Table of Contents

Oral Presentations

Poster Presentations

continued on next page

continued on next page

continued on next page

AI-Enabled Clinical Documentation in Ambulatory Primary Care: A Prescription for What Is Ailing Healthcare?

James J. Geracci, M.D.

Background: The healthcare experience, for both patients and providers, has deteriorated significantly over the past decade or more. 75% of patients report that they wish their healthcare experiences were more personalized and 61% report that they would visit their providers more regularly if that were the case. 42% of physicians report that they are burned out and 64% acknowledge that burnout has worsened since the pandemic. Some would argue that technology has not only done little to enhance the healthcare experience, but may in fact have exacerbated the situation with the advent and ubiquitous adoption of the electronic health record (EHR). Studies show that approximately 50% of physicians' time is now spent using the EHR and half of that time is while they are interacting with patients. AI has significant potential to impact, if not transform, healthcare over the next decade. But are there practical applications that can have an impact now while we are waiting for that potential to be realized?

Methods: We implemented an AI-enabled ambient listening copilot solution for automated clinical documentation across a community-based primary care practice. Utilizing an interventional (pre-post) study design, EHR documentation time, provider burnout, and patient satisfaction were measured before and after implementing the solution.

Results: Our outcomes exceeded our expectations including an average of 40% reduction in documentation time (5 minutes saved per encounter), 50% reduction in feelings of burnout amongst providers, 80% of patients reported their providers were more personable and conversational during their visit, and 50% of providers reported that overall documentation quality was improved.

Conclusions: Utilization of physician time directly impacts not only patient experience and physician professional satisfaction, but also the quality and cost of healthcare as well. Given high rates of physician burnout, declining patient satisfaction, and concerns about rising healthcare costs, optimizing the use of physician time should be a top priority for health systems. The results of our recent experience implementing AI-enabled clinical documentation in the ambulatory primary care setting suggest that there are practical applications of AI technology that can be implemented now to address some of the most significant challenges facing healthcare today.

Automated Task Segmentation of Robotic Gastrojejunostomy Videos Using Advanced Deep Learning

Amr I. Al Abbas M.D.[1]; Huu Phong Nguyen Ph.D.[1]; Shruti R. Hegde, M.D.[1]; Sofia Garces Palacios, M.D.[1]; Andres Abreu M.D.[1]; Patricio M. Polanco M.D.[1]; Amer Zureikat, M.D.[2]; Melissa E. Hogg M.D., M.S.[3]; Herbert J. Zeh III, M.D.[1]; Ganesh Sankaranarayanan, Ph.D.[1]

[1]UT Southwestern Medical Center, Dallas, TX; [2]University of Pittsburgh Medical Center, Pittsburgh, PA; [3]Northshore University Health System, Evanston, IL

Introduction: Minimally invasive surgery allows for video review to evaluate performance. This can be time-consuming and expensive. Deep learning provides an affordable and reproducible alternative. Robotic Pancreaticoduodenectomy (RPD) is a complex procedure. The surgeon's performance during gastrojejunostomy (GJ) has been associated with complications. In this work, we aim to utilize deep learning to automatically segment GJ videos from RPD for future performance analysis.

Methods: Retrospective review of cases at two tertiary referral centers from 2017 to 2021. We adopted a deep Convolutional Neural Network (CNN) for frame-level visual feature extraction and classification. Each robotic GJ video was labeled into 6 tasks in addition to idle time. Training was through measuring the model's error with respect to the ground-truth and optimizing the parameters of the model using backpropagation. The CNN architecture incorporated X3D model trained on RGB and Optical Flow data.

Results: Of the 42 videos included, 30 were used for training and 12 for validating the performance. All frames were extracted (6 frames/second) and annotated. The accuracy and per-class F1-score were 75.51% and 68.55% for tasks. The results for each task are shown in Table 1. The processing time of each frame for the model in ~0.02 seconds, which makes it suitable for real-time application.

Conclusion: Deep learning can be used reliably for automatic task segmentation of robotic GJ videos. Techniques such as few-shot learning will be investigated to improve performance. Future applications of this model include the design of automated systems for skills assessment, provision of real-time feedback, and ultimately outcome prediction.

Table1. Performance of Deep Learning Model by Task

Task	Precision	Recall	F1-score	# of frames
1.1 Stay suture	0.58	0.66	0.62	346
1.2 Inner running suture	0.84	0.93	0.88	3367
1.3 Enterotomy	0.82	0.93	0.88	3155
2.2 Inner running suture	0.70	0.92	0.79	5359
3.1 Inner Layer of Connell	0.73	0.76	0.74	5083
4.1 Outer Layer of Connell	0.84	0.78	0.81	2491
0.1 Idle	0.37	0.04	0.08	2946

Detecting Inconsistent Suicide Cause Annotations Using Large Language Models

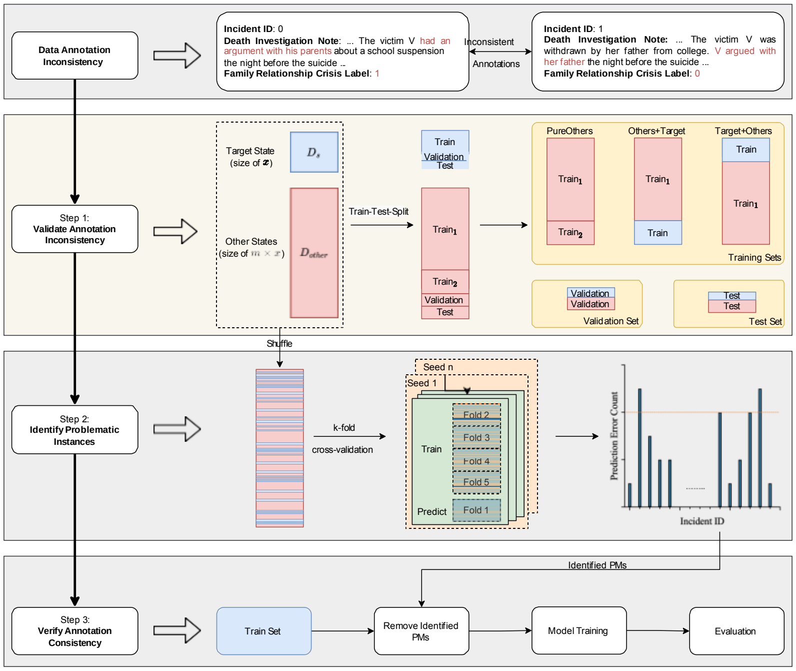
Song Wang[1], Yiliang Zhou[2], Ziqiang Han[3], Cui Tao[4], Yunyu Xiao[2], Ying Ding[1], Joydeep Ghosh[1], Yifan Peng[2]

[1]The University of Texas at Austin, Austin, TX; [2]Weill Cornell Medicine, New York, NY; [3]Shandong University, Qingdao, China; [4] Mayo Clinic, Rochester, MN

Introduction: Data accuracy is essential for scientific research and policy development. The National Violent Death Reporting System (NVDRS) is a state-based reporting system in the U.S. that provides detailed information on violent deaths.[1] Recent studies have suggested annotation inconsistencies between different U.S. states in the NVDRS and their potential impact on erroneous suicide cause attributions. However, existing annotation error detection methods cannot be directly applied to free-text death investigation notes. In this work, we present a novel and powerful approach to detect inter-state annotation inconsistencies. This study is approved by the NVDRS Restricted Access Database (RAD) Proposal.

Methods:

Figure 1 shows a novel framework to detect inter-state annotation inconsistencies utilizing the pre-trained BERT language model as a backbone classifier. Specifically, we refer to the state under evaluation as the "target state" and all other states as the "other states." We focused on three suicide-related factors: Family Relationship Crisis, Mental Health Crisis, and Physical Health Crisis because of their higher prevalence of positive instances in the NVDRS dataset, and their poor classification scores as demonstrated in prior work. We calculated the annotation inconsistencies by determining the degree of decrease in the F-1 score when the model training data was switched from data sampled from the target state to data sourced from other states. We used a cross-validation-like paradigm to identify the problematic instances. We analyzed 267,804 suicide death incidents between 2003 and 2020.



Incident ID: 0
Death Investigation Note: ... The victim V had an argument with his parents about a school suspension the night before the suicide ..
Family Relationship Crisis Label: 1

Incident ID: 1
Death Investigation Note: ... The victim V was withdrawn by her father from college. V argued with her father the night before the suicide ...
Family Relationship Crisis Label: 0

Figure 1. Architecture for inter-state annotation discrepancy detection and remediation.

Results and Conclusions:

Our approach flagged likely inter-state annotation discrepancies, showing that incorporating the target state's data into model training brought an increase of 5.4% to the F1 score on the target state's test set and a decrease of 1.1% on other states' test set. Rectifying the identified problematic data instances resulted in an average F1 increase of 3.85%. This annotation inconsistency detection framework can also be applied to other state-based reporting systems, such as the Fatality Analysis Reporting System (FARS).[2] By addressing such inconsistencies, we hope to pave the way for a more reliable utilization of the NVDRS data in studying suicide causes and suicide preventions.

Bibliography:

1. Wilson RF, Liu G, Lyons BH, Petrosky E, Harrison DD, Betz CJ, Blair JM. Surveillance for Violent Deaths – National Violent Death Reporting System, 42 States, the District of Columbia, and Puerto Rico, 2019. *MMWR Surveill Summ*. 2022 May 20.

2. National Highway Traffic Safety Administration. "Fatality Analysis Reporting System (FARS)." www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars.

Equitable AI: Ensuring Fair Predictions in Distributed Healthcare Systems

Disha Makhija[1], Xing Han[2], Joydeep Ghosh[1, 3]

[1]The University of Texas at Austin, Austin, TX; [2]Johns Hopkins University, Baltimore, MD; [3]Dell Medical School, Austin, TX

Introduction:

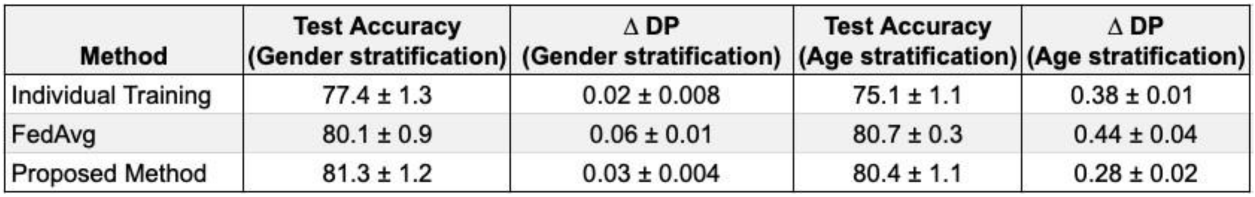
The often observed disproportionate misrepresentation of risks for specific social groups in AI models highlights the critical importance of emphasizing equitable AI in healthcare systems. Neglecting this aspect can lead to the creation, replication, or amplification of bias and discrimination in society. While some approaches have employed federated learning (FL) to tackle the unequal impact by constructing models with broader and more varied patient cohorts, this approach introduces new challenges in addressing algorithmic biases. For example, in multi-site learning on the TCGA data for breast cancer, the majority of the sites consist largely of patients from European ancestry and there is only one site that has a majority representation from African ancestry, leading to racial bias in both local and collective models. This work provides a technique for ensuring algorithmic fairness within the FL framework to mitigate disparities in diagnostic accuracy across different demographic groups.

Methods:

FL refers to privacy-preserving distributed learning achieved by training models on local sites and periodically sending updated weight parameters to a centralized server without sharing any patient-level data. We propose an in-processing method that explicitly incorporates a fairness constraint in the optimization goal at each participating site locally, and then achieving collaboration across sites by appropriately aggregating the local site models at the server to ensure both performance and fairness. By doing this, we are able to achieve a better privacy-utility trade-off with our method as opposed to both local standalone training on sites (which suffers due to the lesser data being available), as well as conventional FL procedures which do not prioritize fairness.

Results and Conclusions:

We tested our method on the Health dataset from HPN that predicts the number of days of hospitalization for each patient based on various parameters. A representative result is shown in Table 1, wherein our method achieves better predictive performance as well as better demographic parity with respect to the age and the gender attributes of the patient. While we demonstrate test accuracy and demographic parity as indicators of performance and fairness, the method can seamlessly work with other performance and fairness evaluation metrics.



Method	Test Accuracy (Gender stratification)	Δ DP (Gender stratification)	Test Accuracy (Age stratification)	Δ DP (Age stratification)
Individual Training	77.4 ± 1.3	0.02 ± 0.008	75.1 ± 1.1	0.38 ± 0.01
FedAvg	80.1 ± 0.9	0.06 ± 0.01	80.7 ± 0.3	0.44 ± 0.04
Proposed Method	81.3 ± 1.2	0.03 ± 0.004	80.4 ± 1.1	0.28 ± 0.02

Table 1. Comparison of the obtained average test accuracy and the difference in demographic parity across different groups.

FunctionalAI4EHR: A Language Model-Empowered AI Framework for Functional Status and Fall Occurrence Ascertainment on Electronic Health Records

Sunyang Fu, Ph.D.[1]; Nahid J Rianon, M.D., Dr.P.H.[2]; Min Ji Kwak, Dr.P.H., M.S.[2]; Hongfang Liu, Ph.D.[1]

[1]McWilliams School of Biomedical Informatics, Center for Translational AI Excellence and Applications in Medicine; [2]Division of Geriatric and Palliative Medicine, McGovern Medical School, The University of Texas Health Science Center at Houston, TX

Introduction: Patients' functional status assesses their independence in their abilities to perform activities of daily living (ADL). Poor functional status is highly associated with a fall, a leading cause of injury, especially among older adults. Identifying individuals at high risk for poor functional status and future falls can provide the opportunity to implement early, tailored, and multifactorial interventions. However, much of the functional status and fall information is stored longitudinally in electronic health records (EHRs) in either semi-structured or free text formats.

Method: We developed and validated FunctionalAI4EHR (Figure 1), a hybrid and federated AI framework to systematically identify impaired functional status event (e.g., cannot take shower independently), history of falls, and risk of falls from unstructured notes. The framework contains a set of pre-trained language representation models for sentence classification, a resource-driven open-source information extraction (IE) framework, a Flower-based federated learning engine and a federated extract, transform and load pipeline. We used correlation analysis and ordinal regression model (ORM) to assess the association between aggregated FunctionalAI4EHR score and established clinical parameters, i) Frailty scores by FRAIL scale, ii) grip strength, iii) 6-meter walking speed and iv) zibrio balance score in older adults (50 years and older) from a geriatric medicine outpatient clinic.

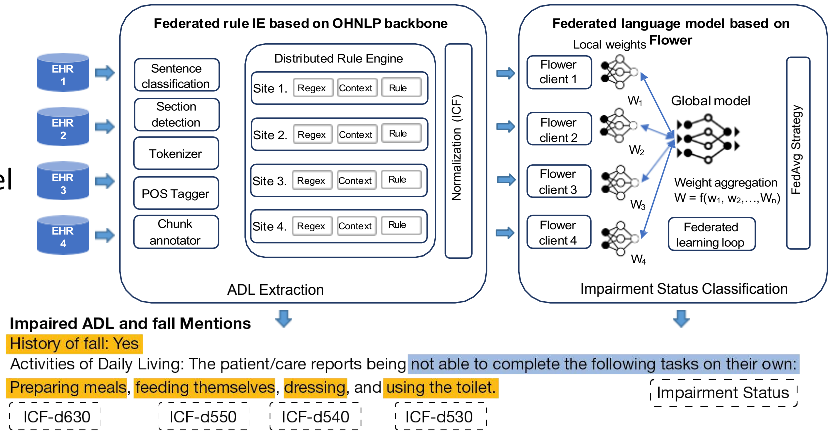

Federated rule IE based on OHNLP backbone
EHR 1
EHR 2
EHR 3
EHR 4
Sentence classification
Section detection
Tokenizer
POS Tagger
Chunk annotator
Distributed Rule Engine
Site 1. Regex Context Rule
Site 2. Regex Context Rule
Site 3. Regex Context Rule
Site 4. Regex Context Rule
Normalization (ICF)
ADL Extraction
Federated language model based on Flower
Local weights
Flower client 1
Flower client 2
Flower client 3
Flower client 4
W1
W2
W3
W4
Global model
Weight aggregation W = f(w1, w2,…,Wn)
Federated learning loop
FedAvg Strategy
Impairment Status Classification
Impaired ADL and fall Mentions
History of fall: Yes
Activities of Daily Living: The patient/care reports being not able to complete the following tasks on their own: Preparing meals, feeding themselves, dressing, and using the toilet.
Impairment Status
ICF-d630
ICF-d550
ICF-d540
ICF-d530

Figure 1. Overview of Function AI4EHR Architecture balance score in older adults (50 years and older) from a geriatric medicine outpatient clinic.

Result: Analysis showed aggregated functional status (poor or better) screening based on Functional AI4EHR on longitudinal EHR data correlated with real-time data reporting 6-meter walking speed in seconds (longer is worse, correlation coefficient [cc]: -0.31, p=0.01) and zibrio balance score (lower is worse, cc: 0.47, p<0.001). The ORM analysis examined the association between frailty status and AI Function score, revealing a statistically significant odds ratio of 1.34 for frailty status (p < 0.001).

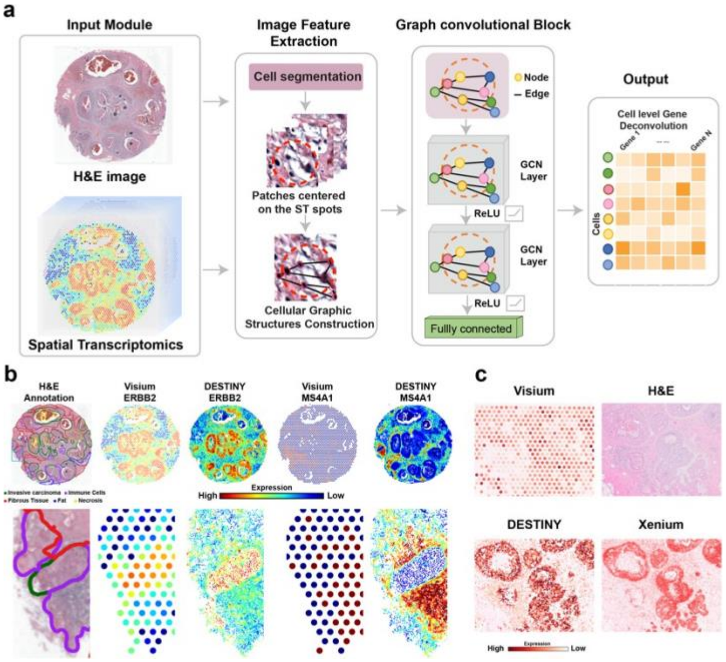
Conclusion: Our preliminary work shows promise that Functional AI4EHR can be used to identify patients with poor functional status that corroborates with clinical assessments documented over time in their health records.

Graphic Convolutional Network that Integrating Histology Information with Spatial Transcriptomics to Uncovering the Spatial Cell-Level Gene Expression

Qin Zhou[1,#]; Shidan Wang[1,#]; Yang Liu[1]; Kenian Chen[1]; Zhuoyu Wen[1]; Tingyi Wanyan[1]; Peiran Quan[1]; Ruichen Rong[1]; Lin Xu[1]; Guanghua Xiao[1,2,3,*]; Yang Xie[1,2,3,*]

[1]Quantitative Biomedical Research Center, Peter O'Donnell Jr. School of Public Health, UT Southwestern Medical Center, Dallas, Texas; [2]Simmons Comprehensive Cancer Center, UT Southwestern Medical Center, Dallas, Texas; [3]Department of Bioinformatics, UT Southwestern Medical Center, Dallas, Texas
*Corresponding authors; #Contributed equaly

Objective: Spatial transcriptomics (ST) has recently made significant progress in capturing gene expression within tissue sections. However, most of the ST platforms only measured the spot-level gene expression, cannot provide cellular-level resolution. This limitation becomes especially significant in studies involving heterogeneous cell types, for instance, tumor microenvironments. To address this critical gap, we introduce DESTINY (Deconvoluting Spatial Transcriptomics Signals to Single-Cell Portrayal), an innovative framework showcasing remarkable efficacy in predicting gene expression of individual cells from ST data.



Method: DESTINY only requires two inputs: ST data and histology image. Through constructing graph convolutional networks, DESTINY integrates cell spatial organization and morphology information from histology images with ST spot gene expression profiles to uncover cell-level gene expression.
Here is how it works (**Figure a**):

1. **Patch Extraction**: Image patches are extracted based on the spot coordinates in the ST data.
2. **Cellular Graphic Structure Construction**: For each patch, a directed cellular graphic structure is constructed using nearest neighbors determined by the Euclidean distance between cells.
3. **Feature Integration:** Features from the directed cellular graphic structures were derived to incorporate the cell spatial and morphology information into the graph convolutional network, with Node features (morphological features) and Edge features (cell-cell spatial interactions features).
4. **Convolutional layer**: All the cell nodes are updated through messages passing from all the neighboring nodes in each convolutional layer, which incorporate both morphological and cell-cell spatial interaction features of neighboring cells.

Results: In this study, we applied the DESTINY model to two breast cancer datasets, one lung cancer dataset, and one ovarian cancer dataset. Across all tissue types, DESTINY demonstrates remarkable advancements in deciphering cell-level gene expression patterns. Notably, DESTINY effectively enhances gene expression resolution (**Figure b**, top), and its cell-level gene expression shows outstanding capability in reliably distinguishing between cell layers (**Figure b**, bottom). Additionally, the cellular gene expression of DESTINY is further validated through the Xenium experiment, exhibiting remarkable consistency (**Figure c**).

Conclusion: We showed the high accuracy and resolution of DESTINY in revealing spatial cell-level gene expression in this study. This model can be readily applied to diverse types of tissues and spatial transcriptomic data resources, offering a comprehensive illustration of tissue heterogeneity and cell-cell communications, and providing valuable insights into tumor immunology studies.

PD-L1 Expression Prediction Using Multi-Scale Ensemble Transformer (SCENT)

Amgad Muneer[1,†], Eman Showkatian[1,†], Maliazurina B. Saad[1], Muhammad Aminu[1], Lingzhi Hong[1,3], Morteza Salehjahromi[1], Sheeba J. Sujit[1], Muhammad Waqas[1], Natalie I Vokes[3], Carol C. Wu[6], Brett W. Carter[6], Joe Y. Chang[4], Xiuning Le[3], Ignacio I. Wistuba[3,7], Caroline Chung[4], David Jaffray[1,9], Don L. Gibbons[3], Ara Vaporciyan[12], J. Jack Lee, John V. Heymach[3], Jianjun Zhang[2,3], Jia Wu[1,3,9]*

[1]Department of Imaging Physics, MD Anderson Cancer Center, Houston, TX; [2]Department of Genomic Medicine, MD Anderson Cancer Center; [3]Department of Thoracic/Head and Neck Medical Oncology, MD Anderson Cancer Center; [4]Department of Radiation Oncology, MD Anderson Cancer Center; [5]Department of Nuclear Medicine, MD Anderson Cancer Center; [6]Department of Thoracic Imaging, MD Anderson Cancer Center; [7]Department of Translational Molecular Pathology, MD Anderson Cancer Center; [8]Department of Biostatistics, MD Anderson Cancer Center; [9]Institute for Data Science in Oncology, MD Anderson Cancer Center; [10]Lung Cancer Genomics Program, MD Anderson Cancer Center; [11]Lung Cancer Interception Program, MD Anderson Cancer Center; [12]Department of Thoracic and Cardiovascular Surgery, MD Anderson Cancer Center

[†] Contributed equally

Background: Immune checkpoint inhibitors (ICIs) offer a durable clinical benefit to approximately 20–30% of non-small cell lung cancer (NSCLC) patients. However, Programmed death-ligand 1 (PD-L1) status, ascertained via immunohistochemistry (IHC) on biopsy specimens, remains the sole NCCN-endorsed biomarker for initiating ICI therapy. This study was conducted to address the need for improved non-invasive prediction of PD-L1 expression in patients with metastatic NSCLC using chest computed tomography (CT) scans.

Purpose: Our objective was to develop and validate SCENT (Scalable Ensemble Transformer), a deep learning model to accurately predict PD-L1 expression from CT imaging, thus reducing the need for invasive biopsies.

Materials and Methods: In this retrospective study, two stage Stage-IV metastatic NSCLC immunotherapy cohorts (*n*= 1080) were analyzed, with cohort 1 utilized for discovery purposes and cohort 2 for validation. The discovery cohort was divided into three subsets for training (*n*=298), tuning and internal validation (*n*=75), and testing (*n*=373). While the second cohort (*n*=332) with no PD-L1 measurements were utilized for validation. These were analyzed using SCENT to accurately predicts PD-L1 expression status. This developed model was further utilized to stratify patients with metastatic NSCLC according to their progression-free survival (PFS) and overall survival (OS). We compared the SCENT model's performance against traditional 2D, 2.5D, and 3D models, as well as the radiomics and clinical models using a cohort of 746 patients.

Results: The SCENT model outperformed conventional models (clinical and radiomics) in predicting PD-L1 expression with higher specificity (81.59%), sensitivity (82.14%), and AUC (80.50%) in the testing cohort. It also showed a consistent improvement across various training set sizes, demonstrating robustness and adaptability (AUC, 82.1%). Significantly, the SCENT model demonstrated comparable performance to Immunohistochemistry (IHC)- derived PD-L1 status in prognosticating OS and PFS, indicating the potential of SCENT model as a substitute for IHC.

Conclusion: The SCENT model offers a significant advance in the non-invasive prediction of PD-L1 expression, with potential implications for the management of Stage-IV metastatic NSCLC patients. This approach can streamline the selection process for immunotherapy, presenting a shift towards more personalized treatment strategies.

Race-Agnostic Machine Learning-Based Models Improve Incident Atrial Fibrillation Prediction

Matthew W. Segar, M.D., M.S.; Neil Keshvani, M.D.**;** Byron Jaeger, Ph.D.; Kershaw Patel, M.D., M.S.C.S.; Shreya Rao, M.D., M.P.H.; Mehdi Razavi, M.D.; Mohammad Saeed, M.D.; Utibe Essien, M.D.; David Kao, M.D.; Ambarish Pandey, M.D., M.S.C.S.

Background: Existing models to predict atrial fibrillation (AF) risk, such as CHARGE-AF, include race as a covariate and may contribute to the existing treatment racial disparities. Machine learning (ML) may help reduce treatment disparities by improving accuracy of AF risk prediction with race- agnostic models.

Objective: To develop and externally validate a race-agnostic, ML-based model that includes social determinants of health (SDOH) to predict incident AF.

