poor interobserver reliability. Preliminary studies have applied computer vision models to segment femoral heads in Perthes radiographs, but there is a gap in the literature on machine-learning models applied to Perthes. We present HIPMETRICS 1.0, our software utilizing a finetuned nnU-Net model to analyze SDS and ATD. We hypothesize that HIPMETRICS 1.0 will segment relevant radiographic features and achieve SDS/ATD accuracy comparable to expert observers while improving efficiency, consistency, and clinical scalability.

Methods: We identified 800 patients from the International Perthes Study Group and Scottish Rite for Children, resulting in AP and lateral radiographs of 1000 diseased and 600 normal hips at skeletal maturity. Bilateral radiographs were bisected, and key features—femoral head, greater trochanter, lesser trochanter, femoral shaft—were manually annotated by three blinded observers. A finetuned nnU-Net model was trained for 1000 epochs to complete convergence. Post-processing of segmentations resulted in keypoints at the femoral head apex, most distant two points along the maximum femoral head diameter, tip of the greater trochanter, and three points along the mid-shaft line, which were used to measure SDS and ATD. Dice scores were calculated to assess the accuracy of model-generated masks, while intra-class correlation (ICC) was used to evaluate the accuracy of SDS and ATD measurements compared to human observers. All steps were integrated into the user-friendly, open-source HIPMETRICS 1.0 software.

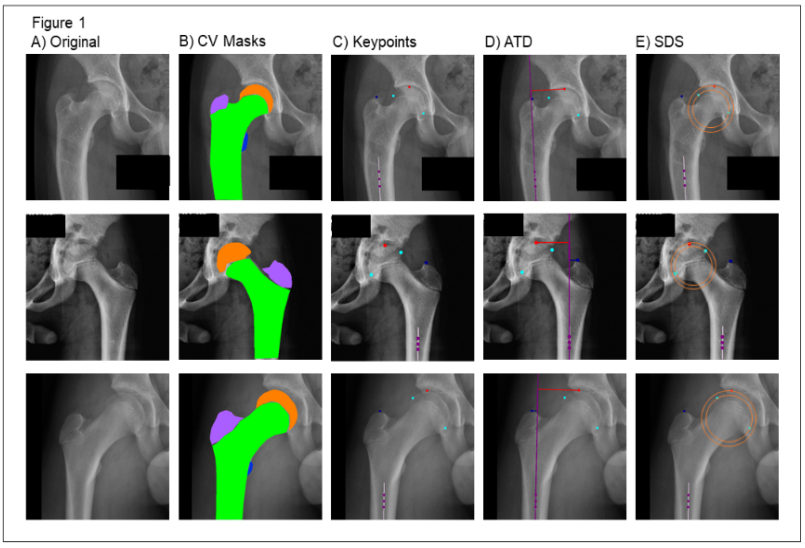


Figure 1: A) Original radiographs of Perthes patients are seen at skeletal maturity. B) CV masks segment the femoral head (orange), greater trochanter (purple), lesser trochanter (blue), and femoral shaft (green). C) Keypoints are drawn in post-processing at the femoral head apex (red), two most distant points along the maximum femoral head diameter (teal), tip of the greater trochanter (blue), and three points along the femoral mid-shaft (purple). D) The ATD is calculated as the shortest distance from the superior aspect of the femoral head to the tip of the greater trochanter (purple). E) The SDS (orange) is calculated as a measure of femoral deformity.

Results: The finetuned nnU-Net model built into HIPMETRICS 1.0 exhibited high accuracy in segmenting the femoral head (Dice 89.93%), greater trochanter (Dice 90.90%), lesser trochanter (Dice 87.39%), and femoral shaft (Dice 93.59%) (Figure 1A-B). These segmentations were then used to create keypoints (Figure 1C), which were used to compute the final SDS (pending final ICC) (Figure 1E), and ATD (ICC 0.99, $p < 0.001$) (Figure 1D), which closely matched human observer measurements.

Discussion: HIPMETRICS 1.0 efficiently segmented radiographic features and calculated SDS and ATD with accuracy comparable to human observers. While clinicians currently assess these metrics subjectively, HIPMETRICS 1.0 uses an objective computer vision model with geometric post-processing to compute the outcome measurements. Given the poor interobserver reliability in manual measurements, an automated tool offers substantial value for clinical decision-making and research standardization. Future versions of HIPMETRICS will measure intermediate-stage markers—lateral pillar classification and epiphyseal quotient—and will integrate radiomics to enhance predictive clinical modeling in Perthes disease.