Methods: The derivation cohort included participants from 2 community studies (ARIC, CHS) who were free of AF at baseline AF outcomes adjudicated on follow-up. A ML model was derived using oblique random survival forests and externally validated among participants from 3 additional cohort studies (MESA, FOS, and FHS Gen 3). Variable importance was assessed using negation method and performance was assessed using Harrell's C-index, index of prediction accuracy (IPA, higher=better), and decision curve analysis. The model was compared to the CHARGE-AF risk score for both performance and quantified bias metrics.

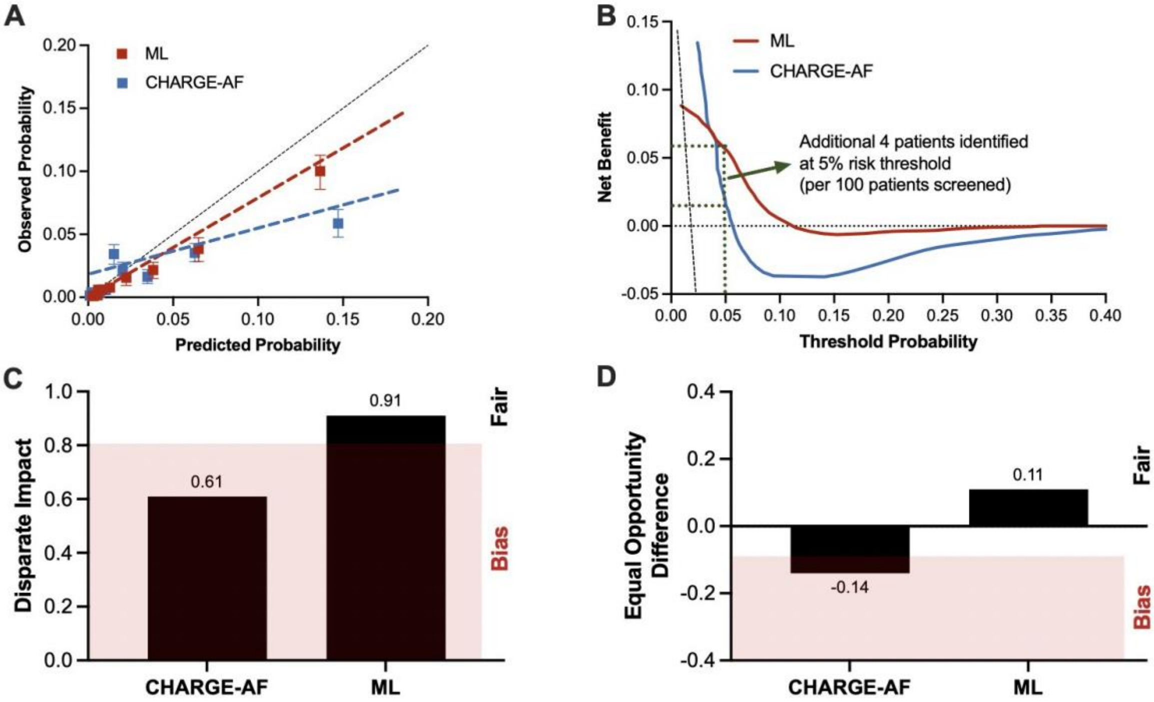
Results: In the derivation cohort (N=16,709), 504 (3.0%) participants developed AF within 5 years. The most important variables for predicting AF included age, SBP, biomarkers (NP-levels, troponin, CRP), lab values (creatinine, FPG), CVD, QRS duration, and SDOH measures (insurance, education). In the validation cohort [N=13,782; 252 (1.8%) AF events), the C-index and IPA for the race-agnostic ML model were 0.83 and 4.0%, respectively, and significantly higher than the CHARGE-AF risk score (0.77 and -2.1%, respectively) (Fig. A). At a 5% risk threshold, an additional 4 individuals per 100 screened would be identified with the ML model (Fig. B). The ML model also displayed significantly improved fairness compared with the CHARGE-AF risk score (Fig. C-D).

Conclusion: A race-agnostic and ML-based AF risk model that integrated SDOH, demographic, and biomarker data demonstrated superior performance when compared with a traditional AF risk equation.

Figure. A) Calibration plot of the ML model compared to the CHARGE-AF risk score. **B)** Decision curve analysis of the ML model compared to the CHARGE-AF risk score. **C-D)** Difference in race bias metrics between the CHARGE-AF and the race-agnostic, ML based on number of participants with incident atrial fibrillation and predicted risk > 5%. Definitions for each bias metric are as follows:

Disparate Impact: ratio of the representation rate of a protected class to the rate of a reference class.

Equal Opportunity Difference: difference in true positive rates between the unprivileged and the privileged groups



Self-Supervised Hybrid Neural Network to Achieve Quantitative Bioluminescence Tomography for Cancer Research

Beichuan Deng[1], Zhishen Ton[1], Xiangkun Xu[1], Hamid Dehghani[2], Ken Kang-Hsin Wang[1]

[1]Biomedical Imaging and Radiation Technology Laboratory (BIRTLab), Department of Radiation Oncology, UT Southwestern Medical Center, Dallas, TX; [2]School of Computer Science, University of Birmingham, Edgbaston, Birmingham, UK

Purpose: Bioluminescence tomography (BLT) enhances commonly-used 2D bioluminescence imaging (BLI) by reconstructing 3D target distributions within tissue, potentially achieving tumor localization and volume estimation—critical for cancer therapy development. However, conventional model-based BLT reconstruction methods struggle with their inherent nature of ill-posedness and noises embedded in measurement data, hampering accurate estimates in tumor's location and geometrical distribution. We innovate a Self-supervised Hybrid Neural Network (SHNN) combining the strengths of model-based methods with machine learning to tackle the challenges and achieve quantitative optical imaging for cancer research.

Method: SHNN offers a comprehensive strategy to achieve quantitative BLT reconstruction. First, the feed-forward neural network is used to approximate the target function (light intensity distribution of tumor) and is imposed by the properties of the target function directly. Second, we propose the converging path on guiding SHNN toward convergence by dividing the optimization process into steps to reach an optimal solution, shown in **Fig. 1**. Both features make SHNN constrain the solution space to a reasonable subset, which mitigates the effect of ill-posedness and improves the robustness to the data noises significantly. Third, the SHNN is specifically designed to the scenarios without training data, commonly encountered in pre-clinical studies, by adopting self-supervised learning, which avoids the common issue of overfitting in machine learning.

Results: The orthotopic glioblastoma (GBM) mouse brain model is used to test the performance of SHNN. In numerical simulations, SHNN outperforms the conventional approach, spectral derivative compressive sensing conjugated gradient (SD-CSCG) method, in accuracy of both tumor localization and size, particularly under strong noise. Moreover, SHNN is capable of accurately reconstructing multiple tumors where SD-CSCG fails. More impressive, our in vivo results indicate that the SHNN can differentiate tumors varying in size, whereas the traditional method is insensitive to such variations.

Conclusions: SHNN demonstrates its high accuracy in solving the strongly ill-posed problem and robustness to data noises, which leads to functionalities of quantitative BLT reconstruction and multi-tumor differentiation.

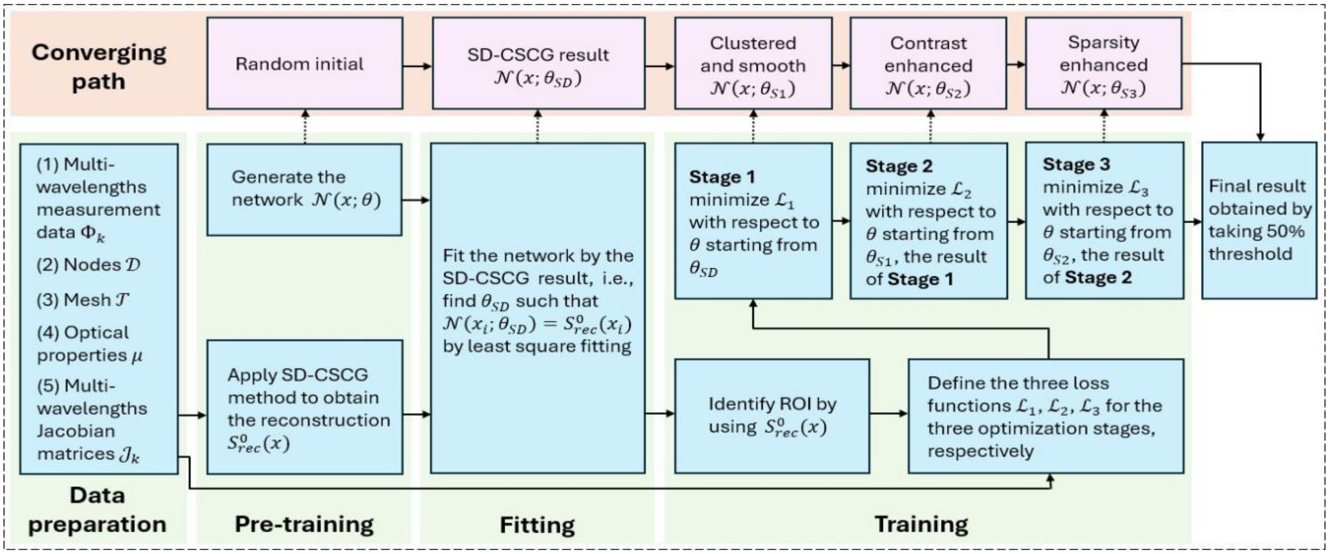

Converging path
Random initial
SD-CSCG result $\mathcal{N}(x;\theta_{SD})$
Clustered and smooth $\mathcal{N}(x;\theta_{S1})$
Contrast enhanced $\mathcal{N}(x;\theta_{S2})$
Sparsity enhanced $\mathcal{N}(x;\theta_{S3})$
(1) Multi-wavelengths measurement data Φ_k
(2) Nodes $\mathcal{D}$
(3) Mesh $\mathcal{T}$
(4) Optical properties μ
(5) Multi-wavelengths Jacobian matrices $\mathcal{J}_k$
Generate the network $\mathcal{N}(x;\theta)$
Apply SD-CSCG method to obtain the reconstruction $S_{rec}^0(x)$
Fit the network by the SD-CSCG result, i.e., find θ_{SD} such that $\mathcal{N}(x_i;\theta_{SD}) = S_{rec}^0(x_i)$ by least square fitting
Stage 1 minimize $\mathcal{L}_1$ with respect to θ starting from θ_{SD}
Stage 2 minimize $\mathcal{L}_2$ with respect to θ starting from θ_{S1}, the result of Stage 1
Stage 3 minimize $\mathcal{L}_3$ with respect to θ starting from θ_{S2}, the result of Stage 2
Final result obtained by taking 50% threshold
Identify ROI by using $S_{rec}^0(x)$
Define the three loss functions $\mathcal{L}_1$, $\mathcal{L}_2$, $\mathcal{L}_3$ for the three optimization stages, respectively
Data preparation
Pre-training
Fitting
Training

Fig. 1 shows the details of SHNN's workflow (green) aligning with the converging path to be traced (orange), which illustrates the essential idea of the optimization strategy.

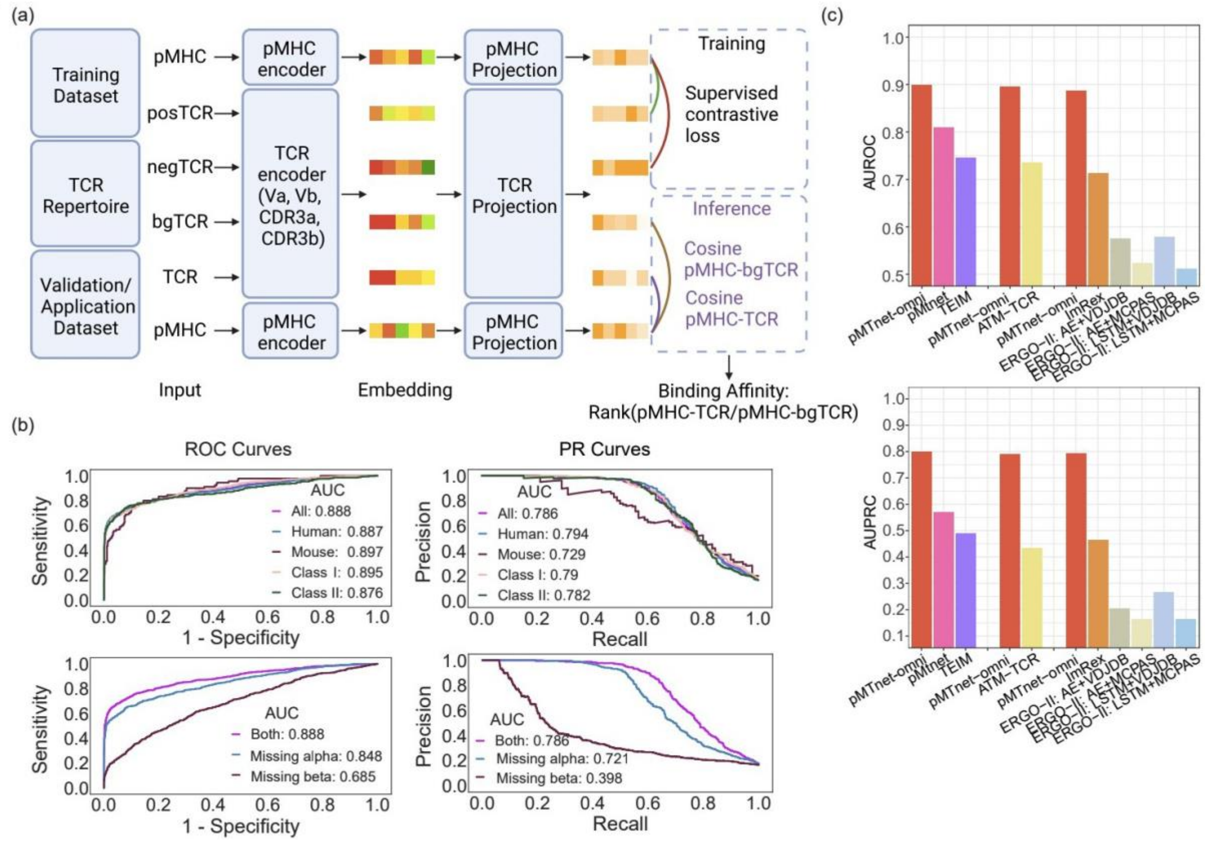
T-Cell Reactivity Biomarkers Empowered by a T-Cell Receptor-Antigen Foundation Model

Yi Han[1,†], Yuqiu Yang[1,†], Yanhua Tian[2,†], Jianjun Zhang[2,*], David E. Gerber[3,4,*], Tao Wang[1,*]

[1]Quantitative Biomedical Research Center, Peter O'Donnell Jr. School of Public Health, UT Southwestern Medical Center, Dallas, TX; [2]Department of Thoracic/Head and Neck Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX; [3]Harold C. Simmons Comprehensive Cancer Center, UT Southwestern Medical Center, Dallas, TX; [4]Department of Internal Medicine, UT Southwestern Medical Center, Dallas, TX

*Corresponding author; †Contributed equally

T-cell-mediated immunity plays key roles in normal development and disease physiology through T-cell receptor (TCR)-dependent activities. It has been challenging to monitor and understand the relevance of the thousands of TCRs in an individual, which could contain important clues for diagnostics, prognostics, and therapeutics. To address this significant gap, we developed a foundation Artificial Intelligence (AI) model, pMTnet-omni, to model the binding of TCRs and pMHCs (peptide-MHC complexes, namely antigens) for both class I and II pMHCs and for both human and mouse pMHCs. This foundation model enabled us to build a suite of biomarkers for assessing the longitudinal evolution and spatial heterogeneity of TCR repertoires, in the context of reactivities towards auto-, viral, and tumor antigens. We tested this method in three clinical scenarios. We developed the first TCR-based biomarker for predicting toxicities of immune checkpoint inhibitor (ICI) treatment in a cohort of 57 ICI-treated patients. Second, the pMTnet-omni biomarker uncovered dominating T- cell responses during COVID-19 infections and vaccination. Finally, this biomarker approach revealed the co-localization of tumor-associated antigen (TAA)-targeting TCRs with TAA-expressing cells in spatial-TCR-sequencing data. Overall, our work provides an expandable and flexible toolkit for easy and effective mapping of the functional relevance of the TCR repertoires for their reactivities towards any given antigens.



Poster #2491 - Early Identification of Patients at Risk for Iron Deficiency Anemia Leveraging Deep Learning

Estefanie Garduno-Rapp, M.D., M.S.H.I. [1]; Yee S. Ng, M.D. [1]; Jenny L. Weon, M.D., Ph.D. [1]; Sameh N. Saleh, M.D., M.B.M.I. [2]; Christoph U. Lehmann, M.D. [1]; Andrew Quinn, M.D. [1]

[1]UT Southwestern Medical Center, Dallas, TX; [2]University of Pennsylvania, Philadelphia, PA

Introduction: The diagnosis of Iron Deficiency Anemia (IDA) can be established with laboratory tests, but usually is only made after a patient becomes symptomatic. Up to 10% of adult patients with incidental IDA findings may have gastrointestinal cancer. The insidious presentation of IDA can cause a significant treatment delay. With advancement in machine learning (ML) and predictive algorithms, we hypothesized that we could reduce the delay to diagnosis by developing an IDA prediction model.

Objective: To develop and compare three predictive deep learning (DL) models using retrospective outpatient laboratory data to identify patients at risk for developing IDA as much as 3-6 months prior to traditional iron studies establishing the diagnosis.

Methods: We analyzed retrospective outpatient electronic health record (EHR) data between 2009 and 2020 from a tertiary care academic medical center in North Texas. We included laboratory features from 30,603 patients to develop four deep learning models including Artificial Neural Networks (ANNs), Long Short- Term Memory (LSTM), Gated Recurrent Unit (GRU), and Transformers using the PyTorch framework. The classifiers were trained using the Adam optimizer across 200 random training-validations splits. We calculated Accuracy (ACC), Area Under the Receiving Operating Curve (AUROC), Sensitivity (SE), and Specificity (SP) in the testing split.

Results: The GRU network outperformed the other models with a final ACC of 0.83, AUROC of 0.89, SE 0.75 and SP of 0.85 across 200 epochs.

Conclusion: Our results demonstrate the feasibility of employing DL techniques to analyze laboratory data for early prediction of patients who will eventually be diagnosed with IDA. Our findings suggest that, although all models achieved similar performance, the GRU model exhibited the highest AUROC (0.89), indicating superior discriminative power. Furthermore, the GRU model demonstrated the highest sensitivity (SE) of 0.75 and specificity (SP) of 0.85, highlighting its effectiveness in identifying positive and negative cases when compared to the other models.

Poster #2510 - WaVNet-Refine: A Joint Framework of an Adaptive WaVNet Model and a Refine Model for Test-Time Adaptation in Medical Image Segmentation

Xiaoxue Qian

UT Southwestern Medical Center, Dallas, TX

Purpose: Medical image segmentation often faces domain gaps between training and testing datasets due to variations in imaging hardware and protocols, leading to performance degradations in deep learning-based segmentation models. We developed a joint framework incorporating an adaptive WaVNet model and a Refine model (WaVNet-Refine) to achieve label-free test-time domain adaptation.

Materials/Methods: WaVNet-Refine has two components: an adaptive WaVNet segmentation model and a subsequent segmentation Refine model. For WaVNet, we introduced multilevel wavelet transforms into a VNet backbone, to improve its robustness against domain variations. Specifically, we calculated multiscale wavelet coefficients from to-be-segmented images; and concatenated selected frequency bands to the VNet encoder at different scales to enable domain-invariant feature extraction and learning. To adapt the WaVNet during test time, we developed a shallow adaptive module positioned at the front of the WaVNet, which helps to further address the domain gaps in testing samples. Following the adaptive WaVNet, we employed a Refine model (trained on noisy/incomplete and reference segmentation pairs) to evaluate/correct the segmentation results. The differences in segmentation results before and after the Refine model enable the generation of an unsupervised loss function to fine-tune the adaptive module of WaVNet during test time.

We evaluated WaVNet-Refine with a multi-site prostate segmentation dataset of 148 T2-weighted MRI sets from seven different sources. WaVNet-Refine was trained on one source domain and applied to the other six domains for test-time adaptation. Different state-of-the-art test-time adaptation methods were also evaluated for comparison.

Results: WaVNet-Refine showed the highest segmentation accuracy among all methods, achieving mean(±s.d.) Dice coefficient (DSC) of 79.36±5.59% and 95th Hausdorff distance (HD95) of 9.1±2.9 mm on six different target domains. In comparison, the no-adaptation baseline method yielded 64.75±11.61% DSC and 39.5±18.6 mm HD95.

Conclusion: WaVNet-Refine achieved significantly higher accuracy compared with other test-time adaptation methods, allowing robust and label-free cross-domain segmentation.

Poster #2515 - Medical Image Segmentation Assisted with Clinical Inputs via Language Encoder in a Deep Learning Framework

Hengrui Zhao, Biling Wang, Deepkumar Mistry, Jing Wang, Michael Dohopolski, Daniel Yang, Weiguo Lu, Steve Jiang*, Dan Nguyen*

Medical Artificial Intelligence and Automation (MAIA) Laboratory and Department of Radiation Oncology, UT Southwestern Medical Center, Dallas, TX
*Corresponding authors

Objective: Current auto-segmentation methods frequently demonstrate limited accuracy in delineating clinical target volumes (CTVs) and some challenging organs at risk (OARs), necessitating time-consuming manual corrections by clinicians. Clinicians typically depend on additional information beyond the imaging data for precise contouring of CTVs and some OARs. In contrast, existing auto-segmentation models primarily focus on the image content alone. This work is aimed at improving the accuracy of auto-segmentation by incorporating information beyond mere imaging data.

Methods: We extracted clinical data from Epic systems, which includes physician notes, medical images and medical bills, then converted it into text with meaningful clinical implications. We used CLIP text encoder to transform these texts into distinctive latent space vectors, enhancing the feature scaling within the U-Net segmentation model. This integration allows the incorporation of multi-modal clinical inputs to guide the image-based auto-segmentation, thereby elevating the segmentation quality substantially.

Results*:* We tested our framework against five factors (CT slice thickness, IV contrast, spacer hydrogel, MRI, physician) that potentially alter the segmentation of prostate in radiation therapy. Our method has a Dice score of 86.4% while the baseline model which does not utilize non-image clinical data has a Dice score of 84.4%.

Conclusion: By integrating non-imaging clinical data into the auto-segmentation process, our model demonstrates a noticeable improvement in segmentation accuracy compared to models using imaging data alone. This serves as a preliminary example to test our concept, and we plan to expand this framework for more complex CTV segmentation tasks in the future.

Poster #2521 - Motion-Resolved Magnetic Resonance Fingerprinting Using Low-Rank Spatiotemporal Implicit Neural Representation (LR-STINR)

Yang Li, Hua-Chieh Shao, Jie Deng, You Zhang

Medical Artificial Intelligence and Automation (MAIA) Lab, Department of Radiation Oncology, UT Southwestern Medical Center, Dallas, TX

Purpose: Magnetic resonance fingerprinting (MRF) allows simultaneous non-invasive quantification of multiple tissue properties. Nevertheless, physiological motion, especially respiratory motion, introduces artifacts and reduces the MRF accuracy. We combined temporal-domain MRF low-rank approximation with a joint reconstruction and motion- compensation strategy (LR-STINR), to achieve high-resolution, motion-resolved MRF.

Materials/Methods: LR-STINR addresses temporal-domain MRF contrast variations via low-rank approximation and the concurrent anatomical motion via a deformable motion model. It uses a spatial implicit neural representation (INR) network to reconstruct a motion-free reference MRF sequence, and a temporal INR to solve intra-scan dynamic motion as deformation-vector-fields (DVFs). The spatial INR maps spatial coordinates to rank-specific coefficients for dictionary-derived low-rank vectors, while the temporal INR reconstructs DVFs using a B- spline motion model learned on the fly. Evaluation was performed on the extended-cardiac-torso (XCAT) phantom with multiple motion scenarios. MRF acquisitions were simulated with the FISP sequence of 13s acquisition time, under a variable-density spiral k-space trajectory. The MRF quantification and motion-tracking accuracy was assessed.

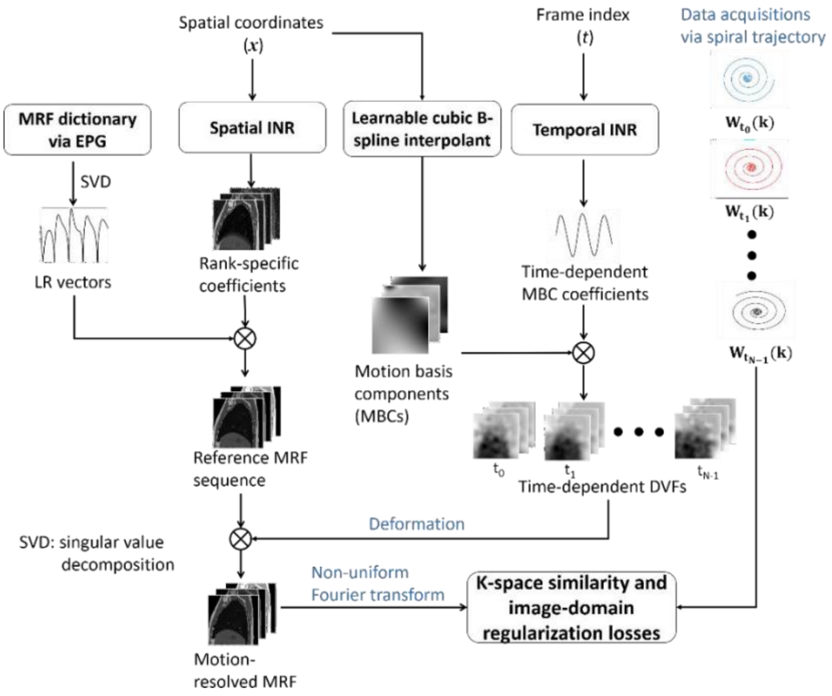


Figure 1. *The workflow of LR- STINR.*

Results: LR-STINR resolves motion-resolved MRFs (one image per spiral spoke: ~13 milliseconds) and tracks the liver tumor motion to an accuracy level of 0.86±0.46mm. It quantifies tissue properties accurately, with mean(±S.D.) absolute percentage errors of 1.77±0.20% and 2.39±0.40% for liver T1 and T2 values, respectively, compared to 7.00±2.60% and 17.02±7.22% for non-motion-resolved reconstructions.

Conclusion: LR-STINR can reconstruct high-quality motion-resolved MRF sequences of regular/irregular motion. It is a one-shot learning technique requiring no prior-training and not affected by generalizability issues.

Poster #2522 - Deep Convolutional LSTM Model for Prediction of Pathological Complete Response to Neoadjuvant Chemotherapy for Breast Cancer Using Multi-Time Point DCE MRI and Clinical Data with Uncertainty Quantification

Bowen Jing, Kai Wang, Erich Schmitz, Shanshan Tang, Yunxiang Li, You Zhang, Jing Wang

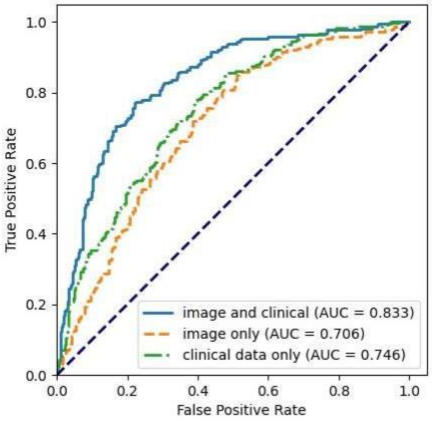
MAIA Lab, Department of Radiation Oncology, UT Southwestern Medical Center, Dallas, TX

Purpose: Early prediction of pathological complete response (pCR) to neoadjuvant chemotherapy for breast cancer patients can enable personalized treatment management for improved outcomes. In this study, we developed a convolutional long short-term memory (LSTM) network to predict pCR in breast cancer patients using dynamic contrast enhanced MR images (DCEMRI) and clinical data from the multi-institutional I-SPY2 clinical trial. Furthermore, prediction uncertainty was estimated for each patient to enable clinicians to prioritize or disregard predictions based on their associated uncertainties.

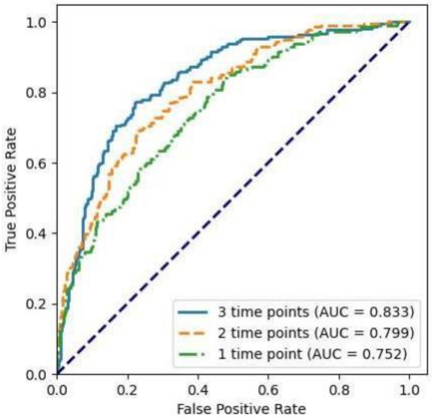
Methods: The dataset included 624 patients with DCEMRI acquired at 3 time points: pretreatment (T0), after 3 cycles of treatment (T1), and after 12 cycles of treatment (T2). A convolutional residual network (ResNet18) was used to learn image representations from multi-phase DCEMRI acquired at each time point. Subsequently, a long short- term memory network was used to predict pCR using the sequential image representations along with the clinical data. Aleatoric uncertainties were estimated using test-time augmentation.

Results: The top-performing model, with an area under the receiver operating characteristic curve (AUROC) of 0.833, was constructed using both clinical data and DCEMRI acquired at the 3 time points before treatment completion (Fig. a). AUROC increased to 0.853 after excluding highly uncertain predictions (~20% of the cohort). Utilizing the images from the first 2 time points along with clinical data yielded an AUROC of 0.799 (Fig. b). When solely clinical data were used, the AUC was 0.746 while using images from all 3 time points alone resulted in an AUROC of 0.706 (Fig. a).

Conclusion: We developed a convolutional LSTM model to predict pCR to neoadjuvant chemotherapy before treatment completion. The reliability of the prediction can be evaluated by measuring the uncertainty via test-time augmentation. By combining clinical data and multi-time point deep image representations, our model outperforms other models built solely on clinical or image data.



(a) ROC curves of prediction models using image data, clinical data or both.



(b) ROC curves of prediction models using images from 3, 2 or 1 time points and clinical data.

Poster #2525 - Mapping Cellular Interactions from Spatially Resolved Transcriptomics Data

James Zhu[1,*], Yunguan Wang[1,2,3,*], Woo Yong Chang[1,*], Alicia Malewska[4], Fabiana Napolitano[5], Jeffrey C. Gahan[4], Nisha Unni[6], Min Zhao[7], Rongqing Yuan[1], Fangjiang Wu[1], Lauren Yue[1], Lei Guo[1], Zhuo Zhao[8], Danny Z. Chen[8], Raquibul Hannan[9], Siyuan Zhang[7], Guanghua Xiao[1,10], Ping Mu[11,12], Ariella B. Hanker[5], Douglas Strand[4], Carlos L. Arteaga[5], Neil Desai[9], Xinlei Wang[13,14,+], Yang Xie[1,10,+], Tao Wang[1,+]

[1]Quantitative Biomedical Research Center, Peter O'Donnell Jr. School of Public Health, UT Southwestern; [2]Division of Pediatric Gastroenterology, Hepatology and Nutrition, Cincinnati Children's Hospital Medical Center; [3]Department of Pediatrics, University of Cincinnati; [4]Department of Urology, UT Southwestern; [5]Harold C. Simmons Comprehensive Cancer Center, UT Southwestern; [6]Department of Internal Medicine, UT Southwestern; [7]Department of Pathology, UT Southwestern; [8]Department of Computer Science and Engineering, University of Notre Dame; [9]Department of Radiation Oncology, UT Southwestern; [10]Department of Bioinformatics, UT Southwestern; [11]Department of Molecular Biology, UT Southwestern; [12]Hamon Center for Regenerative Science and Medicine, UT Southwestern; [13]Department of Mathematics, University of Texas at Arlington; [14]Center for Data Science Research and Education, College of Science, University of Texas at Arlington
*Co-first authors; [+]Corresponding authors

Abstract: Cell-cell communication (CCC) is essential to how life forms and functions. However, accurate, high-throughput mapping of how expression of all genes in one cell affects expression of all genes in another cell is made possible only recently through the introduction of spatially resolved transcriptomics technologies (SRTs), especially those that achieve single cell resolution. Nevertheless, significant challenges remain to analyze such highly complex data properly. Here, we introduce a Bayesian multi-instance learning framework, Spacia, to detect CCCs from data generated by SRTs, by uniquely exploiting their spatial modality. We highlight Spacia's power to overcome fundamental limitations of popular analytical tools for inference of CCCs, including losing single-cell resolution, limited to ligand-receptor relationships and prior interaction databases, high false positive rates, and most importantly, the lack of consideration of the multiple-sender-to-one-receiver paradigm. We evaluated the fitness of Spacia for all three commercialized single cell resolution ST technologies: MERSCOPE/Vizgen, CosMx/Nanostring, and Xenium/10X. Spacia unveiled how endothelial cells, fibroblasts, and B cells in the tumor microenvironment contribute to Epithelial-Mesenchymal Transition and lineage plasticity in prostate cancer cells. We deployed Spacia in a set of pan-cancer datasets and showed that B cells also participate in *PDL1*/*PD1* signaling in tumors. We demonstrated that a CD8+ T cell/*PDL1* effectiveness signature derived from Spacia analyses is associated with patient survival and response to immune checkpoint inhibitor treatments in 3,354 patients. We also revealed differential spatial interaction patterns between γδ T cells and liver hepatocytes in healthy and cancerous contexts. Overall, Spacia represents a notable step in advancing quantitative theories of cellular communications.

Poster: #2527 - Cmai: Predicting Antigen-Antibody Interactions from Massive Sequencing DataBing

Bing Song[1], Kaiwen Wang[2], Saiyang Na[3], Jia Yao[1], Junzhou Huang[3], Tao Wang[1]

[1]Quantitative Biomedical Research Center, Peter O'Donnell Jr. School of Public Health, UT Southwestern Medical Center, Dallas, TX; [2]Department of Statistics and Data Science, Southern Methodist University, Dallas, TX; [3]Department of Computer Science and Engineering, the University of Texas at Arlington, Arlington, TX

ABSTRACT: The interaction between antigens and antibodies is the key step underlying the function of the humoral immune system in various biological contexts. However, current experimental approaches for profiling antibody-antigen interactions are costly and time-consuming, and can only achieve low-to-mid throughput. On the other hand, bioinformatics tools in the field of antibody informatics mostly focus on optimization of baseline antibodies given known binding antigens, which is a related but very different research question. In this work, we developed an Artificial Intelligence tool, Cmai, to address the prediction of the binding between antibodies and antigens. Cmai achieved an AUROC of 0.91 in our validation cohort and can distinguish amino acids on the antibodies and antigens that are in contact. In a cohort of patients on immune- checkpoint inhibitor (ICI) treatments, Cmai found that the elevation of auto-antibodies binding to intracellular auto-antigens during immune-related adverse events (irAEs) is positively correlated with the expression levels of the auto-antigens in the organs affected by the irAEs. In contrast, we found that the binding of antibodies towards extracellular tumor antigens in cancer patients is correlated with the expression of these antigens. We further found that the abundance of these tumor antigen-targeting antibodies is predictive of survival and ICI treatment response. Lastly, we showed that B cells targeting tumor antigens demonstrate spatial co-localization patterns with tumor cells expressing these antigens. Overall, this work provided a powerful tool to profile the landscape of antigen-antibody interactions in biological samples, which reveals novel insights at unprecedented levels and will yield powerful tools for translational development.

Poster #2529 - Real-Time CBCT Imaging via a Single Arbitrarily-Angled X-Ray Projection Using a Joint Dynamic Reconstruction and Motion Estimation (DREME) Framework

Hua-Chieh Shao[1], Tielige Mengke[1], Tinsu Pan[2], You Zhang[1]

[1]Medical Artificial Intelligence and Automation (MAIA) Lab, Department of Radiation Oncology, UT Southwestern Medical Center, Dallas, TX; [2]Department of Imaging Physics, The University of Texas MD Anderson Cancer Center, Houston, TX

Purpose: Real-time cone-beam computed tomography (CBCT) provides instantaneous patient anatomy for image guidance and online treatment adaptation in radiotherapy. However, real-time CBCT imaging faces a stringent temporal constraint (<500 milliseconds) under respiration-induced anatomical motion (~3-5 seconds per cycle), leading to extreme under-sampling (one or a few X-ray projections). We combined a dynamic CBCT reconstruction method (PMF-STINR) with a deep learning (DL)-based, real-time motion estimation model to form a joint dynamic reconstruction and motion estimation (DREME) framework for real-time motion and CBCT estimation.

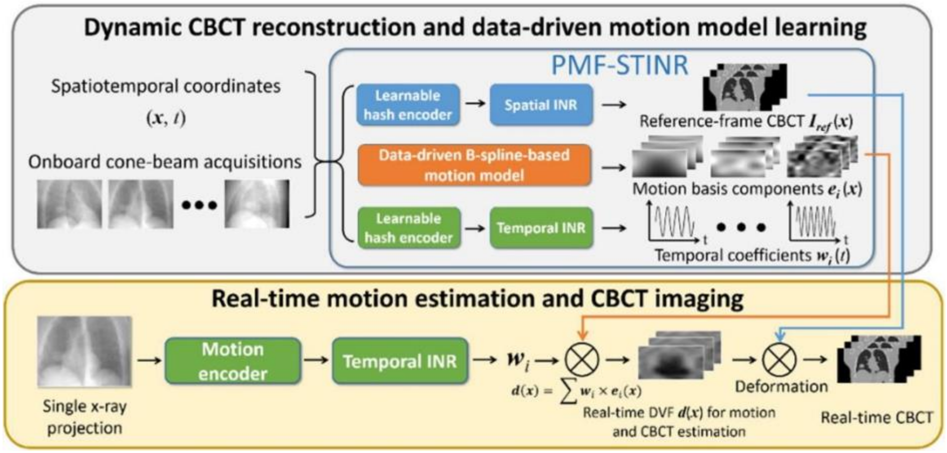

Dynamic CBCT reconstruction and data-driven motion model learning
PMF-STINR
Spatiotemporal coordinates
(x, t)
Onboard cone-beam acquisitions
Learnable hash encoder
Spatial INR
Reference-frame CBCT $I_{ref}(x)$
Data-driven B-spline-based motion model
Motion basis components $e_i(x)$
Learnable hash encoder
Temporal INR
t
t
Temporal coefficients $w_i(t)$
Real-time motion estimation and CBCT imaging
Single x-ray projection
Motion encoder
Temporal INR
w_i
$d(x) = \sum w_i \times e_i(x)$
Real-time DVF $d(x)$ for motion and CBCT estimation
Deformation
Real-time CBCT

***Figure 1**: Workflow of the DREME framework.*

Methods: DREME reconstructs a dynamic CBCT sequence from a pre-treatment CBCT scan, while simultaneously deriving a DL-based, angle-agnostic model for real-time motion and CBCT estimation (Fig. 1). Specifically, DREME uses a prior model-free spatiotemporal implicit neural representation (PMF-STINR) framework to reconstruct dynamic CBCTs and derive a data-driven motion model. The motion model, capturing the up-to-date motion patterns, is subsequently used by DREME to train a DL model. The DL model applies an angle-agnostic encoder to extract motion-related features from a projection and infers corresponding coefficients for the PMF-STINR motion model to construct real-time motion field. The motion field can then morph a reference CBCT volume derived by PMF-STINR into real-time, motion- resolved CBCTs.

Results: DREME can accurately solve 3D motion in real time. A phantom study achieved a mean (±S.D.) center-of-mass-error of 1.0±0.4 mm, and the corresponding result was 1.3±1.2 mm for a patient study. The mean inference time is 1.5 ms per projection, fulfilling the temporal constraint of real-time imaging.

Conclusion: DREME uses the latest onboard CBCT and motion model information, extracted from dynamic CBCT reconstructions, to solve accurate real-time motion and CBCTs from arbitrarily-angled X- ray projections, paving the way for future real-time adaptive radiotherapy.

Poster #2530 - A Comparative Analysis of AI-Generated Practice Questions for USMLE Step 1 Preparation

Samantha Little, B.S., M.S., M.P.H.; Jack Nickles, B.S.; Natasha Gengler, B.S.

Dell Medical School, The University of Texas at Austin

Objective: This study sought to compare the quality and accuracy of USMLE Step 1 practice questions from UWorld to questions generated by ChatGPT 3.5. The goal was to determine whether AI platforms could serve as a viable and cost-effective alternative for USMLE Step 1 preparation, potentially reducing reliance on commercial preparatory materials.

Methods: A quantitative analysis was conducted utilizing USMLE question criteria from the National Board of Medical Examiners (NBME) Item-Writing Guide. Raters determined accuracy by quantifying the number of errors in each question and answer explanation from each source.

Results: ChatGPT-generated questions were more likely to contain inaccuracies ($p < 0.05$), with 45% (9 of 20) of questions being inaccurate and 33% (3 of 9) of those questions containing multiple inaccuracies. ChatGPT-generated questions contained more words than UWorld questions ($p < 0.05$), while ChatGPT was more likely than UWorld to focus on recall rather than application in question format ($p < 0.05$). In addition, ChatGPT and UWorld had a different distribution of first-, second-, and third-order questions ($p < 0.05$), with ChatGPT only producing first- and second-order questions while 45% (9 of 20) of UWorld questions were third-order. However, there was no significant difference in the interpretation of labs/values ($p = 0.32$), the presence of homogenous answers ($p = 0.49$), or the cover of option-style questions ($p = 0.11$). ChatGPT and UWorld also had no significant difference in the amount of NBME elements mentioned in each question ($p = 0.64$).

Conclusion: Our data illustrates that ChatGPT 3.5 currently does not replicate the accuracy and ability to generate complex questions essential for effective USMLE Step 1 examination preparation. As a result, it does not offer a viable alternative to the comprehensive preparatory resources provided by professional platforms such as UWorld.

Poster #2535 - Off-the-Shelf Segmentation Networks as a Tool for Dose-Level Prediction

Qingying Wang[1], Mingli Chen[1], Mahdieh Kazemimoghadam[1], Xuejun Gu[1,2,*], Weiguo Lu[1,*]

[1]Department of Radiation Oncology, UT Southwestern Medical Center, Dallas, TX; [2]Department of Radiation Oncology, Stanford University, Stanford, CA
[*]Corresponding authors

Purpose: Deep convolutional neural networks are prominent in dose prediction tasks, while most studies have attempted to customize network architectures for different diseases and treatment modalities. We propose a universal strategy to predict dose levels with a state-of-the-art segmentation-based network that demonstrates a convenient off-the-shelf application.

Methods: Rather than focusing on the network architecture and learning algorithm, we propose a simple form of inputs, which adopts only two channels independent of the number of organ-at-risks (OARs): 1) the normalized prescription dose, and 2) an avoidance mask with values of 1 for OARs, 0.1 for body, and 0 for the rest. These two channels correspond to the direct clinical goals, the prescription and the avoidance. The outputs are predicted 3D dose level maps with the number of channels defining resolution and can be smoothed by Gaussian kernels in post-processing. A feasibility study was conducted using 3D nn-UNet on 114 clinical Gamma Pod breast cancer cases, which were randomly split into 76/19/19 for training/validation/testing.

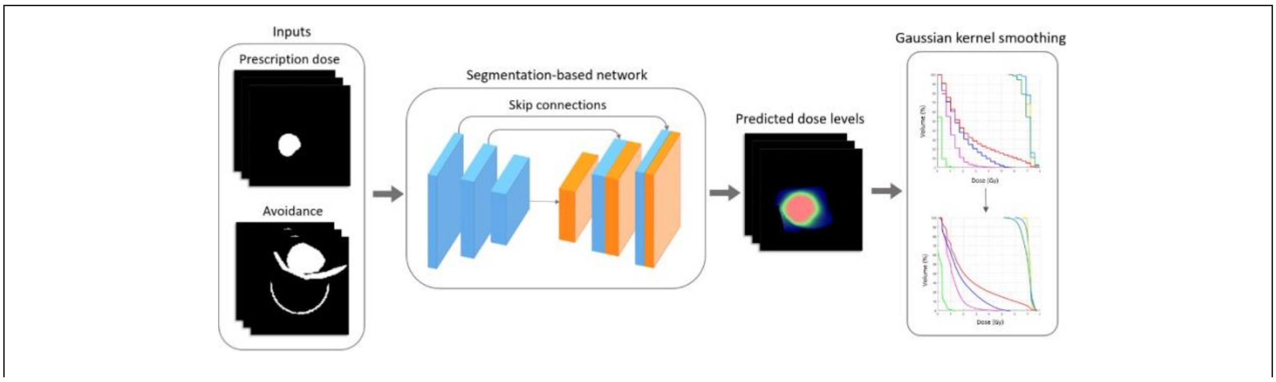


Figure 1: The overview of predicting the dose level by using the segmentation-based network.

Results: The prediction had an average Dice similarity coefficient of 0.89±0.01 compared to clinical isodose volumes between 0% and 120% of the prescribed dose at 5% intervals and a comparable hotspot control (D_5) to clinical plans for PTV and CTV. For gamma analysis with the 3%/3 mm criteria, 79% test cases passed the 90% gamma passing rate (GPR) with average GPR 94.68%±3.16%, while the unpassed cases with the lowest GPR 83% were due to the dose voxels located in the external body or low-dose wash region, which has little impact on plan quality.

Conclusion: This study confirms the feasibility of leveraging the segmentation model for dose prediction tasks, demonstrating the application of an off-the-shelf tool for treatment planning guidance. It can be built as a quick tool for treatment modality selection whenever the patient images are available.

Poster #2536 - AI-Driven Adaptive Radiotherapy Target Segmentation: A Follow-the-Leader Approach

Mahdieh Kazemimoghadam[1], Qingying Wang[1], Mingli Chen[1], Xuejun Gu[1, 2,*], Weiguo Lu[1,*]

[1]Department of Radiation Oncology, UT Southwestern Medical Center, Dallas TX; [2]Department of Radiation Oncology, Stanford University, Stanford, CA
[*]Corresponding authors

Purpose: Adaptive Radiotherapy (ART) requires efficient target delineation for real-time adjustments. Yet, manual contouring's variability, labor intensity, and time consumption hinder ART integration. This study introduces a novel template-guided deep learning model for ART target segmentation.

Methods: The proposed deep learning model was built upon a 3D nnUNet, featuring three input channels: primary CT image, template CT image, and template mask (GTV, CTV, or PTV). The structure allows “Follow-the-Leader" nature of learning in dual aspects (Figure 1). 1) The utilization of templates enables the model to dynamically adapt to various targets, generating precise segmentation based on the type of template selected. 2) The model leverages the stylistic insights from the template to effectively segment the new image. We utilized our in-house retrospective dataset of 114 post-operative breast patients in this study, who received 5-fraction partial breast irradiation (PBI). Patient data were randomly split into training (84), validation (15), and test (15). Within each patient's data, exhaustive exploration of all possible combinations of two fractions for primary CT and template across GTV, CTV, and PTV was performed. Hence, a total of 5040 (20 combinations × 84 patients × 3 targets) training samples were utilized. The model performance was compared against baseline models, each using a single - channel setup (CT image), for GTV, CTV, and PTV segmentation.

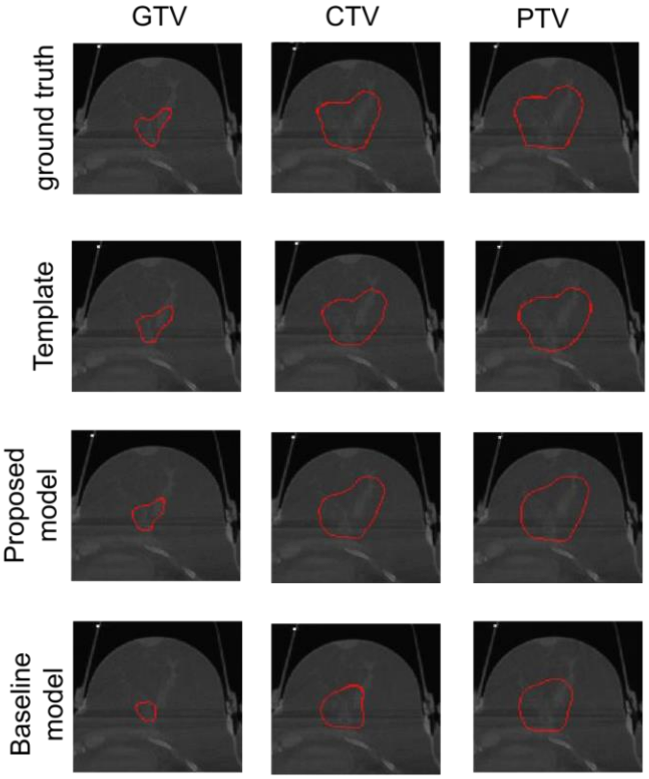


Figure 1. Visual comparison of segmentation results. Each row illustrates the ground truth mask, template, output of the proposed model, and output of the baseline model for GTV, CTV, and PTV, respectively.

Results: The model achieved Dice similarity coefficients (DSC) of 0.74 (±0.06), 0.86 (±0.04), and 0.87(±0.03) for GTV, CTV, and PTV respectively highlighting the versatility of the model in effectively handling diverse targets. The proposed model outperformed the baseline models (Figure 1) with DSC of 0.69 (±0.08), 0.81(±0.07), and 0.82(±0.05) for GTV, CTV, and PTV respectively, exhibiting notably higher accuracy.

Conclusion: The template-guided deep learning model in this study offers a novel approach to ART target segmentation. The model adaptability to chosen templates and utilization of stylistic cues are promising to improve ART workflow.

Poster #2537 - Deep-Learning- Based Spectral Noise Reduction with Synthetic Data for 7T Proton MRSI

Tianyu Wang, Yeison Rodriguez, Anke Henning, Ph.D.

Advanced Imaging Research Center (AIRC), UT Southwestern Medical Center, Dallas, TX

Purpose: Proton Magnetic Resonance Spectroscopic Images (MRSI) provide valuable clinical insights into tumor metabolism but are hindered by inherently low SNR due to metabolite's very low concentration and dense, overlapping resonances in spectra. We aim to leverage the success of deep learning in pattern recognition to reduce noise and perturbation generated during acquisition and thus improve proton spectra SNR for 7T 1H-MRSI data.

Methods: We synthesized 13203 proton spectra with varied metabolite combination and perturbations from simulated spectral basis-set (GAMMA) of 11 different metabolites for a 7T FID 1H-MRSI sequence. Each spectra included core metabolites (NAA, tCho, tCr) and random metabolites selected from [Asp, GABA, Glx, Gly, GSH, Leu, mI, Tau], totaling 7~11 metabolite signals. The analytical signal model posited additive gaussian noise (N(0,θ), θ=1~5% of max signal), and 3 parameters for spectral imperfections at 7T (line broadening 12-15Hz, frequency shifting +/- .1 PPM, and amplitude scaling 80%~120%). Dataset is split into 70% training and 30% validation. An auto-encoder (input/output dim: 3396, hidden layer: 36, Adam optimizer, L2 loss) was trained with 16 batch-size for 25 epochs using only the real part of the signal.

Results: The network effectively reduces noise and perturbations. Reconstructed spectra align closely with ground truth. Some over/under estimation and peak hallucination was observed. It achieved a mean squared error of 2.37e-05 and a peak signal-to-noise ratio of 46.7dB between the reconstructed and clean spectra, marking a 97.7% MSE reduction and a 53.9% PSNR increase over noisy inputs.

Conclusion: This study showed that an autoencoder with synthetic data effectively recovered data obscured by noise and perturbations. Future enhancements could focus on refining peak height accuracy and minimizing false peaks by integrating attention layers that emphasize adjacent inputs, thereby improving the network's understanding of correlations between neighboring peaks.

$$S(f) = \sum_{i=1}^{n} S_i(f) = \mathcal{F}\{g(t)\}, f' = f - \Delta f$$

$$S'(f') = \mathrm{c} \cdot \mathcal{F}\{g(t) \cdot e^{-\alpha t}\}$$

$$\mathrm{X} \ = \ \mathrm{S}' + N(0, \sigma^2), \sigma^2 = \ n\% \ * \ max(S(f))$$

$S(f)$: Clean ground truth, sum of n metabolite spectra.

$S'(f')$: Adjusted spectrum, shifted by Δf, scaled by c, damped by e^{-at}.

X: Observed noisy signal, $S'(f')$ with added Gaussian noise.