An Interpretable Hybrid Machine Learning Tool for Predicting Time-to-Event Risks in Lung Transplant Patients Using UNOS Data

Gaurav Sharma, PhD, MBA[1,2,3,#], Vineet Kumar Kamal, PhD[4], Srinivas Bollineni, MD[5], Irina Timofte, MD[5], Vaidehi Kaza, MD[5], Suresh Keshavamurthy, MD[1], John Murala, MD[1], Matthias Peltz, MD[1], Michael E. Jessen, MD[1#]

[1]Department of Cardiovascular and Thoracic Surgery, University of Texas Southwestern Medical Center, Dallas, TX, United States; [2]Advanced Imaging Research Center, University of Texas Southwestern Medical Center, Dallas, TX, United States; [3]Department of Biomedical Engineering, University of Texas Southwestern Medical Center, Dallas, TX, United States; [4]Department of Biostatistics, All India Institute of Medical Sciences, Kalyani, India; [5]Department of Internal Medicine, Division of Pulmonary and Critical Care Medicine, University of Texas Southwestern Medical Center, Dallas, TX, United States

[#] Presenting Author (Gaurav Sharma, PhD, MBA **or** Michael E. Jessen, MD)

Lung transplantation (LTx) outcomes exhibit significant variability, underscoring the need for precise risk stratification tools to inform clinical decision-making. This study presents the development and validation of an interpretable hybrid machine learning tool designed to predict time-to-event risks, specifically death or retransplantation at 1-, 5-, and 10-years post-transplant in lung transplant recipients. Leveraging data from 44,731 adult patients in the United Network for Organ Sharing (UNOS)-Organ Procurement and Transplantation Network (OPTN) database (2000–2024), this tool integrates advanced computational techniques with traditional statistical methods to enhance prognostic accuracy while maintaining clinical transparency.

The methodology used the AutoScore-Survival framework[1], combining random survival forests for variable selection with Cox regression for score generation. The dataset was partitioned into training (70%), validation (10%), and testing (20%) sets. From 58 candidate variables, six key predictors were identified: length of hospital stay, post-transplant ventilation support, recipient age, type of lung transplant (single vs. double), serum creatinine, and donor age. These predictors were selected based on their predictive importance and clinical relevance[2], ensuring a parsimonious yet effective model. Model performance was rigorously evaluated using discrimination metrics, integrated area under the curve (iAUC: 0.620, 95% CI: 0.609–0.632), C-index (0.624, 95% CI: 0.614–0.632), and calibration measures (calibration slope: 1.006, 95% CI: 0.944–1.068), alongside decision curve analysis, which demonstrated clinical utility across a range of threshold probabilities.

A notable outcome of this research is the development of a web-based risk calculator (accessible at https://risk-calculator.shinyapps.io/lung_trans/), enabling real-time risk assessment for clinicians. This tool provides individualized predictions, facilitating personalized post-transplant management by identifying high-risk patients who may benefit from intensified monitoring or tailored interventions. The hybrid approach ensures interpretability, addressing a critical barrier to the adoption of complex machine learning models in healthcare by allowing clinicians to understand the contribution of each predictor to the overall risk score. By integrating machine learning with traditional statistics, the tool advances predictive analytics in transplant medicine,

offering a scalable framework for other clinical contexts. Its moderate discrimination and strong calibration underscore its reliability, while the decision curve analysis highlights its practical value in decision support. Future efforts will prioritize external validation and the incorporation of additional biomarkers to enhance predictive power. This work demonstrates how AI can transform healthcare by delivering interpretable and actionable insights, ultimately aiming to improve patient outcomes and optimize resource allocation in lung transplantation (LTx).

References:

1. Xie F, Chakraborty B, Ong MEH, Goldstein BA, Liu N. AutoScore: A Machine Learning-Based Automatic Clinical Score Generator and Its Application to Mortality Prediction Using Electronic Health Records. *JMIR Med Inform*. 2020;8(10):e21798.

2. Foroutan F, Clark K, Malik A, et al. Predictors of Mortality Post Lung Transplantation: Systematic Review and Meta-Analysis. *J Heart Lung Transplant.* 2019;38(4, Supplement):S333.

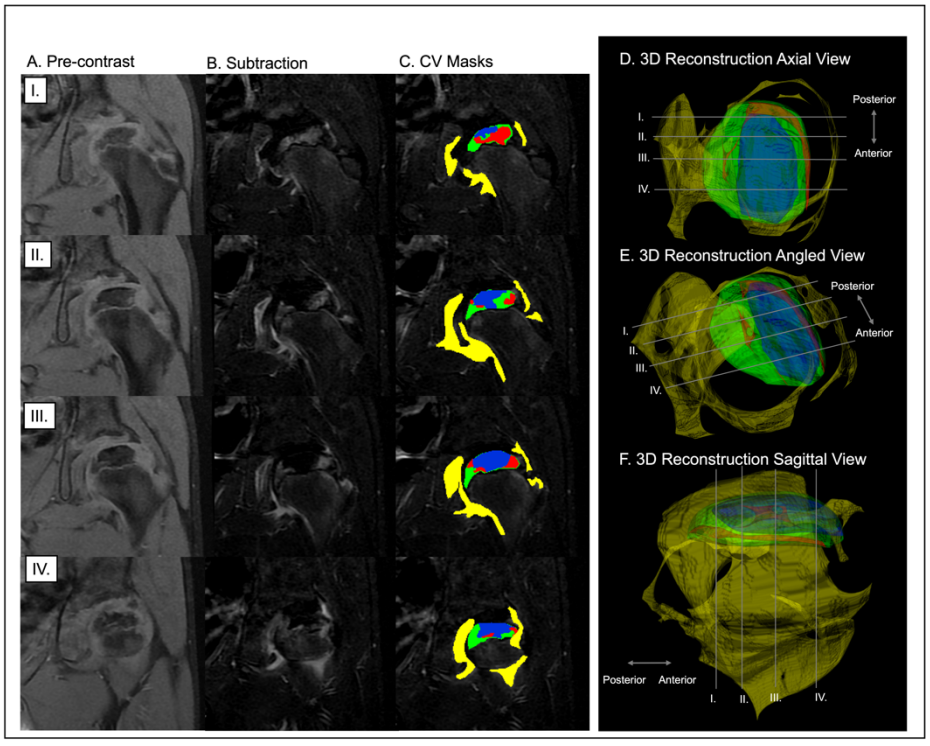
Deep Learning-Based Quantification of Radiographic Prognostic Markers in Legg-Calvé-Perthes Disease via Perfusion MRI

Vishal Gokani, MD[1*], Joonoh Lee, BS[1*], Ahan Mistry, BS[1], Michael Wilde, BS[1], Venkata Repaka, BS [1], Harry Kim, MD[1,2]
1 Center of Excellence in Hip, Scottish Rite for Children, Dallas, Texas, USA
2 Department of Orthopedic Surgery, University of Texas Southwestern Medical Center, Dallas, Texas, USA
*These authors contributed equally

Background: Legg-Calvé-Perthes Disease (LCPD) results in idiopathic osteonecrosis of the pediatric femoral head. Perfusion MRI (pMRI), enhanced by gadolinium contrast, enables quantification of femoral head perfusion and has demonstrated prognostic value, predicting the final femoral head sphericity, an essential outcome measure. Traditionally, expert clinicians visually estimated hypoperfusion percentage, although this method was subjective. HIPVASC 1.0 was released as a platform for objective manual annotation of the hypoperfusion percentage but was not user-friendly or efficient. Additionally, pMRI offers excellent visualization of associated synovitis and granulation tissue, which were not analyzed by these prior methods. Here we present HIPVASC 2.0, a user-friendly computer-vision (CV) software that automates the analysis of the femoral head, hypoperfused regions, synovitis, and granulation tissue using a multi-channel nnU-Net model. We hypothesize that HIPVASC 2.0 will efficiently segment these key features on pMRI images with the accuracy of human observers.

Methods: A total of 922 pMRI scans were retrospectively collected from an international, multicenter database of LCPD patients aged 4 to 11. The training dataset included 158 unaffected and 410 affected hips, and the testing dataset contained 178 unaffected and 176 affected hips that had previously been analyzed by visual estimation and HIPVASC 1.0. Both pre-contrast and post-contrast T1-weighted FS sequences were exported, registered, and processed to create the subtraction sequence, which was fused with the pre-contrast sequence for multichannel input. Two independent observers manually annotated the regions of interest, including the femoral head, hypovascular regions, synovitis, and granulation tissue, by referencing both sequences. The nnU-Net model was trained to 1000 epochs for full convergence with data augmentation and five-fold cross-validation. Dice similarity coefficients quantified segmentation accuracy, while intraclass correlation coefficients (ICC) assessed the accuracy of HIPVASC's hypoperfusion percentage, femoral head volume, synovitis volume, and granulation tissue percentage measurements compared to those of human observers and previous methods.

Results: The nnU-Net segmentation model demonstrated high accuracy for predicting femoral head (Dice = 94%), hypovascular regions (Dice = 97%), synovitis (Dice = 83%), and granulation tissue masks (Dice = 89%), shown in 2D (Figure 1A-C) and on 3D reconstructions (Figure 1D-F). ICC values for each biological metric were consistently high (ICC > 0.99), indicating strong measurement consistency with human observers. Hypoperfusion percentage closely matched clinician visual estimates and HIPVASC 1.0 (ICC > 0.99).



Discussion: These results validate HIPVASC 2.0 in objective segmentation of critical pathological regions in LCPD. Automated, multi-channel segmentation offers efficient, reproducible measurements essential for improved prognostication and patient management in early-stage LCPD. While hypoperfusion percentage has been shown to correlate with sphericity deviation at skeletal maturity, subsequent studies will determine the prognostic value of femoral head volume, synovitis volume, and granulation tissue percentage. Future versions of HIPVASC will be released as open-source applications and will include radiomics and feature extraction for more nuanced, accurate clinical prediction tools.

Significance: This study presents HIPVASC 2.0, which automated analysis of the femoral head, hypoperfusion region, synovitis, and granulation tissue on pMRI. This enhanced quantitative assessment could significantly refine prognosis and guide therapeutic interventions in patients with early-stage LCPD.

Figure 1: A-D) Four 2D coronal subtraction sections are shown, with vascular vessel ingrowth at the medial and lateral aspects of the head. E) The 3D reconstruction is shown from the axial view, with the locations of the four 2D coronal sections marked. The femoral head is colored green, the hypoperfused tissue is blue, granulation tissue is red, and synovitis is yellow. F) An angled 3D reconstruction is shown with synovitis encircling the femoral head. G) The 3D reconstruction is shown from a coronal sagittal view.