Fig 1.Analytical Model for Spectra Simulation

	MSE	PSNR
Input	1.03E-03	3.04E+01
Reconstructed	2.37E-05	4.67E+01
% Change	-97.70%	53.91%

Table 1. Performance Metrics for Reconstructed Spectra

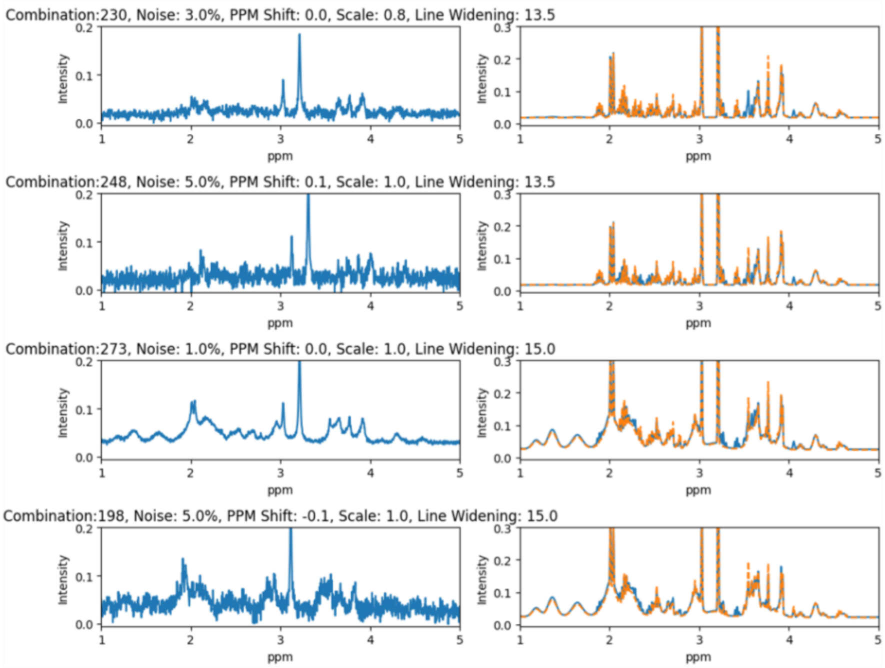


Fig 2. Noisy Input Spectra zoomed in 1~5 PPM region(Left), and Overlay of Ground Truth (Orange) over Reconstructed Spectra in the same region(Blue)

Poster #2538 - Natural Language Processing to Extract Acute Symptom Clusters from Triage Phone Notes with Cancer Patients

Yingzi Zhang, Ph.D., RN[1]; Arthur Hong, M.D.[1]; Kristine Kwekkeboom, Ph.D., RN, FAAN[2]

[1]UT Southwestern Medical Center, [2]University of Wisconsin-Madison School of Nursing

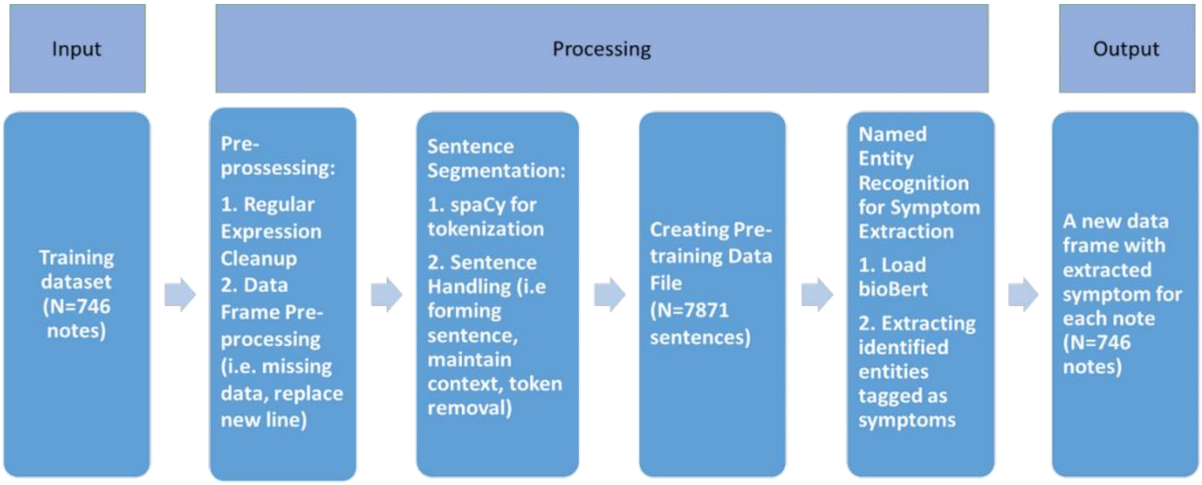
Objective: Patients with cancer undergoing active treatment frequently visit the emergency department (ED) for cancer- or treatment-related symptoms. Although we characterize patients according to a primary complaint, patients often have more than one relevant symptom. The triage conversations documented in the electronic health record provides a rich source of data to better describe the constellation of symptoms leading to an ED visit. The aim of this pilot study was to develop a natural language processing (NLP) methodology that abstracts clusters of symptoms from oncology triage clinical notes.

Methods: A sample of clinical notes (N=746) were used. A comprehensive summary of the data mining process is presented in Figure 1. First, we preprocessed clinical notes using spaCy, then extracted symptoms with a biomedical named entity recognition model, BioBERT. BioBERT is a biomedical language representative model designed for text mining tasks such as Named-entity recognition (NER). BioBERT is pre-trained on biomedical domain corpora including PubMed abstracts and PMC full-text articles. We implemented a model that was trained on a dataset consisting of clinical symptom data from hematological patients.

Results: We implemented the data mining process successfully on the training dataset. Per note, we extracted an average of 3 symptoms, with a maximum of 42 symptoms. Additionally, 27% of the notes contained five or more symptoms. The main issues encountered were false positive labeling (bioBert does not qualify symptoms preceded by "denies," or "not"); and mislabeled abbreviations (Title of "Ms." mislabeled as a symptom).

Conclusion: It is feasible to apply a pre-trained clinical NLP method to gather added richness to acute symptoms in free-text clinical notes. Additional pre-processing such as abbreviations expansion, spell correction, and stop words removal should be incorporated to address the limitations of BioBERT. We plan to compare model accuracy with human clinician coders, before using these symptoms clusters to predict severity of ED visits and likelihood of hospitalization.

Figure 1. Data mining process


Input
Processing
Output
Training dataset (N=746 notes)
Pre-prossessing:
1. Regular Expression Cleanup
2. Data Frame Pre-processing (i.e. missing data, replace new line)
Sentence Segmentation:
1. spaCy for tokenization
2. Sentence Handling (i.e forming sentence, maintain context, token removal)
Creating Pre-training Data File (N=7871 sentences)
Named Entity Recognition for Symptom Extraction
1. Load bioBert
2. Extracting identified entities tagged as symptoms
A new data frame with extracted symptom for each note (N=746 notes)

Poster #2541 - 3D Reconstruction of Spatial Transcriptomics with Spatial Pattern Enhanced Graph Convolutional Neural Network

Chen Tang[1], Lei Dong[1], Xue Xiao[1], Yuansheng Zhou[1], Lei Guo[1], Yunguan Wang[2], Qiwei Li[3], Guanghua Xiao[1], Lin Xu[1]

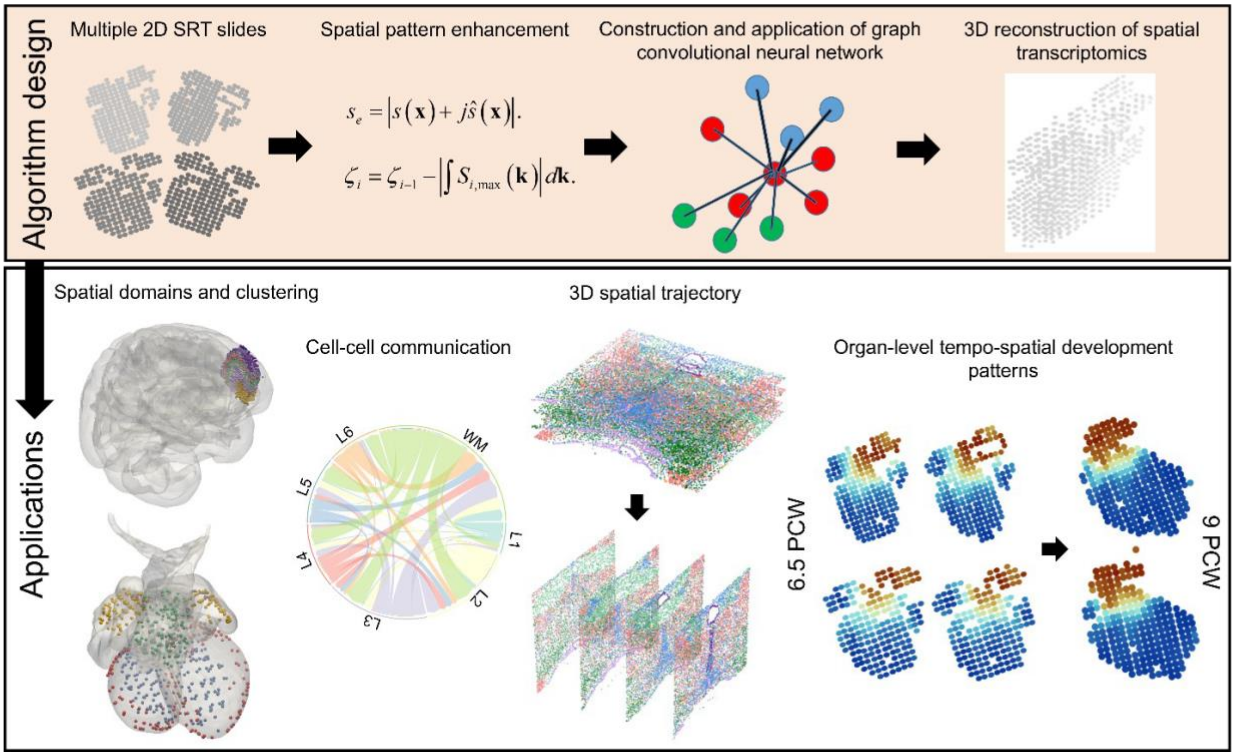
[1]UT Southwestern Medical Center, [2]Cincinnati Children's Hospital Medical Center, [3]The University of Texas at Dallas

Abstract: Spatially Resolved Transcriptomics (SRT) has emerged as a promising technology, enabling the simultaneous analysis of gene expression and spatial information in biomedical research. However, current statistical and deep learning algorithms for SRT data primarily rely on two-dimensional (2D) spatial coordinates. This dependence hinders their ability to accurately identify spatial domains, discern spatially variable genes, map cell-to-cell communications, and trace developmental trajectories in a three- dimensional (3D) spatial context. Addressing this, we've introduced Spa3D, a tool that integrates the anti- leakage Fourier transform and a graph convolutional neural network model to reconstruct 3D spatial structures from multiple 2D SRT slides.

As depicted in Figure 1 (top panel), Spa3D processes multiple 2D SRT slides to produce a detailed 3D spatial structure, seamlessly integrating gene expression and spatial dimensions. To enhance spatial patterns, Spa3D employs two methodologies: the instantaneous amplitude calculation using the Hilbert transform and a modified local anti-leakage Fourier transform for 2D SRT data. The transformed data then facilitates the creation of a 3D graph representation, emphasizing the actual physical distances between consecutive slides, ensuring a more accurate spatial depiction. This approach allows Spa3D to adeptly handle 3D spatial information from multiple SRT slices, even those with dissimilar patterns. Spa3D could yield 3D cellular organization and boast multiple pivotal applications, as showcased in Figure 1 (bottom panel).

When compared with PASTE, a leading method for integrating multiple 2D SRT slides, Spa3D consistently demonstrates its superiority. Across platforms such as 10X Genomics Visium, ST, and MERFISH, Spa3D excels in spatial domain identification, 3D cell-cell communication detection, and revealing 3D spatial trajectories that 2D coordinates might obscure. In conclusion, Spa3D marks a significant advancement in spatial transcriptomic analysis, offering nuanced 3D insights with vast potential for the future of biomedical research and clinical applications.

Figure 1


Algorithm design
Multiple 2D SRT slides
Spatial pattern enhancement
$s_e = \left| s(\mathbf{x}) + j\hat{s}(\mathbf{x}) \right|.$
$\zeta_i = \zeta_{i-1} - \left| \int S_{i,\max}(\mathbf{k}) \right| d\mathbf{k}.$
Construction and application of graph convolutional neural network
3D reconstruction of spatial transcriptomics
Applications
Spatial domains and clustering
Cell-cell communication
L6
WM
L5
L1
L4
L2
L3
3D spatial trajectory
Organ-level tempo-spatial development patterns
6.5 PCW
9 PCW

Poster #2545 - Harnessing Machine Learning for the Selection of Empiric Antibiotics in Urinary Tract Infections

Lauren N. Cooper, M.S.[1]; Alaina M. Beauchamp, Ph.D., M.P.H.[1]; Tanvi A. Ingle, B.S.[1]; Abdi D. Wakene, B.S.[1]; Marlon I. Diaz, B.S.[2]; Chaitanya Katterpalli, M.S.[3]; Tony Keller, B.A.[3]; Clark Walker, M.P.H.[3]; Alexander P. Radunsky, Sc.D., M.P.H.[1]; Zachary M. Most, M.D.[1]; John J. Hanna, M.D.[1,4]; Trish M. Perl, M.D.[1]; Christoph U. Lehmann, M.D.[1]; Richard J. Medford, M.D.[1,4,5]

[1]UT Southwestern Medical Center, Dallas, TX; [2]Paul L. Foster School of Medicine, Texas Tech University Health Sciences Center, El Paso, TX; [3]Texas Health Resources, Arlington, TX; [4]ECU Health, Greenville, NC; [5]Brody School of Medicine, East Carolina University, Greenville, NC

Introduction: The WHO declared that overly broad-spectrum antibiotic (BSAs) use contributes to the global antimicrobial resistance (AMR) crisis [1]. We developed machine learning (ML) models to predict presence and type of AMR bacteria in urine cultures obtained at a first patient encounter prior to the 24-48 hour period required for culture and susceptibility testing to reduce BSA use.

Methods: Using electronic health record (EHR) data from Texas Health Resources and the University of Texas Southwestern Medical Center, we linked urine cultures to four common AMR organisms (AmpC beta-lactamase, Extended spectrum beta-lactamase, Carbapenem resistant *Enterobacterales*, Vancomycin resistant *Enterococcus*) via susceptibility testing. Using geographic, demographic, and socioeconomic data, medical history, and comorbidities, we developed ML models to predict AMR organism presence in the initial urine culture and the most probable AMR organism. We leveraged three classifiers: penalized logistic regression (LR), random forest (RF), and extreme gradient boosting (XGB). We evaluated the models using the area under the receiver operating characteristic curve (AUC-ROC), also considering negative predictive value (NPV) and feature importances.

Results: Binary classifiers predicting presence of an AMR organism in a urine culture had AUC-ROC values of 0.66 for the LR model and 0.68 for the RF and XGB models. The NPVs were 0.961-0.964 for all models. Multi-class classifiers determining the AMR organism saw an increase in AUC-ROC values for all three model types (LR 0.69-0.76, RF 0.72-0.82, XGB 0.74 to 0.81) for all four organisms. For the RF model, factors like the Area Deprivation Index of the patient's address, race, ethnicity, and insurance status were the most important, highlighting the role of socioeconomics in the prediction of AMR, while the XGB model relied mainly on prior infections.

Discussion and Conclusions: With high NPVs, our models can predict the absence of AMR bacteria from EHR data accessible at the time of visit but can also predict the most probable organism when AMR bacteria are present in urine cultures. As decision support, our model would allow clinicians to avoid empiric BSAs, possibly reducing medical costs and future resistance [2]. Further, our work demonstrates the importance of socioeconomic factors in AMR.

References:

1. WHO. Antimicrobial Resistance. Available at: https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance. Accessed 04/17/2024.
2. O'Neill J. Tackling Drug-Resistant Infections Globally: Final Report and Recommendations: Government of the United Kingdom, 2016.

Poster #2547 - AI In-Basket Deep Dive: Who's Busy and Who's Breezy?

Jorge M. Rodriguez-Fernandez

The University of Texas Medical Branch

Introduction and objectives: Discover the differential utilization patterns between high and low utilizers of the AI InBasket draft reply feature. This presentation will delve into a comprehensive retrospective analysis performed before and after the release of AI-generated draft replies. By exploring signal metrics, this study aims to guide on identifying high and low adopters with the aim to develop targeted efforts in each subgroup to enhance utilization. Attendees will identify key signal metric trends and how they can be utilized to refine your department's operational strategies.

Methods: This retrospective analysis utilized median distribution to classify users into high and low utilizers of AI-generated draft replies. Through a detailed signal metric analysis covering inbasket time, appointments per day, and other operational metrics, we discerned clear distinctions in usage. High utilizers demonstrated significant efficiency in daily operations, indicated by metrics such as a higher average draft used percentage and shorter times spent in the system per day. The analysis, encompassing a six-month period before and after AI implementation, offers key insights into enhancing workflow efficiency and AI tool adoption.

Results: The primary outcomes highlight the stark differences between high and low utilizers in several key operational metrics post-AI implementation. High utilizers benefited from the AI features, showing improved efficiency in several areas, including reduced time spent on orders and system use per day. The detailed statistical analysis, with comparative statistics (next page) across subgroups will be presented, such underscores the impact of AI adoption on improving healthcare providers' operational efficiency.

Conclusions: Our retrospective study utilized a six-month pre- and post-implementation signal data analysis to demonstrate the real-world impact of AI on operational efficiency. High utilizers showed notable improvements in various metrics, including average response and draft lengths, indicating enhanced engagement and efficiency. The findings suggest significant operational differences pre-implementation and highlight the need for targeted strategies to boost AI adoption among lower utilizers.

This comprehensive analysis not only underscores the effective utilization of AI tools but also provides actionable insights for other Epic organizations aiming to enhance provider satisfaction and potentially patient care quality. Attendees will gain knowledge on the AI InBasket Epic tool, as well as how to use signal metrics for a strategic implementation.

The retrospective analysis delineates clear distinctions between high and low utilizers in several key metrics: High utilizers demonstrated an average draft used percentage of 34.60% (SD=28.74) compared to 1.49% (SD=2.35) for low utilizers. We observed significant differences in workflow efficiency, such as average response length (299.75 characters, SD=114.21 for high utilizers vs. 259.77 characters, SD=117.68 for low utilizers) and average draft length (428.35 characters, SD=86.76 for high utilizers vs. 420.25 characters, SD=61.68 for low utilizers).

Significant findings pre-implementation include:

- High utilizers had fewer appointments per day (mean=6.27, SD=2.55) than low utilizers (mean=8.24, SD=2.25), $p<0.001$.

- Scheduled hours per day were less for high utilizers (mean=3.28 hours, SD=1.26) compared to low utilizers (mean=4.12 hours, SD=0.95), p<0.001.
- High utilizers spent less time on orders per day (20 minutes, SD=12) vs. low utilizers (26 minutes, SD=15), p=0.049.
- Time in the system per day was lower for high utilizers (101 minutes, SD=59) compared to low utilizers (127 minutes, SD=49), p<0.001.
- Note character length was shorter for high utilizers (3001 characters, SD=1216) compared to low utilizers (3623 characters, SD=1317), p=0.029.
- The percent of appointments closed the same day differed significantly between groups, with high utilizers at 47% (SD=36) compared to low utilizers at 63% (SD=30), p=0.010.

Poster #2549 - Deep-Learning Analysis Uncovers Enhancer-Altering Mutations to Drive Genome-Wide Changes of TF Binding for Tumorigenesis

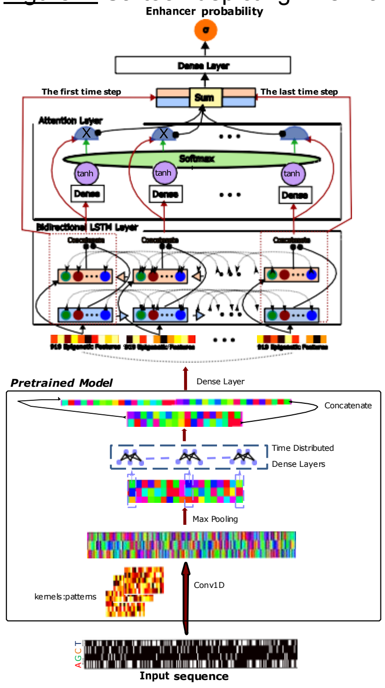
Xue Xiao[1], Lin Xu[1,2,*]

[1]Quantitative Biomedical Research Center, Peter O'Donnell Jr. School of Public Health, UT Southwestern Medical Center; [2]Department of Pediatrics, UT Southwestern
[*]Corresponding author

Abstract: Enhancers are non-coding regulatory DNA sequences that can be bound by proteins to control the target gene transcription. When noncoding mutations occur within enhancer sequences, they can disrupt the original pattern of the enhancer and alter the binding status of transcription factors (TFs). These mutations, referred to as "enhancer-altering mutations", can result in abnormal gene transcription and lead to various Mendelian disorders and common diseases such as cancer. Over the past decade, several computational methods have been developed for predicting enhancer activities. The majority of these methods require epigenomic profiles to build the enhancer predictive models. Although some success has been achieved in predicting enhancers using these algorithms, two significant challenges remain. First, it is well-established that the presence of epigenomic features such as transcription factor binding sites (TFBS), chromatin accessibility, DNA methylation, and histone modification marks does not guarantee that the site is an enhancer region. Second, epigenomics datasets can differ significantly among various cell types and environmental conditions, which limits the applicability of these algorithms based on the cell-type-specific epigenomics datasets used for training. Consequently, these published state-of-art methods have limited generalization capabilities when it comes to unknown cell types and environmental conditions.

Here we have tackled these challenges by introducing a new deep learning-based algorithm for enhancer prediction called AI-enhancer. This algorithm overcomes the limitations of previous methods by combining both enhancer DNA sequences and epigenomic profiles to train a predictive model. AI-enhancer utilizes a pre-trained convolutional neural network with a time distribution layer to extract features from epigenomic profiling data, and a long short-term memory (LSTM) with attention layer to learn features from enhancer DNA sequences. This innovative algorithm design allows AI-enhancer to leverage information from both sources, resulting in higher predictive accuracy for enhancer activities than previously published state-of-the-art methods, as demonstrated by statistical examination and experimental evidence presented in this study. We have provided both computational and experimental evidence to demonstrate the efficacy of AI- enhancer in predicting enhancer activities. We also provided evidence to demonstrate the ability of AI-enhancer to predict enhancer-altering mutations directly, which has not been comprehensively explored in published enhancer predictive models.

Figure 1: Cartoon depicting AI-enhancer algorithm.



Poster #2552 - Development of a Deep Learning Model for Differentiating Vitiligo from Other Pigmentation Disorders

Nneka Ede

Dell Medical School and Cockrell School of Engineering, The University of Texas at Austin, Austin, TX

ABSTRACT: Very little research focuses on deep learning models for the detection of non-melanocytic pigmentation disorders such as vitiligo.[1,2] The differential diagnosis for vitiligo can be quite broad, and it, in itself, can have a variety of clinical presentations; therefore, making it at times difficult to diagnose.[3] The utilization of deep learning has the potential to serve as a helpful clinical diagnostic tool as it may be able to discern fine variations in depigmentation.[4] This study develops and evaluates the accuracy of a deep learning model's ability to discriminate vitiligo from other pigmentation disorders.

A convolutional neural network was trained on over 10,000 images of patients diagnosed with pigmentation diseases. Specifically, the image dataset contained photos of vitiligo and other disorders such as melasma, pityriasis alba, pityriasis versicolor, guttate hypomelanosis, xeroderma pigmentosum, and tinea versicolor. Inception-V3 was used as the base model for transfer learning. The model reached an accuracy of 0.92, an F1 score of 0.95, a recall score of 0.94, and a precision score of 0.95.

These metrics demonstrate the accuracy of the deep learning model used in this study for detecting vitiligo and distinguishing it from other common pigmentation disorders in a controlled experimental setting. While this study shows promising results in its ability to classify vitiligo lesions, there is a need for further research with larger imaging datasets for further clinical validation.

REFERENCES:

1. Chan, S., Reddy, V., Myers, B. et al. Machine Learning in Dermatology: Current Applications, Opportunities, and Limitations. Dermatol Ther (Heidelb) 10, 365–386 (2020). https://doi.org/10.1007/s13555-020-00372-0
2. Zhang J, Zhong F, He K, Ji M, Li S, Li C. Recent Advancements and Perspectives in the Diagnosis of Skin Diseases Using Machine Learning and Deep Learning: A Review. Diagnostics (Basel). 2023 Nov 22;13(23):3506. doi: 10.3390/diagnostics13233506. PMID: 38066747; PMCID: PMC10706240.
3. Goh BK, Pandya AG. Presentations, Signs of Activity, and Differential Diagnosis of Vitiligo. Dermatol Clin. 2017 Apr;35(2):135-144. doi: 10.1016/j.det.2016.11.004. PMID: 28317523.
4. Patel RH, Foltz EA, Witkowski A, Ludzik J. Analysis of Artificial Intelligence-Based Approaches Applied to Non-Invasive Imaging for Early Detection of Melanoma: A Systematic Review. Cancers (Basel). 2023 Sep 23;15(19):4694. doi: 10.3390/cancers15194694. PMID: 37835388; PMCID: PMC10571810.

Poster #2554 - Fast and Reliable Assessment of Hip Dysplasia with Artificial Intelligence-Generated Measurements

Holden Archer[1], Seth Reine[1], Ahmed Alshaikhsalama[1], Louis Vazquez[1], Yin Xi[1], Ajay Kohli[1], Avneesh Chhabra[1], Joel Wells[2]

[1]UT Southwestern Medical Center, Dallas, TX; [2]Baylor Scott & White, McKinney, TX

INTRODUCTION: Hip dysplasia (HD) leads to premature osteoarthritis. Timely detection and correction of HD has been shown to improve pain, functional status, and hip longevity. Several time-consuming radiographic measurements are used to confirm HD. An artificial intelligence (AI) software named HIPPO automatically locates anatomical landmarks on anteroposterior (AP) pelvis radiographs and performs the needed measurements. The primary aim of this study was to assess the reliability of this tool as compared to multi-reader evaluation in clinically proven cases of adult HD for external validation. The secondary aims were to assess the time savings achieved and evaluate inter-reader assessment.

METHODS: This study received institutional review board approval. A consecutive pre-operative sample of 130 hip dysplasia patients (82.3% women and 17.7% men with median patient age at 28.6 years) was used. Three trained readers' measurements were compared to AI outputs of lateral center edge angle (LCEA), caput-collum-diaphyseal (CCD) angle, pelvic obliquity, Tönnis angle, Sharp's angle, and femoral head coverage. Intraclass correlation coefficients (ICC) and Bland-Altman analyses were obtained.

RESULTS SECTION: Among 256 hips with AI outputs, all six hip AI measurements were successfully obtained. The AI-reader correlations were generally good (ICC=0.60 to 0.74) to excellent (ICC> 0.75). There was lower agreement for CCD angle measurement. Most widely used measurements for HD diagnosis (LCEA and Tonnis angle) demonstrated good to excellent inter-method reliability (ICC=0.71- 0.86 and 0.82-0.90). The median reading time for the three readers and AI was 212, 131, 734, and 41 seconds, respectively.

DISCUSSION: This study validated that the AI-based trained software demonstrated significant time savings in reliable radiographic assessment of patients with hip dysplasia.

SIGNIFICANCE/CLINICAL RELEVANCE: In addition to providing extensive time savings, integration of this AI system could provide preliminary measurements to physicians and direction for more thorough assessment for HD, especially in places without access to board-certified radiologists or orthopedic surgeons to conduct the measurements.

Poster #2555 - Enhancing Radiology Workflow via AI-Driven Prediction of Suboptimal Radiographs

Mahan Ghasemi, Mohammad Ghasemi Rad, M.D.

UT Southwestern Medical Center, Dallas, TX

Purpose: This study explores AI's role in predicting inconclusive and repeated radiographs, aiming to mitigate challenges arising from poor visualization and repeat imaging, thereby reducing delays in patient diagnosis and treatment planning. The research investigates AI's potential to monitor suboptimal radiology exams and streamline workflow by providing immediate feedback to technicians.

Methods: We examine recent AI applications in radiology and related fields, highlighting their success in algorithmic function and pattern recognition. By reviewing cases demonstrating AI's effectiveness in algorithmic function and pattern recognition, we propose its utilization for preventive quality control of radiographs and optimizing priority imaging modalities when fulfilling orders. The review explores pattern recognition, anatomical landmark demarcation, and algorithmic prioritization, demonstrating how these can prevent unnecessary repeat imaging.

Results: Machine learning and deep learning have been tested by physicians and proved successful in recognizing anatomical structures, outlining tumors, and even diagnosing low complexity pathologies in various modes of radiology. Basic algorithms of AI can be programmed to evaluate different parameters, the simplest model of which can order items by importance in the same way that physicians have differential preferences for diagnostic tools with different cases.

Conclusions: Machine learning tools can be seamlessly integrated into radiology departments to predict and prevent suboptimal imaging that will yield inconclusive results. They offer instant feedback to technicians through pattern recognition algorithms, reducing unnecessary repeat exams.
Additionally, AI can prevent redundancy when gold standard tools like CT scans have already been used, minimizing delays, and streamlining patient care and treatment planning.

Poster #2559 - Untrained Neural Network for Super-Resolving Non-Contrast-Enhanced 3D Whole-Heart MRI Using REACT

Corbin Maciel[1], Qing Zou[2,3,4]

[1]Department of Biomedical Engineering, UT Southwestern Medical Center; [2]Division of Pediatric Cardiology, Department of Pediatrics, UT Southwestern; [3]Department of Radiology, Department of Pediatrics, UT Southwestern; [4]Advanced Imaging Research Center, Department of Pediatrics, UT Southwestern

Objectives: The main objective is this study is to decrease scan time in 3D whole-heart MRI. Specifically, we propose an unsupervised neural network be used to achieve this objective by super-resolving low-resolution scans.

Methods: An unsupervised method referred to here as the super-resolution neural network (SRNN) was developed to perform super-resolution (SR) on 3D whole-heart cardiac MR images. Parameters were tuned to perform the SR task for a factor of 2 (2X) and the parameters used for a factor of 4 (4X) were based on the existing literature.The SRNN trains for the 1000 or 2000 epochs for 2X and 4X patients respectively, utilizing an initial learning rate of 0.01 and the standard mean-squared-error loss. Parameters are tuned using the Adam optimizer. To evaluate the SRNN against other methods and the ground truth, signal-to-noise ratio (SNR) and a two-tailed, paired t-test were used. Moreover, visual comparisons were created to showcase the SRNN's performance against other methods and the high-resolution ground truth.

Results: The SRNN consistently achieves the highest SNR. Furthermore, the resulting p-values indicate significance in every case (p-value < 0.05 indicates significance). The visual comparisons demonstrate that the SRNN produces images anatomically close to the ground truth and with greater clarity and artifact suppression than the other methods compared.

Conclusions: Based on the results achieved in this study it is concluded that the SRNN can super-resolve low-resolution 3D whole-heart MR images with anatomical accuracy and clarity, thereby decreasing scan time and maintaining image quality.

Poster #2576 - Zero-Shot Multi-modal Questions Answering for Assessment of Medical Student OSCE Physical Exams

Michael J. Holcomb[1], Shinyoung Kang[1], Ameer Shakur[1], Sol Vedovato[1], Thomas O. Dalton[2], Krystle K. Campbell[3], Daniel J. Scott[3,4], Gaudenz Danuser[1], Andrew R. Jamieson[1]

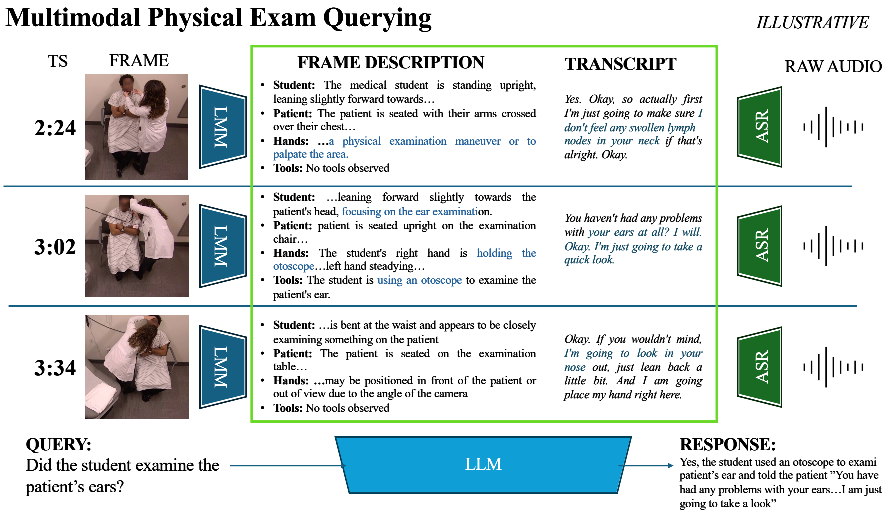
[1]Lyda Hill Department of Bioinformatics, UT Southwestern Medical Center; [2]Department of Internal Medicine, UT Southwestern; [3]Simulation Center, UT Southwestern; [4]Department of Surgery, UT Southwestern

Abstract: An Objective Structured Clinical Examination (OSCE) is a critical component of Medical Education whereby the data gathering, clinical reasoning, physical examination, diagnostic and planning capabilities of medical students are assessed in a simulated outpatient clinical setting with standardized patient actors (SPs) playing the role of patients with a predetermined diagnosis, or case. This study is the first to explore the zero-shot automation of physical exam grading in OSCEs by applying multimodal question answering techniques to the analysis of audiovisual recordings of simulated medical student encounters. Employing a combination of large multimodal models (LMMs), automatic speech recognition (ASR), and large language models (LLMs), this research seeks to assess the feasibility of applying these component systems to the domain of student evaluation without any retraining as illustrated in Figure 1.

A collection of 191 audiovisual recordings of medical student encounters with an SP for a single OSCE case was used as a test bed for exploring relevant features of successful exams. During this case, the students should have performed three physical exams: 1) mouth exam, 2) ear exam, and 3) nose exam. These examinations were each scored by two trained, non-faculty standardized patient evaluators (SPE) using the audiovisual recordings—an experienced, non-faculty SPE adjudicated disagreements.

The percentage agreement between the described methods and the SPEs' determination of exam occurrence as measured by percentage agreement varied from 26% to 83%. The audio-only methods, which relied exclusively on the transcript for exam recognition, performed uniformly higher by this measure compared to both the image- only methods and the combined methods across differing model sizes. The outperformance of the transcript-only model was strongly linked to the presence of key phrases where the student-physician would “signpost” the progression of the physical exam for the standardized patient, either alerting when they were about to begin an examination or giving the patient instructions.

Multimodal models offer tremendous opportunity for improving the workflow of the physical examinations evaluation, for example by saving time and guiding focus for better assessment. While these models offer the promise of unlocking audiovisual data for downstream analysis with natural language processing methods, our findings reveal a gap between the off-the-shelf AI capabilities of available models and the nuanced requirements of clinical practice, highlighting a need for further development and enhanced evaluation protocols in this area. We are actively pursuing a variety of approaches to realize this vision.



2:24

- **Student:** The medical student is standing upright, leaning slightly forward towards…
- **Patient:** The patient is seated with their arms crossed over their chest…
- **Hands:** **…**a physical examination maneuver or to palpate the area.
- **Tools:** No tools observed

Yes. Okay, so actually first I'm just going to make sure I don't feel any swollen lymph nodes in your neck if that's alright. Okay.

3:02

- **Student:** …leaning forward slightly towards the patient's head, focusing on the ear examination.
- **Patient:** patient is seated upright on the examination chair…
- **Hands:** The student's right hand is holding the otoscope…left hand steadying…
- **Tools:** The student is using an otoscope to examine the patient's ear.

You haven't had any problems with your ears at all? I will. Okay. I'm just going to take a quick look.

3:34

- **Student:** …is bent at the waist and appears to be closely examining something on the patient
- **Patient:** The patient is seated on the examination table…
- **Hands:** **…**may be positioned in front of the patient or out of view due to the angle of the camera
- **Tools:** No tools observed

Okay. If you wouldn't mind, I'm going to look in your nose out, just lean back a little bit. And I am going place my hand right here.

QUERY:
Did the student examine the patient’s ears?

RESPONSE:
Yes, the student used an otoscope to exami patient’s ear and told the patient ”You have had any problems with your ears…I am just going to take a look”

Figure 1: System Illustration

Poster #2577 - Revolutionizing Postoperative Ileus Monitoring: Exploring GRU-D's Real-Time Capabilities and Cross-Hospital Transferability

Xiaoyang Ruan[1,2], Sunyang Fu[1,2], Heling Jia[2], Kellie L. Mathis[3], Cornelius A. Thiels[3], Patrick M. Wilson[4], Curtis B. Storlie[4], Hongfang Liu[1,2]

[1]McWilliams School of Biomedical Informatics, University of Texas Health Science Center at Houston; [2]Department of Artificial Intelligence & Informatics, Mayo Clinic, Rochester, MN; [3]Department of Surgery, Mayo Clinic; [4]Department of Quantitative Health Sciences, Mayo Clinic

Background: Postoperative ileus (POI) after colorectal surgery leads to increased morbidity, costs, and hospital stays. Identifying POI risk for early intervention is important for improving surgical outcomes especially given the increasing trend towards early discharge after surgery. While existing studies have assessed POI risk with regression models, the role of deep learning's remains unexplored.

Methods: We assessed the performance and transferability (brutal force/instance/parameter transfer) of Gated Recurrent Unit with Decay (GRU-D), a longitudinal deep learning architecture, for real-time risk assessment of POI among 7,349 colorectal surgeries performed across three hospital sites operated by Mayo Clinic with two electronic health records (EHR) systems. The results were compared with atemporal models on a panel of benchmark metrics.

Results: GRU-D exhibits robust transferability across different EHR systems and hospital sites, showing enhanced performance by integrating new measurements, even amid the extreme sparsity of real-world longitudinal data. On average, for labs, vitals, and assisted living status, 72.2%, 26.9%, and 49.3% respectively lack measurements within 24 hours after surgery. Over the follow-up period with 4-hour intervals, 98.7%, 84%, and 95.8% of data points are missing, respectively. A maximum of 5% decrease in AUROC was observed in brutal-force transfer between different EHR systems with non-overlapping surgery date frames. Multi-source instance transfer witnessed the best performance, with a maximum of 2.6% improvement in AUROC over local learning. The significant benefit, however, lies in the reduction of variance (a maximum of 86% decrease). The GRU-D model's performance mainly depends on the prediction task's difficulty, especially the case prevalence rate. Whereas the impact of training data and transfer strategy is less crucial, underscoring the challenge of effectively leveraging transfer learning for rare outcomes. While atemporal Logit models show notably superior performance at certain pre-surgical points, their performance fluctuate significantly and generally underperform GRU-D in post-surgical hours.

Conclusion: GRU-D demonstrated robust transferability across EHR systems and hospital sites with highly sparse real-world EHR data. Further research on built-in explainability for meaningful intervention would be highly valuable for its integration into clinical practice.

Poster #2581 - Real-Time Secondary Dose Verification for Online Adaptive Radiotherapy (ART) via Geometry-Encoded U-Net

Shunyu Yan, Austen Maniscalco, Biling Wang, Nguyen Dan, Steve Jiang, Chenyang Shen

Medical Artificial Intelligence and Automation (MAIA) Laboratory, Department of Radiation Oncology, UT Southwestern Medical Center, Dallas, TX

Purpose: For online ART treatment planning workflow with a patient already being set on the treatment couch, computational-based secondary dose verification is strongly desired to ensure accuracy and quality of the ART plans. However, it often involves extensive computations, substantially prolonging the time patients spend on the couch, and thereby limiting the efficiency and accuracy of ART. This study aimed at developing a real-time deep-learning (DL) based secondary dose verification algorithm to accurately predict dose distributions based on CT images and fluence maps (FMs). It innovatively integrates FMs into CT image domain by explicitly resolving the geometry of treatment delivery to enhance model training efficacy and efficiency.

Methods and Materials: For each gantry angle, FM was constructed based on the optimized multi-leaf-collimator apertures and the corresponding monitoring units. FMs of different arcs were summed together as a unified representation. To effectively encode the treatment beam configuration, the constructed FMs were encoded at 30cm from the isocenter following the exact geometry on treatment machines. For dose estimation, a $3D$ U-Net was employed to take the integrated CT and FM volume as input to generate a dose distribution. Training and validation were conducted on 381 prostate cancer patients treated in our institution. Another 40 testing cases were retrieved to independently evaluate the performance of the established model.

Results: The proposed model can perform secondary dose verification within ~15ms for each patient. The average γ passing rate (3%/2mm, 10% threshold) of the estimated dose is 99.9% ± 0.16% on testing patients. Dose coverage for the target and OAR sparing showed minimal deviation, with most parameters within 1% error. To assess the model's effectiveness, we utilized CT scans as input and expected a zero- dose output. Our results showed a negligible average dose discrepancy across all cases, demonstrating the model's precision and potential clinical utility.

Conclusions: The proposed DL-based dose verification model can accurately estimate 2D dose distribution in real-time for ART.

Poster #2583 - Artificial Intelligence Chatbots as Sources for Patient Education Material on Child Abuse

Lily Nguyen; Viet Tran; Joy Li; Denise Baughn, M.D.; Joseph Shotwell, M.D.; Kimberly Gushanas, Ph.D.; Sayyeda Hasan, M.D.; Lisa Falls, M.D.; Rocksheng Zhong, M.D.

University of Texas Medical Branch, Department of Psychiatry and Behavioral Sciences

Background: Childhood maltreatment, such as child abuse or neglect, critically impacts neurodevelopment and increases risk of psychiatric disorders in adulthood. With the increasing power and accessibility of artificial intelligence (AI) Large Language Models (LLMs), patients may turn to these platforms as sources of medical information, however their efficacy remains unexplored.

Methods: Eight questions on child abuse were taken from National Child Traumatic Stress Network (NCTSN) webpages and inputted into ChatGPT, Google Gemini, and Microsoft Copilot. Child and Adolescent psychiatrists and a pediatric psychologist were blinded to the sources and independently scored responses for quality, understandability, and actionability. Secondary outcomes included misinformation, readability, word count, and top reference sources.

Results: Analysis of 32 queries from LLMs and NCTSN about child abuse showed good quality (mean DISCERN, 51.7 [range 45.4-55.5]), with little to no misinformation. Understandability was moderate (mean PEMAT, 76.5% [range 73-80%]), and actionability was poor (mean PEMAT, 64% [range 52%-72%]. Responses were written at a tenth-grade level at best based on the Flesch-Kincaid Grade Level score. ChatGPT responses were more difficult to read than NCTSN ($p<0.05$). AI chatbots produced significantly longer responses ($p<0.001$).

Conclusions: AI chatbots can provide accurate information on child abuse akin to authoritative sources but significantly lengthier. However, all sources lack actionable guidance and exceed recommended reading levels, which limits their effectiveness. AI chatbots should complement rather than replace primary medical information sources. Improvements are needed to enhance accessibility, readability, and actionability of patient education materials on topics like child abuse and neglect.

Poster #2584 - A Language Model Fusion Approach for Identifying Patient Primary Concerns in Patient Portal Messages

Yang Ren, M.S.[2]; Yuqi Wu, Ph.D.[1]; Jung-wei Fan, Ph.D.[1]; Aditya Khurana, M.D.[1]; Sunyang Fu, Ph.D.[3]; Dezhi Wu, Ph.D.[2]; Hongfang Liu, Ph.D.[3]; Ming Huang, Ph.D.[1,3]

[1]Mayo Clinic, Rochester, MN; [2]University of South Carolina, Columbia, SC; [3]University of Texas Health Science Center, Houston, TX

Abstract: The increasing volume of patient portal messages (PPMs) creates a challenge for healthcare providers who need to efficiently triage messages and respond to patients promptly[1]. Existing AI solutions focus on streamlining workflows, but less attention has been paid to accurately identifying patients' primary concerns within PPMs [2,3]. Understanding these primary concerns is crucial for delivering high-quality, patient-centered care. This study proposes a novel fusion framework that leverages pretrained language models (LMs) with complementary strengths, as shown in **Figure 1**. We combine these LMs using a Convolutional Neural Network (CNN) for multi-class classification, enabling precise identification of patient primary concerns. We compared our framework to traditional machine learning models, individual BERT-based models, and various ensemble models.

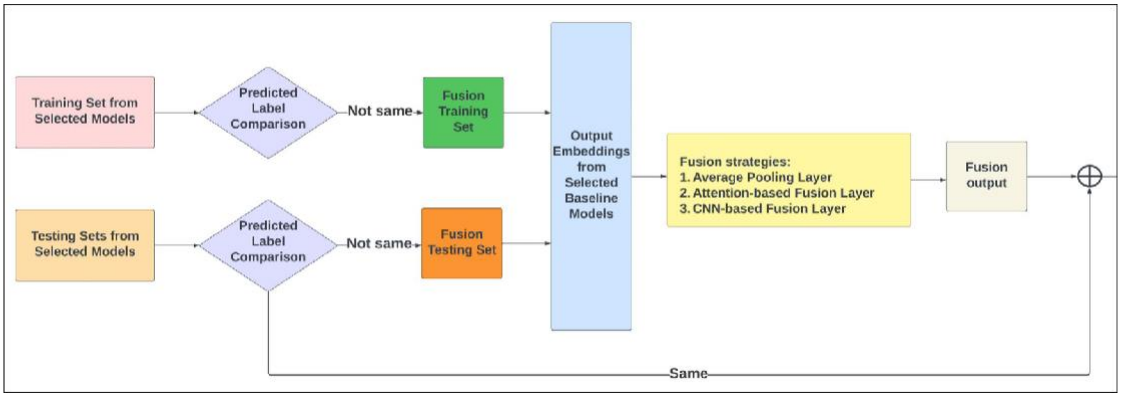

Training Set from Selected Models
Predicted Label Comparison
Not same
Fusion Training Set
Testing Sets from Selected Models
Predicted Label Comparison
Not same
Fusion Testing Set
Output Embeddings from Selected Baseline Models
Fusion strategies:
1. Average Pooling Layer
2. Attention-based Fusion Layer
3. CNN-based Fusion Layer
Fusion output
Same

Figure 1. An overview of the fusion strategies of pretrained language models.

The results demonstrate that BERT-based models outperform traditional methods. Notably, our proposed fusion model achieved the highest accuracy (77.67 ± 2.74%) and F1 score (74.37 ± 3.70%) in macro average. These findings highlight the effectiveness of multi-class classification for patient concern detection and the potential of our fusion framework to significantly improve accuracy. By combining multiple LMs and leveraging multi-class classification, this approach not only improves the accuracy and efficiency of identifying patient concerns in PPMs, but also helps healthcare providers manage the growing volume of messages and ensure timely responses to critical communications.

Table 1. Performance of the developed models.

Group	**Model Name**	**Accuracy%**	**Precision%**	**Recall%**	**F1 score%**
Generic language models	BERT	73.17 ± 1.88	69.65 ± 2.22	68.98 ± 2.72	68.93 ± 2.27
	RoBERTa	**74.70 ± 2.63**	71.49 ± 3.66	71.68 ± 3.69	**70.93 ± 3.54**
	ALBERT	72.13 ± 3.86	68.68 ± 4.30	69.14 ± 4.28	68.08 ± 4.25
Domain-specific language models	BioBERT	**73.50 ± 2.30**	69.76 ± 3.34	69.49 ± 3.18	**69.23 ± 3.11**
	BioClinicalBERT	72.30 ± 2.89	68.65 ± 3.36	69.87 ± 2.99	68.62 ± 3.07
	PubMedBERT	72.97 ± 2.40	69.40 ± 3.11	69.23 ± 3.38	68.78 ± 3.08
Source-specific language models	BERTweet	73.33 ± 3.68	69.98 ± 3.36	71.64 ± 3.56	**69.96 ± 3.32**
	TWhinBERT	71.87 ± 2.64	68.33 ± 3.29	67.84 ± 3.66	67.57 ± 3.31
	RedditBERT	**73.60 ± 2.55**	70.10 ± 3.58	69.45 ± 3.48	69.41 ± 3.44
Fusion models	2BERT(d,s)+CNN*	75.12 ± 2.62	71.54 ± 3.81	70.83 ± 3.32	70.85 ± 3.31
	2BERT(g,s)+CNN*	75.83 ± 3.10	72.04 ± 4.02	72.71 ± 4.13	72.07 ± 3.95
	2BERT(g,d)+CNN*	75.26 ± 2.74	71.73 ± 3.45	71.80 ± 3.80	71.44 ± 3.51
	3BERT+Aver	76.55 ± 2.64	74.10 ± 3.91	70.95 ± 4.21	71.89 ± 3.71
	3BERT+Attn	76.17 ± 3.30	72.48 ± 4.16	72.47 ± 4.39	72.27 ± 4.14
	3BERT+CNN	**77.67 ± 2.74**	74.20 ± 3.47	75.05 ± 4.49	**74.37 ± 3.70**

* The g, d, and s denote generic, domain-specific, and source-specific language models, respectively.

Poster #2585 - Diagnostic Accuracy of GPT-4 in Differential Diagnosis Generation of Brain MRI Lesions

John Pruitt[1], Mitul Gupta[1], R. Nick Bryan[1], Taiyeb Rangwalla[1], Shawn Lyo[2], Felipe Rosero Castro[2], Suyash Mohan[2]

[1]Dell Medical School, The University of Texas at Austin; [2]Perelman School of Medicine, University of Pennsylvania

Purpose: The study seeks to establish benchmarking data for GPT-4's diagnostic accuracy in generating differential diagnoses from brain MRI examinations based solely on characteristic imaging features. It evaluates the performance of GPT-4 against human evaluators, attending physicians, fellows, and residents, aiming to identify the utility and limitations of LLMs in a real-world setting.

Materials and Methods: Data from a comprehensive single-center multi-site study at a large urban US academic health system was utilized. 476 clinical brain MRI scans representing 120 distinct diseases were selected randomly from clinical RIS/PACS archives. A subset of 32 cases with the highest inter-reader agreement on predefined characteristic imaging features was selected. GPT-4 and human evaluators were then tasked with generating a ranked list of top three differential diagnoses based on these features and the patient age. Diagnostic accuracies were quantitatively assessed using a scoring system where responses were scored as 0 if the correct diagnosis was not included, 1 if it was ranked first, and 3 if included but not first. A composite score of 1+3 was calculated to determine if the clinical diagnosis was "at least" included in the top 3.

Results: GPT-4 achieved a composite score of 18.8% through direct API queries and 31.3% using ChatGPT-4 interface, indicating modest performance. In contrast, human evaluators showed significantly

higher accuracies with fellows leading at 74.6%, followed by attendings at 72.1%, and residents at 60.9%.

Discussion: The results underscore GPT-4's limitations in clinical decision support, particularly in generating differential diagnoses from imaging features without direct image access. Factors such as prompt sensitivity and lack of domain-specific training may contribute to its lower performance compared to human readers. Establishing robust benchmarking data and continuous validation are essential steps before LLMs can be safely integrated into clinical decision-making processes.

Clinical Relevance Statement: This study provides comparative benchmarking data on the diagnostic accuracy of LLMs and physicians on brain MRI pathology using key characteristics, highlighting the current limitations and the need for ongoing LLM validation.

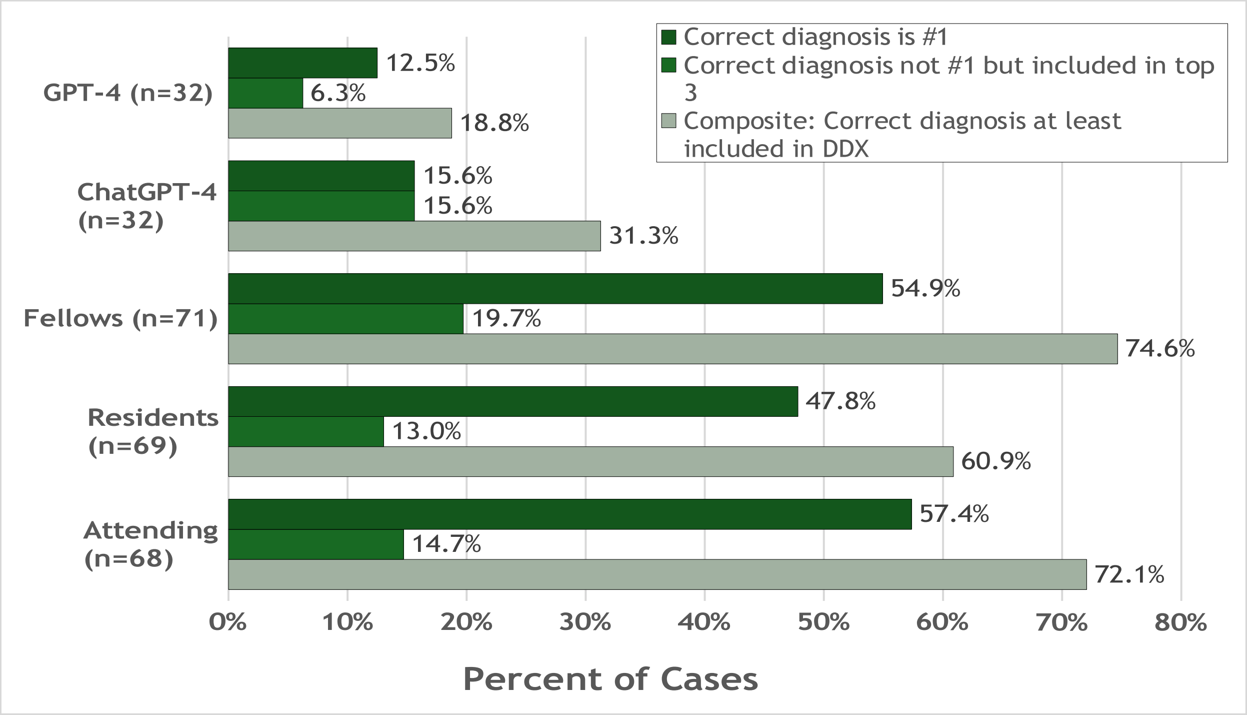


Figure. *Diagnostic accuracy scores broken down by evaluator classification. Notably, GPT4 (API access) performed worse than ChatGPT-4 (chat interface), but both performed worse than the human evaluators (residents, fellows, and attendings*).