Redesigning Teaching and Learning for the AI Era: A Faculty Development Course on Artificial Intelligence in Clinical Education

Yulia Piller, MAT, PhD; Ramona Dorough, PhD.

UT Southwestern Medical Center, Dallas, TX

The integration of Artificial Intelligence (AI) into health professions education requires faculty to not only adopt new technologies but also critically examine their pedagogical utility. This project presents the design and implementation of a graduate-level course focused on teaching and learning in clinical education, enhanced with AI-centered content. Originally a traditional pedagogy course, it was redesigned to include AI applications within every module.

Participants engaged with a foundational AI-in-education text, applying its principles to real-world teaching contexts. Key topics included:

- Writing learning objectives using AI-supported tools
- Designing and adapting case studies with AI input
- Creating assignments, rubrics, and evaluation plans enhanced by generative AI
- Exploring ethical concerns, limitations, and risks of AI in academic and clinical contexts

The course emphasized hands-on application, requiring students to complete teaching tasks using AI, reflecting on its effectiveness, and lead discussions on AI's role in faculty workload, academic integrity, and instructional quality. Final deliverables included a critique of the course text and a summary of AI use cases in their own instructional practice.

This poster highlights the structure, outcomes, and lessons learned from the course. It also explores how graduate health education programs can equip faculty to become informed, ethical, and innovative users of AI in teaching and curriculum development.

Using Artificial Intelligence to Support Curriculum Review and Improvement in the School of Health Professions

Ramona Dorough, PhD
University of Texas Southwestern Medical Center, Dallas, Texas, United States

Oral Presentation Abstract

Health professions education must continually adapt to meet the demands of clinical practice and the needs of students entering increasingly complex healthcare environments. At the School of Health Professions, we examined the delivery and effectiveness of four core science courses: Anatomy, Physiology, Neuroscience, and Pathology. These courses are offered to students in the Physician Assistant, Physical Therapy, and Prosthetics and Orthotics programs. Because they are taught by basic scientists, questions emerged about how well the course content aligns with professional healthcare competencies.[1]

This project aimed to evaluate the courses using a combination of stakeholder feedback and artificial intelligence tools to support timely and structured curriculum review. Feedback was collected through student focus groups and discussions with course directors and program chairs. Participants raised concerns about clarity of instruction, clinical relevance of content, and whether assessments reflected course objectives.[2],[3]

To manage the feedback efficiently, we used a large language model to summarize and organize written comments. This allowed faculty to identify recurring issues more quickly than through traditional manual review.[4],[5] The output was reviewed by faculty to ensure accuracy and context, and then used to support course planning and improvement.

Faculty also reviewed course objectives, test materials, and instructional strategies. A multi-year analysis of exam data is underway to evaluate question difficulty, alignment with learning outcomes, and patterns in student performance over time. Bloom's Taxonomy is being used to examine the cognitive level of assessments and ensure that they reflect appropriate expectations for graduate students preparing for clinical practice.[6],[7]

Outcomes from this review include the adoption of a shared textbook in Anatomy, improved alignment between lecture and lab sessions, and earlier outreach to students with academic concerns. Additional clinical examples were incorporated into instruction to address concerns about limited connection between foundational science and applied healthcare practice.[8],[9]

This project shows how artificial intelligence can be used to support curriculum improvement in graduate health professions education.[4] By summarizing and organizing qualitative feedback, the process helped faculty make quicker and more informed decisions. Faculty expertise remained essential throughout the process to interpret feedback accurately and implement meaningful changes.[10]

This work offers a model for other academic programs seeking to improve the alignment between instruction, assessment, and professional expectations. It also invites broader discussion on how technology can support faculty development, assessment review, and educational decision making in healthcare education.[10]

References

1. Mahzari M, AlNahedh T, Ahmed AA, Al Rumyyan A, Shaban S, Magzoub ME. Practical guide to undergraduate medical curriculum alignment and mapping. Adv Med Educ Pract. 2023;14:1001-1012. doi:10.2147/AMEP.S424815

2. Macabrey J, Wuest LL, Buetti D. Who's at the table? A scoping review of stakeholder engagement in evaluating medical education programs. Eval Health Prof. 2024. doi:10.1177/01632787241286911

3. Khan HF, Qayyum S, Beenish H, et al. Determining the alignment of assessment items with curriculum goals through document analysis by addressing identified item flaws. BMC Med Educ. 2025;25:200. doi:10.1186/s12909-025-06736-4

4. Parker MJ, Anderson C, Stone C, Oh Y. A large language model approach to educational survey feedback analysis. Int J Artif Intell Educ. 2024;34(2):1-38. doi:10.1007/s40593-024-00414-0

5. Castellanos A, Jiang H, Gomes P, Vander Meer D, Castillo A. Large language models for thematic summarization in qualitative health care research: comparative analysis of model and human performance. JMIR AI. 2025;4:e64447. doi:10.2196/64447