Poster #2587 - Integration of External Knowledge into Pre-Trained Language Models Capturing Physician Reasoning from Electronic Health Records

Qiuhao Lu, Ph.D.; Andrew Wen, M.S.; Hongfang Liu, Ph.D.

McWilliams School of Biomedical Informatics, University of Texas Health Science Center, Houston, TX

Objective: This study aims to enhance the clinical relevance and interpretability of pre-trained language models (PLMs) by integrating external knowledge in a manner that respects the diversity and proprietary nature of healthcare data. We hypothesize that domain knowledge, if captured and distributed as standalone modules, can be effectively re-integrated into PLMs to improve their adaptability and utility in healthcare settings.

Methods: We demonstrate that via adapters, small and lightweight neural networks that enable the integration of extra information without full model fine-tuning, we can inject diverse sources of external domain knowledge into language models and improve the overall performance with an increased level of interpretability. As a practical application of this methodology, we introduce a novel task, structured as a case study, that endeavors to capture physician knowledge in assigning cardiovascular diagnoses from clinical narratives, where we extract diagnosis-comment pairs from Electronic Health Records (EHRs) and cast the problem as text classification.

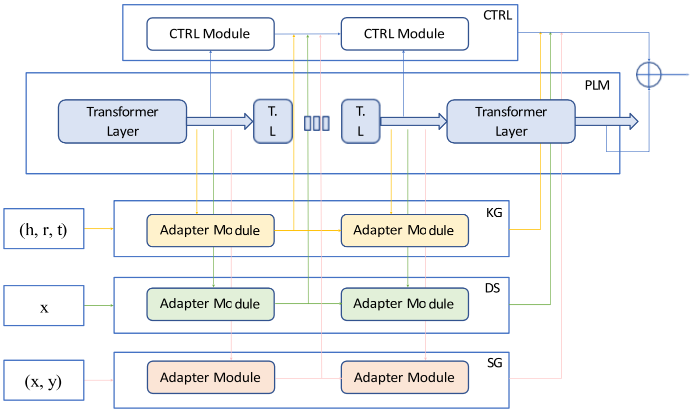


Figure 1: Diverse Adapters for Knowledge Integration (DAKI).

Results: The study demonstrates that integrating domain knowledge into PLMs significantly improves their performance. While improvements with ClinicalBERT are more modest, likely due to its pre-training on clinical texts, BERT adaptations surprisingly match or exceed ClinicalBERT in several metrics. This underscores the effectiveness of knowledge adapters and highlights their potential in settings with strict data privacy constraints. This approach also increases the level of interpretability of these models in a clinical context.

Conclusion: This research provides a basis for creating health knowledge graphs infused with physician knowledge, marking a significant step forward for PLMs in healthcare. Notably, the model balances integrating knowledge both comprehen- sively and selectively, addressing the heterogeneous nature of medical knowledge and the privacy needs of healthcare institutions.

Poster #2590 - Diagnostic and Prognostic Evaluation of an Echocardiography-Based Artificial Intelligence Algorithm for Detecting HFpEF: A Case-Control Analysis

Vinayak Subramanian, Katarina Yaros, Matthew Segar, Alvin Chandra, Thomas Koshy, Ross Upton, Ashley Ackerman, Ambarish Pandey

Background: Heart failure (HF) with preserved Ejection Fraction (HFpEF) is common among older adults and is associated with a high burden of morbidity and mortality. Despite its increasing prevalence, the diagnosis of HFpEF remains challenging and often requires the assessment of left ventricular filling pressure by invasive or non-invasive approaches. Recently, the FDA cleared an echocardiography-based AI HFpEF model that utilizes a 3-dimensional convolutional neural network to detect HFpEF using a single 4-chamber clip from a resting echocardiogram. However, the external validation of this algorithm against clinically adjudicated and confirmed HFpEF cases is limited.

Purpose: To evaluate the diagnostic and prognostic performance of the echocardiography-based AI HFpEF model in a cohort of HFpEF patients and matched controls.

Methods: The study included patients referred for HFpEF work-up to the UT Southwestern HFpEF clinic and age, sex, and BMI-matched control participants without HF and a normal echocardiogram. The HFpEF cases were clinically adjudicated based on the clinical history, signs and symptoms of HF, normal ejection fraction (>45%), and objective evidence of elevated filling pressures by resting (PCWP > 15 mm Hg) or exercise invasive hemodynamics (PCWP > 25 mm Hg) or echocardiogram (E/e' >14) in a subset. The performance of the AI HFpEF model was evaluated using receiver operator curves. Among patients with clinically adjudicated HFpEF, the association of the AI-HFpEF phenotype with elevated resting/exercise PCWP and peak exercise oxygen uptake (VO_{2peak}) was assessed using multivariable logistic and linear regression models adjusting for age, sex, race, BMI, and comorbidities (diabetes, hypertension, kidney disease, atrial fibrillation)

Results: Of the 166 patients referred for evaluation of HFpEF, 82% had clinically adjudicated HFpEF, and 69.8% had elevated LV filling pressure at rest or exercise. In the matched cohorts of patients with clinically adjudicated HFpEF and matched control individuals (N = 122 each), the AI algorithm-based probability of HFpEF demonstrated good performance in identifying clinically adjudicated and hemodynamically confirmed HFpEF (AUROC: 0.75 for each) that was greater than the widely used H2FpEF score (0.69 and 0.70, **Figure**). In the HFpEF referral cohort, a higher probability of HFpEF based on the AI- algorithm phenotype was significantly associated with lower VO_{2peak} (b [95% CI] per 5% higher probability: -0.11 [-0.21 to -0.01, P-value: 0.03] and greater odds of elevated PCWP (Odds ratio [95% CI] per 5% higher probability: 1.07 [1.01 – 1.15, P-value: 0.04] at rest or exercise after accounting for other confounders. Based on Youden's index, the AI algorithm- based probability threshold of >0.75 was identified as the optimal cutoff for detecting HFpEF by the AI algorithm, with high sensitivity (0.85) and accuracy (0.74) and, adequate specificity (0.66).

Conclusion: The echocardiography-based AI HFpEF model demonstrated excellent sensitivity and discrimination in identifying patients with vs. without clinical HFpEF. Furthermore, the AI

HFpEF model also had prognostic utility such that individuals with a higher probability of AI-HFpEF phenotype had more severe HFpEF.

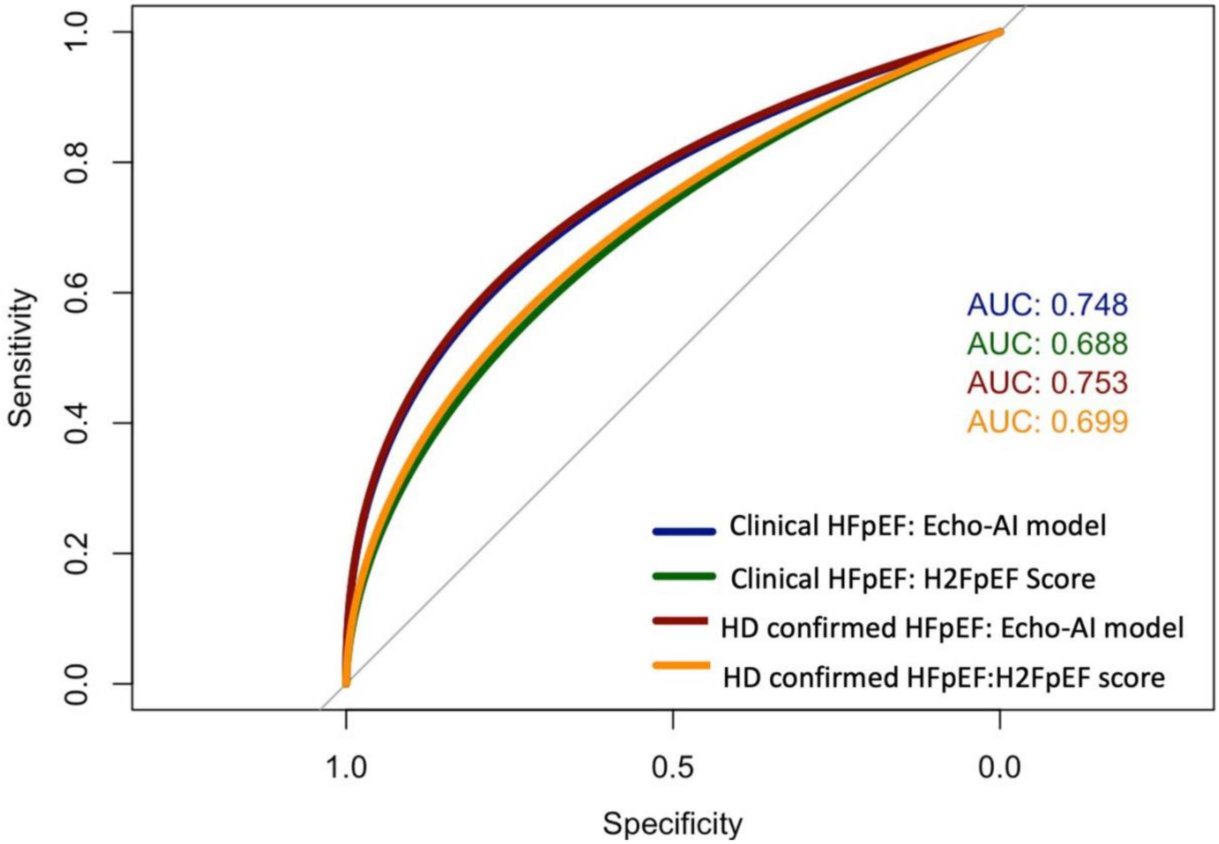


Figure: *Area under the receiver operating curve for predicting HFpEF outcome using the Echo AI model and H2FpEF score.*

Poster #2592 - Understanding the Implications of Systemic Biases in Generative Artificial Intelligence in Multiple Sclerosis

Mahi Patel, Francisco Villalobos, Kevin Shan, Lauren Tardo, Lindsay Horton, Peter Sguigna, Kyle Blackburn, Shanan Munoz, Tatum Moog, Alexander Smith, Katy Burgess, Morgan McCreary, Darin Okuda

Department of Neurology, UT Southwestern Medical Center

Background: Recent observations within our clinic suggest that Generation Z (Gen Z) people with MS utilize online generative artificial intelligence (AI) platforms for personalized medical advice before their first visit with a neuroimmunology specialist. Our objective was to determine if ChatGPT (Generative Pre-trained Transformer) could diagnose MS earlier than their clinical timeline, and to assess if accuracy differed based on age, sex, and race/ethnicity.

Methods: The clinical timeline for people diagnosed with MS, above the age of 18, was retrospectively identified and simulated using ChatGPT-3.5 (GPT-3.5). Chats were conducted using both actual and derivatives of their age, sex, and race/ethnicity to test diagnostic accuracy. Logistic regression (subject- specific intercept) was used to capture intra-subject correlation to test the accuracy prior to and after the inclusion of MRI data.

Results: The study cohort included 75 unique people with MS. Of those, 50 were members of Gen Z (38 female; 22 White; mean age at first symptom was 20.6 years (y) (standard deviation (SD)=2.2y)), and 25 were non-Gen Z (15 female; 16 White; mean age at first symptom was 34.2y (SD=10.2y)) with 386 digital simulations of these people created. Median time to diagnosis in clinic was significantly longer (0.39y (95% CI=[0.29, 0.58])) versus ChatGPT (0.08y (95% CI=[0.04, 0.28]) (p<0.013)). Prior to
including the MRI data, males had a 47.2% less likely chance of a correct diagnosis versus females (p=0.04). Post-MRI data inclusion, the odds of an accurate diagnosis was 3.8-fold greater for Gen Z, relative to non-Gen Z (p=0.007) with the diagnostic accuracy being 72.9% less in males versus females (p=0.004), and 69.8% less for White versus non-White subjects (p=0.003).

Conclusion: Although generative AI platforms enable rapid information access, obtained responses may not be generalizable to all users and bias may exist in select groups.

Poster #2593 - Zero-Shot Cone-Beam Computed Tomography (CBCT) to CT Conversion Using a Denoising Diffusion Wavelet Model (DDWM)

Yunxiang Li, Jiacheng Xie, You Zhang

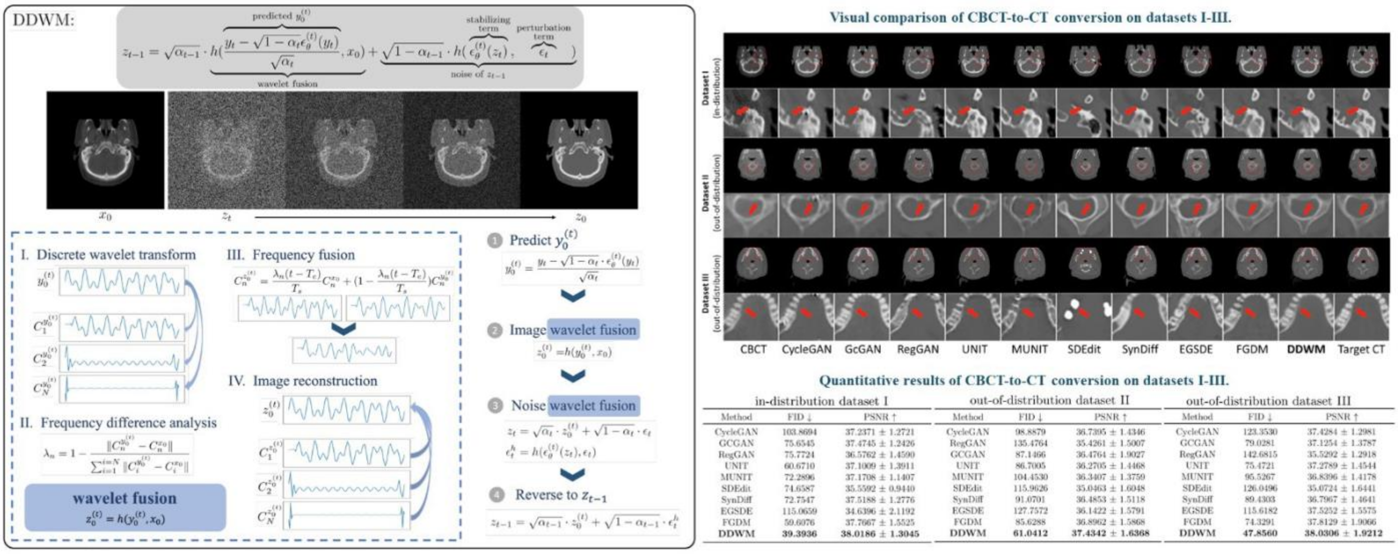
Medical Artificial Intelligence and Automation (MAIA) Laboratory, Department of Radiation Oncology, UT Southwestern Medical Center, Dallas, TX

Purpose: CBCT-to-CT conversion can reduce artifacts and enable accurate dose calculation for adaptive radiotherapy. Diffusion models allow accurate image conversion, but usually require paired datasets for model training. We proposed DDWM, a diffusion model which is solely trained on the CT domain and can be directly applied for zero-shot precise CBCT-to-CT conversion.

Methods: Diffusion models can learn data distributions by a reverse diffusion process. By DDWM, we trained a diffusion model solely using the CT data to learn CT distribution, and used wavelet transform to adaptively merge CBCT information into the diffusion model to achieve zero-shot CBCT-to-CT conversion. Specifically, we decomposed a CBCT into different frequency wavelet bands, and merged band-specific CBCT information into the reverse diffusion process to guide CBCT-to-CT conversion. For frequency bands with similar CBCT/CT information (for instance, high frequency anatomical details), the CBCT band information is merged with a higher weighting to preserve anatomical structures. For frequency bands with differing CBCT/CT information, the CBCT information is automatically weighted lower to allow replacement by the diffusion model-learned CT distribution.

Results: DDWM was trained on brain CT images (dataset I) and evaluated on three CBCT-to-CT conversion datasets (I-III). The converted CTs were evaluated against reference CTs, measured by Frechet-inception-distance (FID, lower is better) for image quality and peak signal-to-noise-ratio (PSNR, higher is better) for conversion faithfulness. As shown in Fig. 1, DDWM exhibits superior performance in both FID (39.39) and PSNR (38.02). In addition, FGDM clearly outperformed the other models when tested on never-before-seen out-of- distribution data without fine-tuning.

Conclusion: DDWM allows zero-shot, anatomy-preserving CBCT-to-CT conversion, which is essential for medical applications where paired datasets are scarce and anatomical integrity is critical.



$$z_{t-1} = \sqrt{\alpha_{t-1}} \cdot h(\underbrace{\overbrace{\frac{y_t - \sqrt{1-\alpha_t}\epsilon_\theta^{(t)}(y_t)}{\sqrt{\alpha_t}}}^{\text{predicted } y_0^{(t)}}, x_0)}_{\text{wavelet fusion}} + \underbrace{\sqrt{1-\alpha_{t-1}} \cdot h(\overbrace{\epsilon_\theta^{(t)}(z_t)}^{\text{stabilizing term}}, \overbrace{\epsilon_t}^{\text{perturbation term}})}_{\text{noise of } z_{t-1}}$$

$$\lambda_n = 1 - \frac{\|C_n^{y_0^{(t)}} - C_n^{x_0}\|}{\sum_{i=1}^{i=N} \|C_i^{y_0^{(t)}} - C_i^{x_0}\|}$$

$$z_0^{(t)} = h(y_0^{(t)}, x_0)$$

$$C_n^{z_0^{(t)}} = \frac{\lambda_n(t-T_e)}{T_s} C_n^{x_0} + (1 - \frac{\lambda_n(t-T_e)}{T_s}) C_n^{y_0^{(t)}}$$

$$y_0^{(t)} = \frac{y_t - \sqrt{1-\alpha_t} \cdot \epsilon_\theta^{(t)}(y_t)}{\sqrt{\alpha_t}}$$

$$z_0^{(t)} = h(y_0^{(t)}, x_0)$$

$$z_t = \sqrt{\alpha_t} \cdot z_0^{(t)} + \sqrt{1-\alpha_t} \cdot \epsilon_t$$

$$\epsilon_t^h = h(\epsilon_\theta^{(t)}(z_t), \epsilon_t)$$

$$z_{t-1} = \sqrt{\alpha_{t-1}} \cdot z_0^{(t)} + \sqrt{1-\alpha_{t-1}} \cdot \epsilon_t^h$$

Visual comparison of CBCT-to-CT conversion on datasets I-III.

Quantitative results of CBCT-to-CT conversion on datasets I-III.

in-distribution dataset I			out-of-distribution dataset II			out-of-distribution dataset III		
Method	FID ↓	PSNR ↑	Method	FID ↓	PSNR ↑	Method	FID ↓	PSNR ↑
CycleGAN	103.8694	37.2371 ± 1.2721	CycleGAN	98.8879	36.7395 ± 1.4346	CycleGAN	123.3530	37.4284 ± 1.2981
GCGAN	75.6545	37.4745 ± 1.2426	RegGAN	135.4764	35.4261 ± 1.5007	GCGAN	79.0281	37.1254 ± 1.3787
RegGAN	75.7724	36.5762 ± 1.4590	GCGAN	87.1466	36.4764 ± 1.9027	RegGAN	142.6815	35.5292 ± 1.2918
UNIT	60.6710	37.1009 ± 1.3911	UNIT	86.7005	36.2705 ± 1.4468	UNIT	75.4721	37.2789 ± 1.4544
MUNIT	72.2896	37.1708 ± 1.1407	MUNIT	104.4530	36.3407 ± 1.3759	MUNIT	95.5267	36.8396 ± 1.4178
SDEdit	74.6587	35.5592 ± 0.9440	SDEdit	115.9626	35.0463 ± 1.6048	SDEdit	126.0496	35.0724 ± 1.6441
SynDiff	72.7547	37.5188 ± 1.2776	SynDiff	91.0701	36.4853 ± 1.5118	SynDiff	89.4303	36.7967 ± 1.4641
EGSDE	115.0659	34.6396 ± 2.1192	EGSDE	127.7572	36.1422 ± 1.5791	EGSDE	115.6182	37.5252 ± 1.5575
FGDM	59.6076	37.7667 ± 1.5525	FGDM	85.6288	36.8962 ± 1.5868	FGDM	74.3291	37.8129 ± 1.9066
DDWM	**39.3936**	**38.0186 ± 1.3045**	**DDWM**	**61.0412**	**37.4342 ± 1.6368**	**DDWM**	**47.8560**	**38.0306 ± 1.9212**

Fig. 1. *Our DDWM framework and experimental results.*

Poster #2595 - Orthopedic AI

Blake Martin; Michael Sander, M.D.; Jorge Zwir

UT Rio Grande Valley School of Medicine

Background (Literature Review): Orthopedic conditions, particularly those needing hip and knee arthroplasty procedures, pose challenges in predicting clinical trajectories and optimizing treatment outcomes. Current approaches rely on generalized treatment pathways, leading to suboptimal resource utilization and compromised patient outcomes. The ORTHOPEDIC AI project aims to personalize treatment strategies using AI predictors and recommendation systems. The 2022 American Joint Replacement Registry (AJRR) annual report contains 2,550,532 valid primary and revision hip and knee arthroplasty cases submitted from 2012-2021 with majority of the cases being either total knee replacements (1,372,186 procedures, 53.8%) or total hip replacements (951,348 procedures, 37.3%). The AJRR is the largest arthroplasty registry, by volume, in the world. Integrating this repository of comprehensive patient data, ORTHOPEDIC AI seeks to optimize treatment selection, reduce hospitalization duration, minimize post-discharge complications, and enhance overall clinical outcomes.

Objectives/Aims:

1. Develop a comprehensive AI-driven predictive model for orthopedic patients undergoing hip and knee arthroplasty procedures.
2. Customize treatment recommendations based on individual patient profiles and clinical trajectories.
3. Implement recommendation systems to facilitate personalized interventions and improve clinical outcomes.

Hypotheses:

1. Integration of diverse patient data into an AI-driven predictive model will significantly improve outcome predictions.
2. Tailored treatment recommendations will lead to reduced hospitalization duration, fewer complications, and improved patient-reported outcomes.
3. Implementation of recommendation systems will optimize resource utilization and enhance patient satisfaction.

Results: The AI model is anticipated to provide accurate, quantifiable risks of total joint replacement surgeries across diverse patient characteristics.

Conclusions: The ORTHOPEDIC AI project aims to enhance knowledge of risks associated with total joint replacement surgeries, potentially leading to better outcomes for patients, clinicians, and their families.

Poster #2596 - Brain Cell Type Atlas Prediction by High-Resolution Diffusion MRI

Xinyue Han[1], Zhuoheng Liu[1], Jie Chen[1], Nian Wang[1,2,3]

[1]Advanced Imaging Research Center, UT Southwestern Medical Center, Dallas, TX; [2] Department of Biomedical Engineering, UT Southwestern; [3] Peter O'Donnell Jr. Brain Institute, UT Southwestern

Introduction: Cell types group together cells that exhibit the same anatomies and functions[1]. Brain cell types have been classified using costly single-cell sequencing technologies[2], which require heavy experimental and analytical steps. Magnetic resonance imaging (MRI), especially diffusion MRI, is a unique technique for probing brain tissue microstructure[3]. The quantitative metrics derived from diffusion MRI have been demonstrated to be sensitive imaging biomarkers for various neurodegenerative diseases[4,5]. However, whether MRI can be used to predict cell types in the brain is still unknown. In this study, we developed an artificial intelligence (AI) brain cell-type prediction model based on high-resolution diffusion magnetic resonance imaging (MRI). We also provided a single-cell MRI-based brain cell atlas at the Allen Mouse Brain Common Coordinate Framework (CCFv3) space.

Methods: We acquired high-resolution (50 um isotropic) adult mouse brains *ex vivo* diffusion MRI images (b value = 1000, 4000, 6000, 8000 s/mm^2, TR = 100 ms, TE = 15.2 ms). As shown in Figure 1, we extracted dMRI metrics through the whole brain. We then registered dMRI maps into CCFv3 space using Advanced Normalization Tools and used each dMRI metric as a predicting feature. Single- cell cell type data was imported from Allen Brain Cell (ABC) Atlas as responses, which has a hierarchical structure of 7 cell type neighborhoods and 34 cell type classes. By using one-versus- all random forest (RF) classifiers, we built classification models to classify firstly the cell type neighborhoods and secondly the cell type classes under each neighborhood. Classification performance was evaluated by splitting data by 75% /25% as training/validation set.

Results: The brain cell type neighborhood classification yields an average accuracy of 85.2% and an average precision of 84.5% across all neighborhoods, with the "TH-EPI-Glut" neighborhood having the highest accuracy. The brain cell type class classification under each neighborhood yields an average accuracy of 86.6% and an average prevision of 86.3%. The highest accuracy is from cell type class "MH-LH Glut".

Conclusions: High-resolution diffusion MRI can characterize the cell type distributions across the whole brain and can predict cell type atlas at single cell level.