6. Zahoor AW, Farooqui SI, Khan A, Kazmi SAM, Qamar N, Rizvi J. Evaluation of cognitive domain in objective exam of physiotherapy teaching program by using Bloom's taxonomy. J Health Allied Sci NU. 2023;13:289-293. doi:10.1055/s-0042-1755447

7. Rosenberger K, Skinner D, Monk J. Ready for residency: a Bloomian analysis of competency-based osteopathic medical education. J Am Osteopath Assoc. 2017;117(8):529-536. doi:10.7556/jaoa.2017.103

8. Cervantes J, Dudrey E, Baatar D, et al. Improving integration of basic science into clinical medicine: vertical integration into clinical education (VICE) activity. Med Sci Educ. 2022;32(1):47-50. doi:10.1007/s40670-021-01485-7

9. Daniel M, Morrison G, Hauer KE, et al. Strategies from 11 U.S. medical schools for integrating basic science into core clerkships. Acad Med. 2021;96(8):1125-1130. doi:10.1097/ACM.0000000000003908

10. van Schaik SM. Accessible and adaptable faculty development to support curriculum reform in medical education. Acad Med. 2021;96(4):495-500. doi:10.1097/ACM.0000000000003804

DeepHistomorphometry: An End-to-End Deep-Learning Platform for Quantitative Bone Histomorphometry and Phenotypic Profiling in Osteonecrosis of the Femoral Head

Vishal Gokani, MD; Joonoh Lee, BS†; Michael Wilde, BS†; Ahan Mistry, BS*; Salar Sanati, BS*; Alex Kim, BS; John Carter, BS; Harikrushnaa Srikanth, BS; Venkata Repaka, BS_1; Daniel Trautmann, BS_1 Garrett Croley, BS; Phillip Taboada, BS_1; Harry K.W. Kim, MD

$_1$ Center of Excellence in Hip, Scottish Rite for Children, Dallas, Texas, USA

$_2$ University of Texas Southwestern Medical Center, Dallas, Texas, USA

†These authors contributed equally.

*These authors contributed equally.

Vishal.gokani@tsrh.org

INTRODUCTION: The porcine and murine models of ONFH involve disruption of blood flow, producing histomorphometric changes on whole-slide imaging (WSI). Manual analysis of histomorphometry is unreliable, time-consuming, and limited to traditional metrics: bone surface (BS), trabecular bone area normalized to total area (BA/TA), osteoclast (OC/BS) and osteoblast (OB/BS) counts, osteoid surface (Osteoid/BS), and bone formation rate (BFR/BS). Here we present DeepHistomorphometry, an open-source platform for end-to-end automated analysis of WSI, incorporating machine-learning models trained on the largest multi-species dataset collected with over 500,000 features annotated. The purpose was to assess the accuracy of DeepHistomorphometry and to incorporate phenotypic profiling, which includes cell morphology parameters involving cell size, aspect ratio, circularity, and nuclearity. We hypothesized that DeepHistomorphometry would provide highly accurate measurements and would detect that ONFH produces traditional histomorphometry consistent with bone death and phenotypic profiling consistent with highly active phenotypes.

METHODS: 50 piglets and 50 mice underwent unilateral surgery to induce ONFH, providing histologic sections from 100 necrotic and 100 normal femoral head samples from each animal. Slides were stained with TRAP or McNeal or left unstained for fluorochrome double-labeling (DL) analysis to measure bone formation. WSI was completed at 10x magnification. Annotation by three blinded observers required over 2000 hours. 140 femoral head samples were used for training and validation while 60 were used for testing (30 pig samples shown, 30 mouse samples not shown). Each nnU-net model was trained with five-fold cross-validation, data augmentation, and early stopping. The DICE score assessed the quality of the models' masks. Intra-class correlation (ICC) assessed the accuracy of the CV models' measurements compared to human observers. Two-tailed, unpaired t-tests and Principal Component Analysis (PCA) quantified histomorphometric and phenotypic differences between the control and ONFH groups.

RESULTS: On TRAP-stained slides, DeepHistomorphometry segmented BA (Dice 94.45%), OC (Dice 84.49%), and nuclei (Dice 79.72%) with high accuracy (Fig. 1A). On McNeal-stained slides, BA (Dice 97.33%), osteoid layers (DICE 74.87%) and OB (DICE 65.67%) were outlined despite cell crowding (Fig. 1B). DeepHistomorphometry segmented red- (Dice 81.45%) and green- (Dice 85.31%) single-labeled surfaces and DL surfaces (Dice 94.49%) (Fig. 1C). ICC exceeded 0.95 ($p < 0.001$) for all measurements. DeepHistomorphometry demonstrated increased OC/BS ($p < 0.001$), decreased OB/BS ($p = 0.009$), decreased BFR/BS ($p = 0.007$), and decreased BA/TA ($p < 0.001$) in ONFH compared to normal controls and revealed increased OC size ($p = 0.018$), circularity ($p = 0.009$), aspect ratio ($p < 0.001$), and nuclearity ($p = 0.003$) (Fig. 1A) and increased OB size ($p = 0.004$), circularity ($p = 0.007$), and aspect ratio ($p = 0.003$) (Fig. 1B) in ONFH. PCA clustering demonstrated unique phenotypes (Fig. 1D-E).