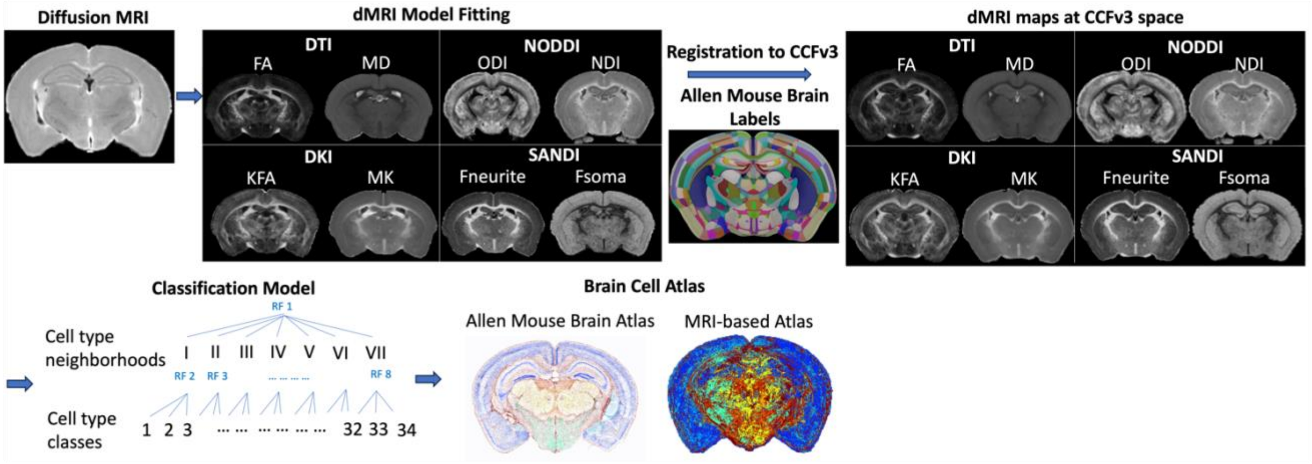


Figure 1. Workflow of the study.

References:

1 Zeng, H. What is a cell type and how to define it? *Cell*, 185, 2739-2755, (2022).

2 Zeng, H. & Sanes, J. R. Neuronal cell-type classification: challenges, opportunities and the path forward. *Nature reviews*, 18, 530-546, (2017).

3 Lerch JP, van der Kouwe AJW, Raznahan A, et al. Studying neuroanatomy using MRI. *Nat Neurosci*. 20, 314-326, (2017).

4 Andica, C., Kamagata, K., Hatano, T., Saito, Y., Ogaki, K., Hattori, N., & Aoki, S. MR biomarkers of degenerative brain disorders derived from diffusion imaging. *Journal of Magnetic Resonance Imaging*, *52*, 1620-1636, (2020).

5 Goveas, J., O'Dwyer, L., Mascalchi, M., Cosottini, M., Diciotti, S., De Santis, S., ... & Giannelli, M. Diffusion-MRI in neurodegenerative disorders. *Magnetic resonance imaging*, *33*, 853-876, (2015).

Poster #2597 - Rates of AI ‘Hallucination’ in Case-Based Learning Tasks Assigned to Medical Students

Benjamin Popokh

UT Southwestern Medical Center, Dallas, TX

Research Question: Based on Case-Based Learning (CBL) prompts provided to UTSW medical students, how often does AI “hallucinate” references for Question-Resource-Answer (QRA) tasks?

Objectives: We quantify the instances of AI hallucination (generation of realistic, but fake, outputs) when assigned QRA tasks based on CBL prompts assigned to medical students.

Methods: From 2023 – 2024, medical students at UTSW were assigned 6 CBL sessions, with 3 cases in each. ChatGPT was prompted with the exact QRA task as assigned to students, and all data from the simulated patient note was input into the AI. In total, 18 QRA responses were obtained. The resources provided by ChatGPT were searched online in various databases. If no parts of the reference could be found, this was classified as “Entirely Hallucinated”; if only some parts were accurate, this was “Partially Hallucinated”, and if the reference was entirely accurate, we labeled this as “Accurate”. We also categorized the types of questions generated.

Results: Out of the 18 references provided by AI in response QRA tasks, 12 (67%) were entirely hallucinated, 2 (11%) were partially hallucinated, and 4 (22%) were accurate. Questions generated AI included the following topics: disease etiology (n = 7, 39%), clinical manifestations and complications (n = 6, 33%), risk factors (n = 3, 17%), associations between risk factors and disease markers (n = 1, 5%), and clinical management strategies (n = 1, 5%).

Conclusions: Although AI tools such as ChatGPT 3.5 can generate well-written questions and answers based on simulated patient notes, the rates of AI hallucination when citing references are remarkably high. This finding should caution both students and educators when considering the use of AI to complete or supplement CBL activities, such as the QRA task. The fact that it can generate any, if only a minority, of accurate references is a testament to its future potential in medical education. Nevertheless, this study demonstrates that at this time and the foreseeable future, learners must continue to develop independent strategies for literature review and citation.

Poster #2598 - Deep Learning-Based H-Score Quantification of Immunohistochemistry-Stained Images

Zhuoyu Wen[1], Danni Luo[1], Shidan Wang[1], Ruichen Rong[1], Yang Xie[1,2,3], Guanghua Xiao[1,2,3]

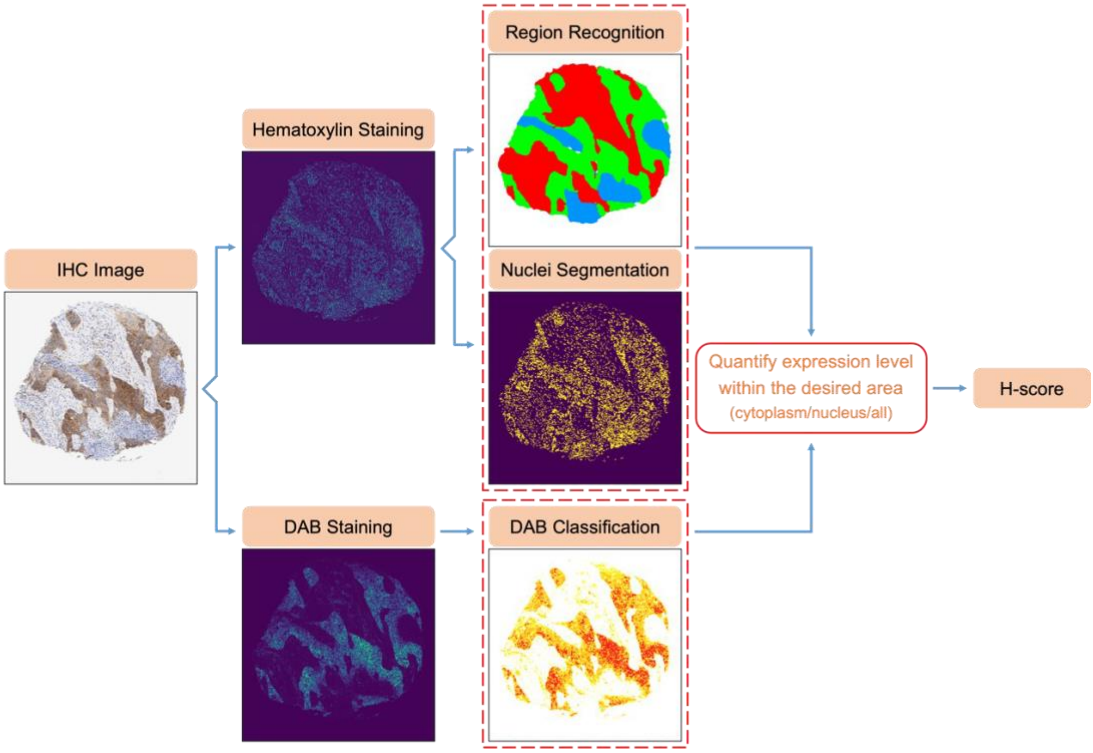
[1]Quantitative Biomedical Research Center, Peter O'Donnell Jr. School of Public Health, UT Southwestern Medical Center, Dallas, TX; [2]Simmons Comprehensive Cancer Center, UT Southwestern Medical Center, Dallas, TX; [3]Department of Bioinformatics, UT Southwestern Medical Center, Dallas, TX

Abstract: Immunohistochemistry (IHC) is a well-established and commonly-used staining method for clinical diagnosis and biomedical research. For most IHC images, the target protein is conjugated with a specific antibody and stained by Diaminobenzidine (DAB) as brown, while hematoxylin serves as a blue counterstain for cell nuclei. The protein expression level is quantified via the H-score, calculated from DAB staining intensity within the target cell region. Traditionally, this process requires evaluation by two expert pathologists, which is both time-consuming and subjective. To improve the efficiency and accuracy of this process, we developed an automatic algorithm to quantify the H-score of IHC images (Figure 1). To characterize the protein expression in specific cell regions, a deep learning model for region recognition was trained based on hematoxylin staining only, whose pixel accuracy for each class ranges from 0.92 to 0.99.

Within the desired area, the algorithm categorizes DAB intensity of each pixel as negative, weak, moderate, or strong staining, and calculates the final H-score based on the percentage of each intensity category.

Overall, this algorithm can use an IHC image as input and directly output the H-score within a few seconds, which greatly enhances the speed of IHC image analysis. This automated tool delivers H-score quantification with precision and consistency that is comparable to experienced pathologists, but at a significantly reduced cost during IHC diagnostic workups. It has significant potential to advance biomedical research reliant on IHC staining for protein expression quantification.

Figure 1. *Flowchart of automatic H-score quantification of IHC imag*es.



Poster #2600 - Automatic Identification of the Principle Compressive and Tensile Groups in the Human Femoral Head Using Artificial Intelligence-Based Segmentation

Dan Nguyen[1,2], Alireza Zomorodian[3,4], Chenyang Shen[1,2], Eleanor Lederer[3,4,5], Orson Moe[3,4], Naim Maalouf [3,4], Megan Sorich[6], Steve Jiang[1,2], Khashayer Sakhaee[3,4]

[1]Medical Artificial Intelligence and Automation (MAIA) Laboratory, UT Southwestern Medical Center, Dallas, TX; [2]Department of Radiation Oncology, UT Southwestern; [3]Department of Internal Medicine, UT Southwestern; [4]Charles and Jane Pak Center for Mineral Metabolism and Clinical Research, UT Southwestern; [5]VA North Texas Health Sciences Center;

[6]Department of Orthopedic Surgery, UT Southwestern

Rationale: Hip fracture (HFx) is a public health epidemic. Current bone imaging techniques do not assess bone architecture and inadequately evaluate fracture risk. We hypothesized that use of AI techniques to analyze data-intensive microCT scans would enable evaluation of bone microarchitecture.

Methods: We used a deep learning-based U-net style model to identify principal compressive and tensile groups in microCT scans of femoral heads from 2 HFx patients, containing 1685 and 1121 slices each, with voxel sizes of 30 x 30 x 30 um. We used data from Patient 1 (1685 slices) for model development, creating supervised learning data by manually contouring 200 segmentations each of compressive groups and tensile groups in slices, removing continuity bias by designing a random slice display contouring platform. We divided the 400 segmentations into 380 training and 20 validation. We designed a U-net style architecture to train on 2D slices of our labeled data, and trained the model for 1000 epochs, the final model showing the lowest validation loss. We used data from Patient 2 (1121 slices) to test the model, using 11 slices of the test data volume, evenly spaced apart along the slice direction, and segmented each compressive group for each slice twice. One segmentation was labeled ground truth, while the other was used to quantify intra- human variability as a baseline of the model performance. We calculated the Dice coefficient between ground truth data, and between ground truth and the 2nd segmentation.

$$\left(Dice(A, B) = \frac{2(A \cap B)}{\mathrm{A+B}} \right)$$

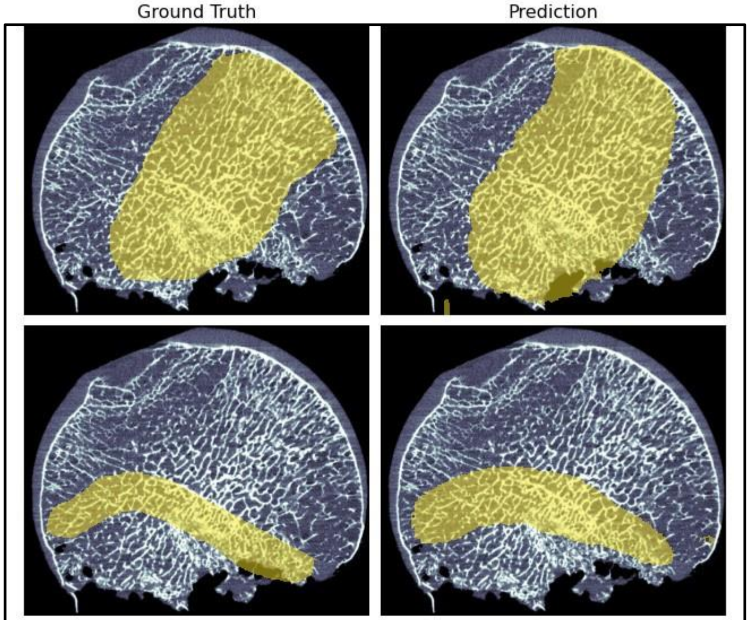


Figure 1: Principal compressive group (top) and tensile group (bottom) segmentations, with ground truth (left) and AI-based prediction (right), for an example slice

Results: The model achieved the lowest validation score at the 37th epoch before overfitting. For principal compressive and tensile groups, we achieved a Dice score of 0.87 ± 0.08 and 0.60 ± 0.16, respectively. In comparison the intra-human Dice score was 0.88 ± 0.05 and 0.64 ± 0.17.

Conclusion: The AI model performed similarly to a human contouring the groups, confirming the feasibility of using AI to accurately identify key elements of bone microstructure. Future plans include achieving super-human-level performance by increasing the labeled dataset and using a 3D model and expanding to identify other patterns and structures. This method will enable *in vivo*, noninvasive assessment of human bone microstructure at a level never before achieved, improving clinical capabilities to identify and reduce fracture risk.

Poster #2601 - Automated Video-Based Analysis for Assessment of Skills in Advanced Laparoscopic Suturing Using Deep Learning

Huu Phong Nguyen, Ph.D.[1]; Sofia Garces Palacios, M.D.[1]; Darian Hoagland, M.D.[2]; Sai Abhinav Pydimarry[1]; Shekhar Madhav Khairnar, M.S.[1]; Madhuri Nagaraj, M.D.[1]; Daniel J. Scott, M.D.[1]; Dmitry Nepomnayshy, M.D.[2]; Ganesh Sankaranarayanan, Ph.D.[1]

[1]UT Southwestern Medical Center, Dallas, TX; [2]Lahey Hospital and Medical Center, Burlington, MA

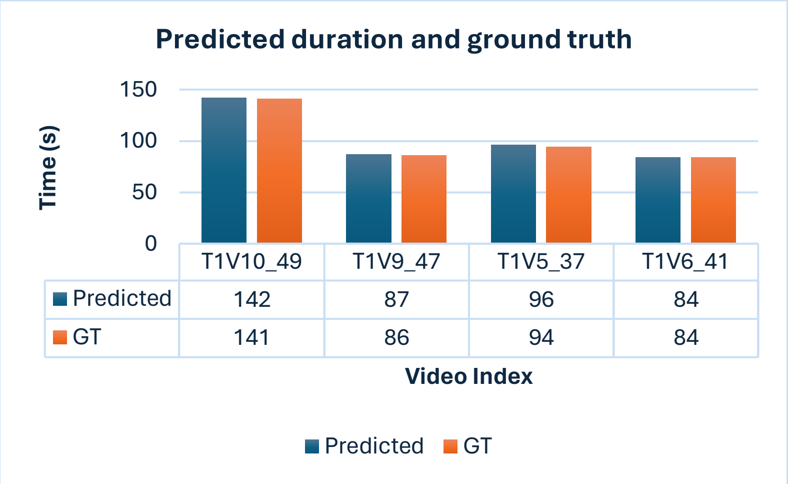
Introduction: Teaching advanced laparoscopic suturing remains challenging despite the proliferation of training programs. A national survey found that 73% of Fellowship Council program directors and fellows recognize the critical need for such a curriculum. To address this gap, the Simulation Committee of the Association for Surgical Education (ASE) collaborated with experts to develop the Advanced Training in Laparoscopic Suturing (ATLAS) program in 2022. ATLAS offers a structured curriculum with six proficiency-based tasks. In this work, we propose to utilize deep learning to automatically segment videos for future performance analysis.

Methods: Retrospective review of videos collected from a tertiary center from 2022 to 2024 was used in this study. We adopted a Deep Convolutional Neural Network (CNN) for frame-level visual feature extraction and classification. Detailed task segmentation was performed for the needle handling, the first of the five tasks of the ATLAS. The resulted 10 steps were annotated and timing was also recorded. End-to-end training of the frames extracted from the videos were performed using the X3D model.

Results: Out of the 12 videos analyzed, 8 were reserved for training with the remaining 4 were utilized for validating performance. Each video was extracted at 6 frames per second and annotated. The accuracy and per-class F1-score for tasks were 81.25% and 73.36%, respectively. Additionally, the predictions for durations compared against ground truth are plotted in Figure 1, showing an average error of approximately 1-2 seconds over a 100-second span. The model processed each frame in approximately 0.01 second, which makes it suitable for real-time applications.

Conclusion: Deep learning can be reliably used for automatic task segmentation of ATLAS videos. Techniques to detect needle drops and bends will be investigated to compute performance score automatically. Future applications of this model include the design of automated systems for skills assessment and provision of real-time feedback.

Figure 1. *Task duration prediction versus ground truth*



Poster #2602 - Enhancing Operations and Health Care Through Integrated AI: A Comprehensive Overview of UTMB's Advanced Infrastructure

Mark Schultze

The University of Texas Medical Branch

Abstract: The University of Texas Medical Branch (UTMB) is at the forefront of integrating artificial intelligence (AI) and machine learning (ML) technologies to revolutionize healthcare practices. Our state-of-the-art technological framework, supported by over 600 Azure resources, demonstrates uncompromised adherence to HIPAA and ISO 27001 standards, ensuring top-tier data security and patient privacy.

Infrastructure and Data Management: At the heart of our AI-driven initiatives is a comprehensive data warehouse that seamlessly interfaces with an analytical model, creating a robust environment for data-driven insights and decision-making. This is complemented by sophisticated data factory pipelines that efficiently process diverse data inputs from a myriad of sources including Epic, PeopleSoft, MD Staff, Taleo, UTIMCO, and Elsevier. This extensive data integration supports complex web applications and underpins our operational and research activities.

AI Integration and Applications: Central to our technological advancements is the deployment of Azure Machine Learning and OpenAI solutions, applied to operational data and an expansive patent dataset. These platforms have been pivotal in enhancing our operational efficiencies and fostering groundbreaking projects in AI and ML that promise substantial improvements in financial management and patient outcomes.

A flagship innovation within our suite of applications is Power-CoPilot, integrated into our Power application. This tool allows for real-time queries across all faculty operational data, demonstrating practical AI utility in everyday administrative and clinical decision-making.

Comprehensive Patient Data Vectorization: A significant stride in our data strategy includes the development of a comprehensive vectorized patient dataset, encompassing patient outcomes, encounter notes, lab results, and prescribed medications. This dataset has been meticulously curated to provide a granular view of patient histories, enhancing the ability of healthcare providers to access and utilize historical data for informed patient care.

The integration of this dataset into our OpenAI instance allows for sophisticated querying capabilities against the OpenAI large language models (LLMs). This functionality enables healthcare providers and staff to rapidly retrieve historical patient information, supporting a more informed and responsive patient care process.

Enhanced Healthcare Delivery and Provider Support: The deployment of AI technologies is a transformative step in healthcare delivery, where advanced AI solutions are being employed to potentially enhance disease diagnosis, predictive analytics, quality surveillance, and support for clinical decision-making. Furthermore, our AI-driven initiatives aim to enhance health system operations, including patient communications, documentation workflows, and the efficiency of care monitoring and billing systems, with anticipated improvements across these areas.

C**onclusion:** UTMB’s comprehensive AI infrastructure not only demonstrates our capability in managing extensive datasets and deploying advanced AI tools but also showcases our commitment to enhancing healthcare delivery, supporting educational endeavors, and ensuring operational excellence. This symposium presentation will detail the integration of these technologies and discuss the profound impacts they have had on our institution's efficiency and the quality of care provided to patients.

Poster #2603 - Assess Disease Progressive via Mixed Probability Density Function and Multilayer Perceptron

Jie Chen, Xinyue Han, Zhuoheng Liu, Nian Wang

Advanced Imaging Research Center, UT Southwestern Medical Center, Dallas, TX

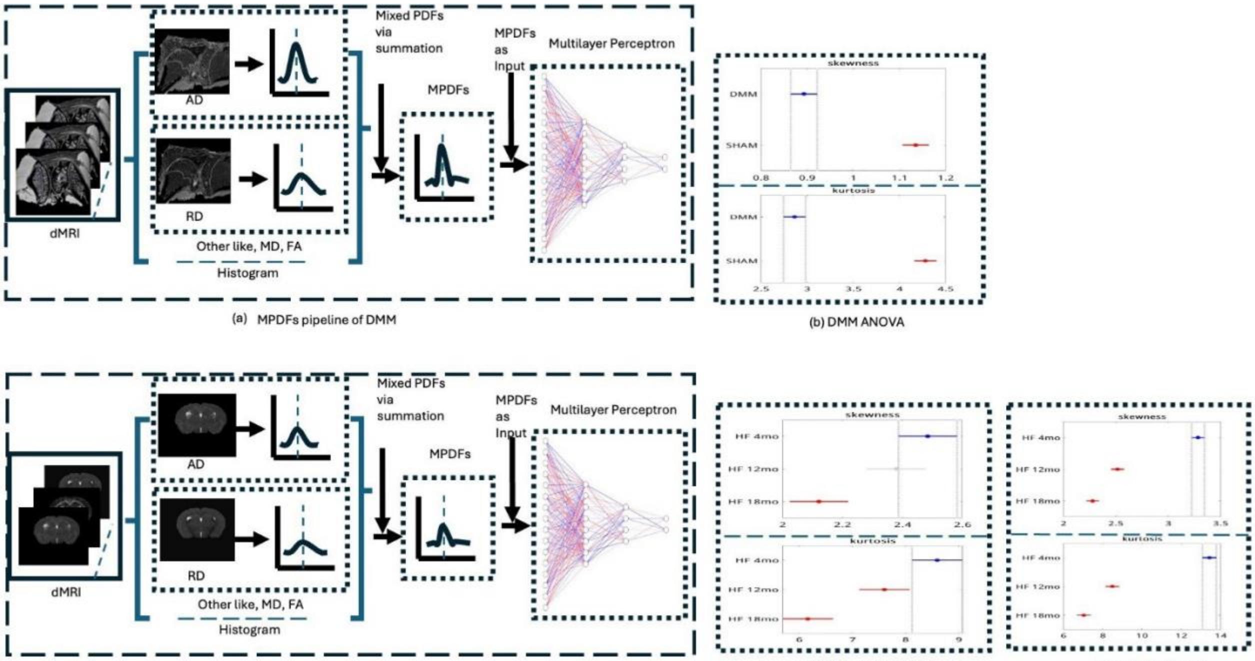
Introduction: Diffusion tensor imaging (DTI) metrics (AD, FA, MD, RD) effectively analyze tissue microstructures, aiding medical diagnosis and research. However, integrating DTI metrics for disease diagnosis and progression assessment requires further exploration. This study aims to integrate DTI metrics using mixed probability density functions (MPDFs) and feed them into a multilayer perceptron (MLP) for assessing disease progress.

Methods: Two diseases were investigated: destabilization of the medial meniscus (DMM) surgery in rats for Osteoarthritis and MODEL-AD with a high- fat diet in mice. The analysis involved 2,240 data samples for DMM and SHAM groups, 160 for female mice and 200 for male mice for MODEL-AD, each with four DTI metric images. MPDFs were generated from individual DTI metric histograms (a, c), and then fed into a four-layer MLP. Training parameters included a learning rate of 0.001, 3000 epochs, batch size 64, and Adam optimization with 75% training, 15% validation, and 15% test data. Evaluation metrics comprised accuracy, precision, recall, and F1 score. Statistical significance was determined using One-way ANOVA on MPDF skewness and kurtosis, with 95% CIs and $p \leq .05$.

Results: For DMM, accuracy reached 99.7%, with precision, recall, and F1 score at 1.0. Despite the MLP's simple architecture, it achieved high performance, likely due to significant skewness ($p \leq 0.0001$) and kurtosis ($p \leq 0.0001$) differences between DMM and SHAM groups, as shown by ANOVA (b).

For MODEL-AD, male mice achieved 93.3% accuracy, with precision, recall, and F1 scores at 0.94. Female mice reached 83.3% accuracy, with precision, recall, and F1 scores at 0.76. Skewness and kurtosis of female mice were statistically significant ($p \leq 0.0001$) across HF 4-month, HF 12-month, and HF 18- month groups, while no significant skewness difference was observed between HF 4-month and HF 12-month groups (d). Male mice exhibited significant skewness and kurtosis ($p \leq 0.0001$) across the three groups, indicating varying susceptibility between genders, with males showing higher susceptibility to the high-fat diet (e).

Conclusion: Combining MPDFs of DTI metrics images with MLP has the potential to serve as a valuable screening tool for disease diagnosis.



(a) MPDFs pipeline of DMM

(b) DMM ANOVA

(c) MPDFs pipeline of MODEL-AD

(d) HF Female ANOVA

(e) HF Male ANOVA

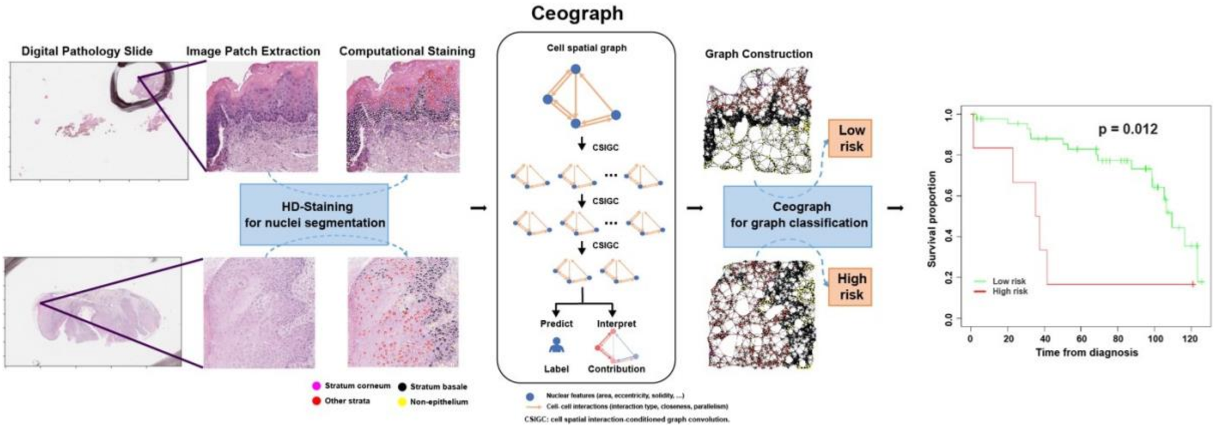
Poster #2607 - Deep Learning of Cell Spatial Organizations Identifies Clinically Relevant Insights in Tissue Images

Shidan Wang[1,*], Ruichen Rong[1], Qin Zhou[1], Donghan M. Yang[1], Xinyi Zhang[1], Xiaowei Zhan[1], Justin Bishop[2], Zhikai Chi[2], Clare J. Wilhelm[3], Siyuan Zhang[2], Curtis R. Pickering[4], Mark G. Kris[3], John Minna[5,7,8], Yang Xie[1,6,9], and Guanghua Xiao[1,6,9,*]

[1]Quantitative Biomedical Research Center, Peter O'Donnell Jr. School of Public Health, UT Southwestern, Dallas, TX; [2]Department of Pathology, UT Southwestern; [3]Department of Thoracic Oncology, Memorial Sloan Kettering Cancer Center, New York, NY; [4]Department of Surgery, Yale School of Medicine, New Haven, CT; [5]Hamon Center for Therapeutic Oncology Research, UT Southwestern; [6]Simmons Comprehensive Cancer Center, UT Southwestern; [7]Department of Pharmacology, UT Southwestern; [8]Department of Internal Medicine, UT Southwestern; [9]Department of Bioinformatics, UT Southwestern

*Corresponding author

ABSTRACT: Recent advancements in tissue imaging techniques have facilitated the visualization and identification of various cell types within physiological and pathological contexts. Despite the emergence of cell-cell interaction studies, there is a lack of methods for evaluating individual spatial interactions. In this study, we introduce Ceograph, a cell spatial organization-based graph convolutional network designed to analyze cell spatial organization (for example, the cell spatial distribution, morphology, proximity, and interactions) derived from pathology images. Ceograph identifies key cell spatial organization features by accurately predicting their influence on patient clinical outcomes. In patients with oral potentially malignant disorders, our model highlights reduced structural concordance and increased closeness in epithelial substrata as driving features for an elevated risk of malignant transformation. In lung cancer patients, Ceograph detects elongated tumor nuclei and diminished stroma-stroma closeness as biomarkers for insensitivity to EGFR tyrosine kinase inhibitors. With its potential to predict various clinical outcomes, Ceograph offers a deeper understanding of biological processes and supports the development of personalized therapeutic strategies.