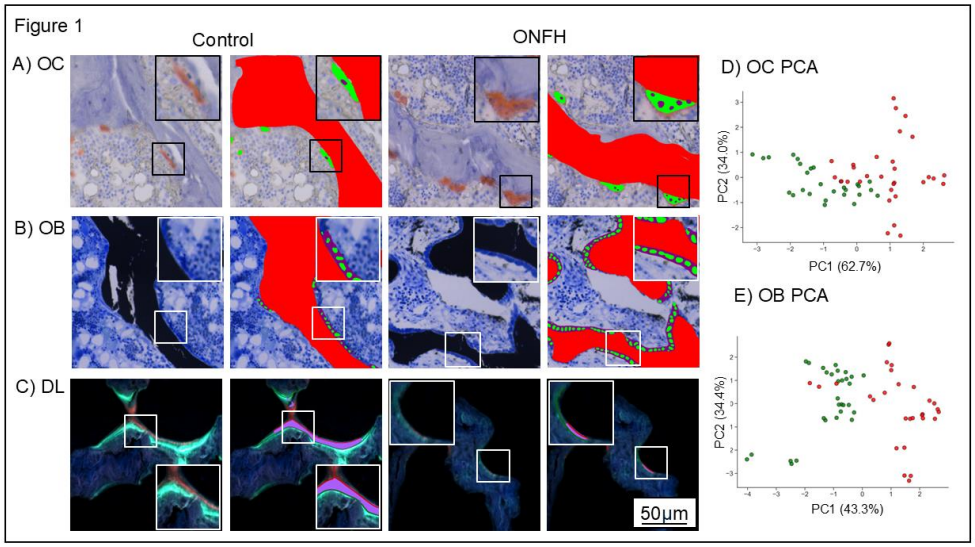


Figure 1: Original pig histology is shown alongside DeepHistomorphometry masks. Mouse histology (not shown) also supported these findings. A) In controls, OC (green) appear small with minimal nuclei (purple). These cells are largely inactive (inset). OC become enlarged, multinucleate, and elongated, resorbing trabecular bone (red) in ONFH. B) OB (green) appear small and inactive in controls. In ONFH, necrotic regions exhibit few surviving cells, although surviving OB in scant regions of bone growth appear enlarged, elongated, and hyperactive with a remodeling phenotype. C) Fluorescent DL regions (purple) span between green and red labels deposited 4 days apart to measure bone formation. Controls exhibit significant bone formation, while ONFH exhibits minimal growth. PCA incorporating traditional histomorphometry and novel phenotypic metrics reveals distinct D) OC and E) OB phenotypes in controls (green) and ONFH (red).

DISCUSSION: DeepHistomorphometry provided a user-friendly, open-source, species-agnostic, high-throughput, end-to-end automated platform for bone histomorphometry analysis of WSI. DeepHistomorphometry also added phenotypic profiling: OC size, aspect ratio, circularity and nuclearity increased in ONFH, which induced a resorptive phenotype causing OC elongation and enlargement with increased resorptive activity. Surviving OB exhibited enlargement and elongation, suggesting a hyperactive remodeling phenotype. Future studies will apply phenotypic profiling to ONFH treatments and to additional bone diseases.

Utilizing GPT-Derived Embeddings for Drug Synergy Score Prediction

Juyeon Kim, MPH[1], Peifeng Ruan, PhD[1,2], Yang Xie, PhD[1,2,3,4]

[1]Quantitative Biomedical Research Center, Peter O'Donnell Jr. School of Public Health, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA
[2]Department of Health Data Science & Biostatistics, O'Donnell School of Public Health, UT Southwestern Medical Center, Dallas, Texas, 75390, USA
[3]Simmons Comprehensive Cancer Center, UT Southwestern Medical Center, Dallas, Texas, 75390, USA
[4]Department of Bioinformatics, UT Southwestern Medical Center, Dallas, Texas, 75390, USA

INTRODUCTION. Identifying synergistic drug combinations is crucial for effective cancer therapies but remains challenging due to the vast combinatorial space and complex drug interactions. Existing in silico approaches predominantly use molecular fingerprints or the Simplified Molecular Input Line Entry System (SMILES) as drug structural features and transcriptomic profiles as cell line features, typically requiring complex models to capture interactions between drugs and cells. Embedding-based methods have recently emerged but remain underexplored in drug synergy prediction. Text-derived embeddings from large language models (LLMs), which encode diverse biomedical knowledge, can capture nuanced interactions missed by molecular fingerprints, SMILES, or transcriptomic profiles. This study assesses whether text-derived embeddings from LLMs provide a more expressive feature representation and simplify the modeling process.

METHODS. Embeddings were generated from GPT-3.5-based descriptions of drugs and cell lines using systematic prompt engineering. OpenAI's text-embedding-ada-002 model produced fixed-length embeddings, which served as input features for a simple two-hidden-layer, feed-forward neural network. For classification tasks, synergy scores were binarized at a cutoff of 10. The network was trained independently for each of five drug synergy metrics using the DrugComb dataset, encompassing over 410,000 unique combinations across various cancer studies. The model was evaluated using leave-drug-pair-out cross-validation to assess its performance on unseen drug combinations. Performance was benchmarked on an independent O'Neil dataset against three comparison models: 1. Fingerprint-based model (DeepSynergy): same two-hidden-layer architecture, but using molecular fingerprints and transcriptomic profiles—testing whether embeddings add expressiveness beyond traditional structural inputs. 2. Multi-task learning model (BAITSAO): shared two hidden layers with our GPT-derived embeddings as inputs—assessing if architectural complexity further boosts performance. 3. Baseline machine learning models: linear or logistic regression and random forest applied to the same embeddings—measuring their standalone predictive strength.

RESULTS. On the Bliss score from the DrugComb dataset, the embedding-based model achieved a Pearson correlation of 0.89 for regression tasks. For classification, it reached an AUC of 0.93 and an F1 score of 0.73. On the Loewe score, benchmark comparisons demonstrated that the embedding-based model outperformed the fingerprint-based model, the multi-task learning model, and baseline machine learning models for the regression task.

DISCUSSION AND CONCLUSIONS. Text-derived embeddings can capture rich chemical and biological knowledge, enabling a simple feed-forward network to achieve high predictive performance. Leveraging LLM-derived drug and cell line features can streamline computational workflows and accelerate the identification of potent drug combinations, advancing personalized cancer therapies.

A deep learning algorithm to predict atrial fibrillation during sinus rhythm in electrocardiograms and its association with ischemic stroke outcomes

Zifan Gu, MS[1], Kuroush Nezafati, MD[1], Jingwei Huang PhD[1], Ann Marie Navar, MD, PhD[4], Eric D. Peterson, MD, MPH[4], Donghan M. Yang, PhD[1], Ruichen Rong, PhD[1], Guanghua Xiao, PhD[1,2,3], Yang Xie, PhD[1,2,3]

[1]Quantitative Biomedical Research Center, Peter O'Donnell Jr. School of Public Health, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA
[2]Simmons Comprehensive Cancer Center, UT Southwestern Medical Center, Dallas, Texas, 75390, USA
[3]Department of Bioinformatics, UT Southwestern Medical Center, Dallas, Texas, 75390, USA
[4]Department of Internal Medicine, The University of Texas Southwestern Medical Center, Dallas, TX 75390, USA

IMPORTANCE. Early detection of atrial fibrillation (AF) before clinical symptoms appear could help prevent thromboembolic events such as stroke. Deep neural networks have demonstrated the ability to predict various forms of AF, with the potential to identify AF-related ischemic stroke risks.

OBJECTIVE. To assess the feasibility of using deep learning to detect paroxysmal AF and assess ischemic stroke risk in asymptomatic high-risk patients.

DESIGN, SETTING, AND PARTICIPANTS. This retrospective study analyzed 12-lead ECGs from the MIMIC-IV database, collected from the Beth Israel Deaconess Medical Center between 2008 and 2019. The dataset includes ECGs from outpatient visits, the emergency department (ED), and the intensive care unit (ICU). Paroxysmal AF was identified through cardiologist reports, and ischemic stroke was identified through diagnostic codes.

MAIN OUTCOMES AND MEASURES. A convolutional neural network was trained with a ResNet backbone using ECG traces as the sole input to predict paroxysmal AF from normal sinus rhythm (NSR). Model performance was validated on a held-out MIMIC-IV dataset. This model's ability to identify patients at high risk for ischemic stroke was also assessed.

RESULTS. A total of 28,367 ECGs were included in the stroke prediction analysis. The model achieved an area under the receiver operating characteristic (AUROC) curve of 0.76 and an F1 score of 0.68. The odds ratio for high- versus low-risk AI-ECG groups over one year was 1.76, with a hazard ratio of 1.55 over 11 years.

CONCLUSIONS AND RELEVANCE. This deep learning algorithm enables a two-stage risk assessment, detecting paroxysmal AF and stratifying ischemic stroke risk. Implementing this tool in clinical practice could reduce the public health burden associated with AF and stroke.

Title: Performance of an artificial intelligence foundation model for prostate radiotherapy segmentation

Matt Doucette, BS[1]; Chien-Yi Liao, PhD[1]; Mu-Han Lin, PhD[1,2]; Steve Jiang, PhD[1,2]; Dan Nguyen, PhD[1,2]; Daniel X. Yang, MD[1,2]

[1]Medical Artificial Intelligence and Automation (MAIA) Laboratory, University of Texas Southwestern Medical Center, Dallas, TX
[2]Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, TX

Abstract

Background: Radiation therapy is a major treatment modality for localized prostate cancer, and accurate target segmentation is critical aspect of radiation delivery that can directly impact patient outcomes. While artificial intelligence (AI) segmentation tools exist, foundation models have recently emerged as a novel paradigm for specialized tasks with minimal domain-specific retraining or fine-tuning. However, their performance in clinical applications like radiotherapy planning is not well understood. We therefore evaluated the performance of the SAM 2 foundation model in segmenting pre-operative intact prostate and post-operative prostate fossa targets for prostate radiotherapy planning.

Methods: Retrospective cohort study deploying and testing a foundation model for AI segmentation for prostate radiotherapy planning. CT simulation images and radiation plans were obtained from a single academic institution for 282 patients (139 pre-operative, 143 post-operative) undergoing prostate cancer treatment. Data analysis was performed from September 2024 to February 2025. AI segmentation was evaluated with varying levels of human intervention, ranging from intervals of every 2nd to every 10th ground truth slice provided as input. Segmentation accuracy was measured by Dice Similarity Coefficient (DSC) and Hausdorff Distance (HD) for intact and post-operative prostate target delineation.

Results: While SAM 2 outperformed interpolation in DSC and HD for both intact and post-operative prostate cancer patient cases, the AI segmentation accuracy was significantly better in the intact pre-operative patient cases where anatomic boundaries were better defined than post-operative patient cases. This is especially evident when sparse ground truth was provided simulating lower levels of human intervention, as demonstrated by DSC values decreasing from 0.956 to 0.856 in the pre-operative setting and from 0.952 to 0.751 in the post-operative setting when the interval between ground truth slices was increased from every 2nd to every 10th slice. Correspondingly, HD increased from 3.666 to 7.450 in the pre-operative setting and from 4.020 to 15.36 in the post-operative setting under the same conditions.

Conclusions: AI foundation models show promising application for specialized medical tasks such as prostate cancer radiotherapy segmentation with limited need for fine-tuning or retraining, although their clinical application will require further understanding of task-specific performance.

Assessing User Perceived Value of Ambient Note Technology in Academic Centers: Usability, Time Efficiency, and Willingness to Invest

Ling Chu MD, Samuel McDonald MD, **Justin Rousseau MD, Catherine Chen MD, Robert Turer MD, DuWayne Willett MD**
University of Texas Southwestern, Dallas, Texas

Background: Generative AI has been shown to decrease clinician documentation burden and is being increasingly adopted across the industry. Previous studies have demonstrated that such tools reduce cognitive load and increase clinician satisfaction. However, cost remains a significant concern for widespread implementation. This pilot study evaluates the impact of Microsoft Nuance DAX (Dragon Ambient eXperience) on outpatient clinicians' documentation complemented by a qualitative assessment to gauge physician perception of the tool's value.

Methods: This pilot study was conducted at an academic center to measure documentation time and physician perceived value of DAX from July to December 2024. Documentation time and DAX usage data were extracted from Epic's Clarity Database for all physicians in the pilot. A mixed-effects model was developed to evaluate the impact of DAX use on documentation times. Fixed effects included DAX use, adoption level (high vs. low based on a 50% usage threshold), and their interaction to assess how adoption influenced documentation time. Random intercepts for physicians accounted for clustering and variability. Absolute time differences were calculated to complement relative percentage changes. Physicians were enrolled in the pilot via REDCap survey and received standardized surveys after two months. Usability was measured using Nielsen's System Usability Scale (SUS). The survey also assessed the number of additional patients each physician would be willing to see (0-6) and their willingness to self-fund DAX access ($0-2000). Results are presented using median/IQR and mean/SD.

Results: 59 physicians (24 specialties) participated in the pilot encompassing 51,330 encounters. DAX-assisted documentation had a median duration of 709 seconds (IQR: 211 – 1735, n = 10,268) compared to 828 seconds (IQR 244 – 1749, n=41062) for standard documentation, with an adoption rate of 0.39 (IQR 0.16 - 0.73). High-adopters showed a 16.8% longer documentation time for non-DAX notes compared to DAX notes, corresponding to an absolute increase of 125 seconds. Among low adopters, DAX-enabled notes took 54.4% longer compared to high adopters, with the gap between non-DAX and DAX documentation times narrowing by approximately 49 seconds (interaction term $\beta = -0.607$, $p < 0.001$). 26 physicians responded to the survey with a SUS mean score of 83 (SD 14.5). The survey indicated 42% of respondents were willing to see more patients per visit (median: 1, IQR 0 – 3), and 58% were willing to fund a portion of the monthly cost (mean: $440, SD $525) to maintain DAX access.

Discussion: Generative AI tools offer a transformative opportunity to reduce clinician documentation burden and enhance workflow efficiency. High DAX adoption correlates with improved documentation times, suggesting a learning effect among consistent users. With a SUS score of 83, DAX demonstrates excellent usability. Cost remains a key barrier to implementation. The willingness

of users to fund part of the tool's cost and to see additional patients underscores its value. Identifying high-adopter physicians and specialties that benefit most from DAX can optimize return on investment and make adoption more cost-effective, guiding broader deployment, alleviating burnout, and driving adoption across health systems.