Poster #2608 - Potential Role of Multiomics in Predicting Treatment Outcomes for Brain Metastases with Personalized Ultra-Fractionated Stereotactic Adaptive Radiotherapy (PULSAR)

Haozhao Zhang[1]; Michael Dohopolski, M.D. [1]; Strahinja Stojadinovic, Ph.D. [2]; Heejung Kim, Ph.D. [2]; Arnold Pompos, Ph.D. [1]; Andrew R. Godley, Ph.D. [1]; Steve Jiang, Ph.D. [1]; Zabi Wardak, M.D. [2]; Robert Timmerman, M.D.[2]; Hao Peng, Ph.D.[1]

[1]Medical Artificial Intelligence and Automation (MAIA) Lab, Department of Radiation Oncology, UT Southwestern Medical Center, Dallas, TX; [2]Department of Radiation Oncology, UT Southwestern

Objectives: PULSAR represents a paradigm shift in radiotherapy pioneered at UTSW. In this study, we developed a data-driven multiomics approach to predict treatment response at an earlier stage (intra-treatment) for brain metastases patients.

Methods: A retrospective analysis of 39 patients with 69 brain metastases treated with PULSAR was conducted. Radiomics, dosiomics, and delta features were extracted from pretreatment and intra-treatment MRI scans alongside dose distributions. Six individual models and an ensemble feature selection (EFS) model were constructed using support vector machines (SVM). The classification task aimed to differentiate between lesion groups based on whether they exhibited a volume reduction of more than 20% at follow-up. Performance metrics including sensitivity, specificity, accuracy, precision, F1 score, and area under the ROC curve (AUC) were assessed. Various feature extraction and ensemble feature selection scenarios were explored to enhance model robustness and prevent overfitting.

Results: The EFS model integrated crucial features from pre-treatment radiomics, pre-treatment dosiomics, intra-treatment radiomics, and delta-radiomics, outperforming six individual models with an AUC of 0.983, accuracy of 0.962, and F1 score of 0.903. Among the top 15 features of the EFS model, 14 were derived from post-wavelet transformation and 1 from original images. Discrete wavelet transform facilitated a more comprehensive characterization of underlying structures by decomposing volumetric images into multi- resolution components.

Conclusion: Our study highlights the feasibility of employing a data-driven multiomics approach to predict tumor volume changes in brain metastases patients undergoing PULSAR treatment. The EFS model exhibited superior performance compared to individual models, underscoring the significance of integrating both pretreatment and intra-treatment data. This application of multiomics alongside SVM classification for intra-treatment decision support in PULSAR shows promise for optimizing brain metastasis management and potentially mitigating risks associated with under- or over-treatment.

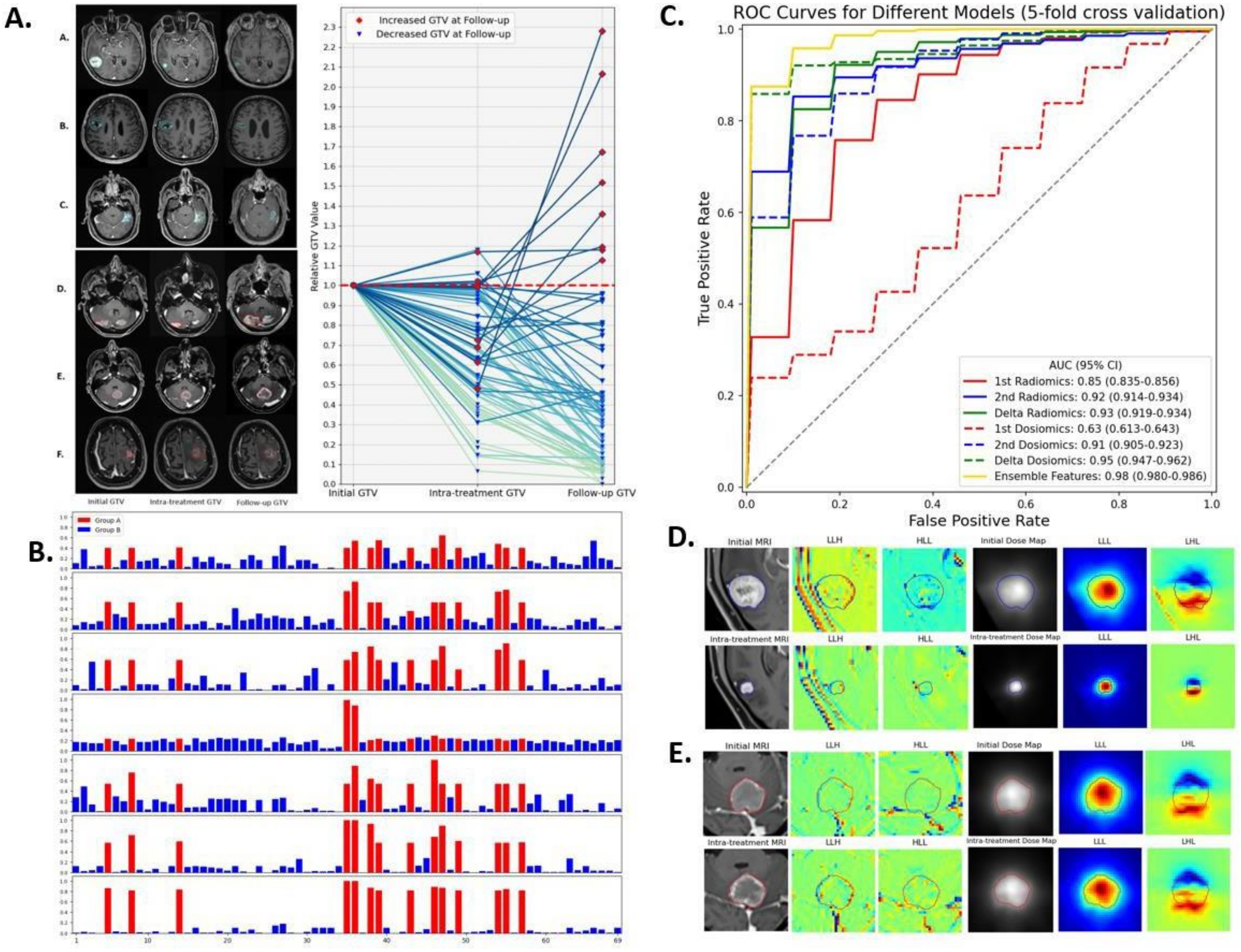


Fig 1. (A)Tumor volume analysis. Right plot: Lesion volumetric changes are examined at three time points. Left Plot: In Group 1, lesions exhibit a decreased GTV during follow-up, with intra-treatment variations categorized as A decreased, B unchanged, and C increased GTV. In contrast, Group 2 depicts lesions with an increased GTV during follow-up, featuring intra-treatment assessments of D decreased, E unchanged, and F increased GTV. (B) The classification scores of lesions among all 7models. (C) ROC curves for 7 models. (D) One with decreased GTV selected for exemplifying feature extraction through four wavelet transforms. (E) One with non-decreased GTV selected for exemplifying feature extraction through four wavelet transforms

Poster #2611 - A Critical Assessment of Using ChatGPT for Extracting Structured Data from Clinical Notes

Jingwei Huang[1], Donghan M. Yang[1], Ruichen Rong[1], Kuroush Nezafati[1], Colin Treager[1], Zhikai Chi[2], Shidan Wang[1], Xian Cheng[1], Yujia Guo[1], Laura J. Klesse[3], Guanghua Xiao[1], Eric D. Peterson[4], Xiaowei Zhan[1], Yang Xie[1]

[1]Quantitative Biomedical Research Center, Peter O'Donnell Jr. School of Public Health, UT Southwestern Medical Center; [2]Department of Pathology, UT Southwestern; [3]Department of Pediatrics, UT Southwestern; [4]Department of Internal Medicine, UT Southwestern

Objective: Existing natural language processing (NLP) methods to convert free-text clinical notes into structured data often require problem-specific annotations and model training. This study aims to evaluate ChatGPT's capacity to extract information from free-text medical notes efficiently and comprehensively, and evaluate how well can ChatGPT do and what are its strength and weakness?

Methods: We developed a large language model (LLM)-based workflow, utilizing systems engineering methodology and spiral "prompt engineering" process, leveraging OpenAI's API for batch querying ChatGPT. We evaluated the effectiveness of this method using a dataset of more than 1,000 lung cancer pathology reports and a dataset of 191 pediatric osteosarcoma pathology reports, comparing the ChatGPT (gpt-3.5-turbo) outputs with expert-curated structured data.

Findings: ChatGPT-3.5 demonstrated the ability to extract pathological classifications with an overall accuracy of 89%, in lung cancer dataset, outperforming the performance of two traditional NLP methods. The performance is influenced by the design of the instructive prompt. Our case analysis shows that most misclassifications were due to the lack of highly specialized pathology terminology, and erroneous interpretation of TNM staging rules. Reproducibility shows the relatively stable performance of ChatGPT-3.5 over time. In pediatric osteosarcoma dataset, ChatGPT-3.5 accurately classified both grades and margin status with accuracy of 98.6% and 100% respectively.

Interpretation: Our study shows the feasibility of using ChatGPT to process large volumes of clinical notes for structured information extraction without requiring extensive task-specific human annotation and model training. The results underscore the potential role of LLMs in transforming unstructured healthcare data into structured formats, thereby supporting research and aiding clinical decision-making.

Poster #2612 - Characterizing Tumor Heterogeneity Integrating Cell Differentiation Status and Spatial Evolution Trajectory of Lung Cancer

Yang Liu[1], Ruichen Rong[1], Shidan Wang[1], Liwei Jia[2], Peiran Quan[1], Tingyi Wanyan[1], Guanghua Xiao[1,3,4,#], Yang Xie[1,3,4,#]

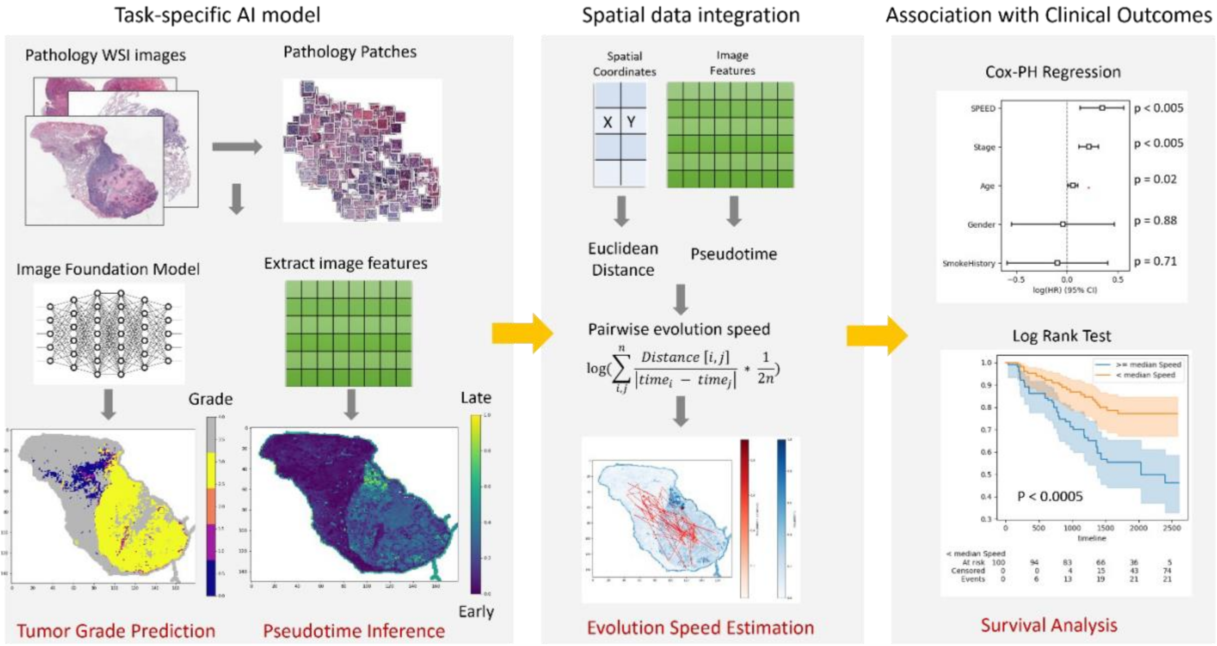
[1]Quantitative Biomedical Research Center, Peter O'Donnell Jr. School of Public Health, UT Southwestern Medical Center, Dallas, TX; [2]Department of Pathology, UT Southwestern; [3]Department of Bioinformatics, UT Southwestern; [4]Simmons Comprehensive Cancer Center, UT Southwestern
[#]Corresponding authors

Abstract: A key challenge for successful cancer treatments is intra-tumor heterogeneity due to the morphological and genetic diversity of cells in the tumor microenvironment (TME). The status of tumor differentiation from the normal cells (tumor grade) and their evolution trajectory within TME are essential factors in determining tumor progression and treatment strategy. Traditionally, the tumor grade diagnosis is based on only specific regions of interest (ROIs) on the histopathology whole-slide images (WSI), which ignores information from the rest of the areas. And the process requires decisions from several experienced pathologists. In this work, we developed a deep- learning computing model to identify different tumor grades efficiently, and we also demonstrated the efficacy of inferring evolution trajectory using only image features. To characterize the intra-tumor heterogeneity and the association with clinical outcomes, we also derived a signature, integrating the histological spatial information and trajectories, describing the tumor differentiation speed at the WSI level.

As depicted in Figure 1, we trained our task-specific model based on the pre- trained residual network on the extracted image patches from the NLST dataset. Pseudotime was calculated using the PAGA algorithm from the features extracted from the model. The average evolution speed is calculated according to the pairwise Euclidean distance and pseudotime.
Validated across multiple lung cancer datasets, our results show the proposed speed signature derived from the tumor grade prediction model is a potential biomarker of patient survival.

Overall, we developed the first *in silico* AI-assisted approach to help efficiently assess the tumor grade from the histopathology images. Together with potential application in discovering novel biomarkers and further understanding the tumor microenvironment, our work will positively facilitate the cancer research and clinical practice of cancer treatment.

Figure 1.


Task-specific AI model
Pathology WSI images
Pathology Patches
Image Foundation Model
Extract image features
Grade
Late
Early
Tumor Grade Prediction
Pseudotime Inference
Spatial data integration
Spatial Coordinates
Image Features
X
Y
Euclidean Distance
Pseudotime
Pairwise evolution speed
$\log(\sum_{i,j}^{n} \frac{Distance\,[i,j]}{|time_i - time_j|} * \frac{1}{2n})$
Evolution Speed Estimation
Association with Clinical Outcomes
Cox-PH Regression
SPEED
Stage
Age
Gender
SmokeHistory
p < 0.005
p < 0.005
p = 0.02
p = 0.88
p = 0.71
−0.5
0.0
0.5
log(HR) (95% CI)
Log Rank Test
>= median Speed
< median Speed
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
P < 0.0005
0
500
1000
1500
2000
2500
timeline
< median Speed
At risk 100 94 83 66 36 5
Censored 0 0 4 15 43 74
Events 0 6 13 19 21 21
Survival Analysis

Poster #2616 - Rubrics to Prompts: Deploying Zero-Shot Large Language Models for Automatic, Expert-Level Grading of Medical Student Encounter Notes

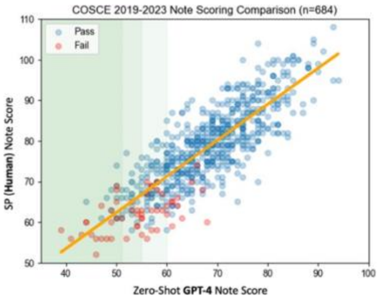
Andrew R. Jamieson, Ph.D. [1,*]; Michael J. Holcomb, M.S. [1,*]; Thomas O. Dalton, M.D. [2]; Kystle K. Campbell, D.H.A. [3]; Sol Vedovato, M.S. [1]; Gaudenz Danuser, Ph.D. [1]; Daniel J. Scott, M.D.[3,4]

[1]Lyda Hill Department of Bioinformatics, UT Southwestern Medical Center; [2]Department of Internal Medicine, UT Southwestern; [3]Simulation Center, UT Southwestern; [4]Department of Surgery, UT Southwestern

Objective: The post-exam learner note is an important component of the Objective Structure Clinical Exam (OSCE), measuring medical student competence via participation in a live-action simulated patient encounters with human actors. Assessment of student performance on this task is labor intensive, requiring documents to be reviewed by a specially trained evaluator. We demonstrate that large language models (LLMs) enable a new era of automatic, rubric-based grading, *without* the requirement of *any* prior training data or labels (i.e. “zero-shot”).

Methods: Notes from 684 3rd year UTSW medical students across 4 years (2019-2023) captured at UTSW’s Simulation Center during the “Comprehensive” OSCE (COSCE) were analyzed, composed of 10 stations leading to over 6840 encounters/notes. Grading rubrics, consisting of a total of 165 individually scored components (for a grand total of over 100,000 items), were provided for each of the stations along with human assessed/labeled scores for all students. First, base-line performance using supervised machine learning models (fundamentally reliant sufficient training data with ground-truth labels) was established. Next, we bypassed the need for prior training data by adopting zero-shot inference with large-language models (LLMs) (e.g. GPT-4). Inspired by the concept of textual-entailment, we developed a straight-forward framework for mapping rubric items into a series of minimally engineered prompts. To validate our approach, a *completely new set* of OSCE stations/scenarios *with different rubrics* were evaluated (Fall 2022 cohort, n=229, 10 stations, 2320 notes). Experiments extended to open-source, smaller foundation models to compare performance to proprietary frontier models (e.g., GPT-4, GPT-3.5), both in the zero-shot regime with *Mixtral-8x7B-Instruct-v0.*1 and with supervised fine-turning by using GPT-4 outputs from the separate COSCE data to train Llama-2-7B-chat.

Results: Human-level grading of OSCE notes was achieved: zero-shot GPT-4 had 83.2% Item-Level Agreement with standard patient evaluator (SPE) grades and a Spearman Rank correlations of 0.82 ($p <= 0.001$) across *all students and years* for the “COSCE” notes. On the 2nd set of OSCE stations with different rubrics/scenarios, the open-source, *locally tuned model, llama- 2-7B-chat (86.6%), nearly matched the zero-shot GPT-4 item- level performance (89.7%)*. Mixtral (83.0%) outperformed GPT3.5 (73.6%) on the 2nd OSCE set with zero-shot inference.



Conclusions & Impact: This technology enables deployment, in *near-real-time*, of bespoke evaluation schemes for high-fidelity assessment of human performance, unlocking pedagogical innovations within the medical educational arena and with practical implications for performance assessment more generally. Further, smaller, locally tunable foundation models can achieve high- performance when provided high-quality examples from a teacher model. Finally, UTSW piloted concurrent deployment of this AI system for grading medical student notes during the Fall 2023 OSCEs, replacing >90% of human grading and achieving cost-saving, rapid (from *months to days*), and accurate note assessment -- improving the overall quality of the OSCE process.

Poster #2617 - RadPointGPT: A Chatbot Enhanced by Retrieval Augmented Generation and a Large Language Model for Improved Access to Internal Documents

Paulo E.A. Kuriki, Fernando Kay, Cecelia Brewington, Ronald Peshock

Department of Radiology, UT Southwestern Medical Center, Dallas, TX

Background/Current Status: Standard Operating Procedures (SOPs) are crucial for establishing guidelines and protocols in radiology. However, they often result in an overwhelming number of documents, making it challenging to locate specific information. Traditional search methods struggle with natural language, highlighting the need for a more effective retrieval system.

Goal/Analysis: The goal is to create a chatbot that allows users to consult information from SOP documents using natural language conversation.

Intervention/Results: We developed a chatbot that combines Large Language Models (LLMs) with Retrieval Augmented Generation (RAG). We tokenized and stored 252 SOP documents in a vector database. A chatbot interface was developed to allow users to interact using natural language. Questions are queried based on similarity in the vector database, retrieving relevant documents. These documents are then sent to a local Mistral-7B model to generate an answer. This approach was chosen for its speed and security in handling sensitive information. An Evaluation Process was implemented using GPT-4. 10-20 questions per document were generated and submitted to our pipeline. For each response, the retrieval accuracy and response relevance were calculated. This enabled us to refine our method, including embedding optimization, retrieval re-ranking, and metadata rewriting. As a result, the global F1-Score reached 93.5% with an accuracy of 87.8%.

Conclusions: Combining the RAG technique with LLMs represents a significant advancement in document retrieval, offering an efficient and secure method to query SOPs. The Evaluation Pipeline was pivotal in identifying inefficiencies. Future efforts will focus on refining the retrieval process, mitigating biases, and incorporating user feedback for continuous system improvement.

Poster #2619 - Identifying Social Determinants of Health in Medical Notes: A Large Language Model Approach

Zifan Gu[1], Jingwei Huang[1], Ann Marie Navar[2], Eric D. Peterson[2], Guanghua Xiao[1,3,4], Yang Xie[1,3,4], Donghan M. Yang[1]

[1]Quantitative Biomedical Research Center, Peter O'Donnell Jr. School of Public Health, UT Southwestern Medical Center; [2]Department of Internal Medicine, UT Southwestern; [3]Department of Bioinformatics, UT Southwestern; [4]Simmons Comprehensive Cancer Center, UT Southwestern

Objectives and Research Questions: Social determinants of health (SDOH) have been shown to affect health outcomes across a range of clinical conditions. However, research on SDOH and application of risk models that incorporate SDOH are significantly hindered by the difficulty of acquiring these data on a patient level. Medical text notes contain a wealth of SDOH information in an unstructured format that could be utilized for these purposes. Conventional natural language processing (NLP) techniques require laborious and expensive manual labeling, rule design, and model training, limiting their application scale. The advent of large language models (LLMs) offers an efficient potential solution for extracting medical concepts from notes via zero- shot/few-shot prompting. Here, we explore the feasibility of harnessing LLMs to extract SDOH data from real-world medical notes at practically acceptable accuracy levels.

Methods: In our pilot study, we utilized a dataset consisting of 5,321 sentences from 200 clinical notes (65 by nurses; 70 by physicians; 65 by social workers) of 183 patients in the Medical Information Mart for Intensive Care (MIMIC-III) database which is pre-labelled for SDOH[1]. Two generative pre-trained transformers (GPTs) were tested ("GPT-3.5-turbo" and "GPT-4") using
few-shot prompting with an iterative prompt engineering approach. We tasked the models to determine whether a sentence contains information on two SDOH categories: marital status and employment status. Results are evaluated by precision, recall, and F1. We further compared the performance of these GPTs against the best performing model (fine-tuned Flan-T5 XL) previously reported on the same dataset[1].

Results and Discussion: Our approach achieved F1 of 0.90 and 0.65 in identifying adverse marital status (widowed, divorced, single) and employment status (unemployed, underemployed, disability), respectively, outperforming the previously reported fine-tuned Flan-T5 XL model (Table 1). Performance slightly decreased in identifying non-adverse marital status (married, partnered) and employment (employed, retired, student) status. A substantial improvement in performance was achieved with GPT-4 compared to GPT-3.5 using identical prompts. High F1s (> 0.99) for identifying non-SDOH-containing sentences, coupled with high recalls (> 0.71) in detecting SDOH evidence, underscore a use case for automated selection of SDOH-relevant notes. Future work will evaluate the GPT approach on a larger independent testing dataset and a broader array of SDOH categories[2].

Conclusions: Our few-shot prompting approach leveraging GPT-4 demonstrates the effectiveness of LLMs in identifying adverse SDOH data elements from real-world medical notes, highlighting its potential to extract and augment structured SDOH data for clinical research.

***Table 1**. Model performance for extracting SDOH attributes.*

			GPT-4			*GPT-3.5*	*Flan-T5 XL*[1]
	Attribute	Count (sentence)	Precision	Recall	F1	F1	F1
Marital status	Adverse	23	0.84	0.96	**0.90**	0.29	0.68
	Non-adverse	157	0.77	0.74	0.76	0.86	na*
	No evidence	5141	1.00	0.99	1.00	0.99	na*
Employment status	Adverse	22	0.57	0.77	**0.65**	0.59	0.55
	Non-adverse	45	0.40	0.71	0.51	0.44	na*
	No evidence	5256	1.00	0.99	0.99	0.99	na*

*na – Not reported in previous work[1].

References:

1 Guevara, M. *et al.* Large language models to identify social determinants of health in electronic health records. *NPJ Digit Med* **7**, 6 (2024). https://doi.org:10.1038/s41746-023-00970-0

2 Ahsan, H., Ohnuki, E., Mitra, A. & Yu, H. MIMIC-SBDH: A Dataset for Social and Behavioral Determinants of Health. *Proc Mach Learn Res* **149**, 391-413 (2021).

Poster #2620 - A Deep Learning Model for Clinical Outcome Prediction Using Longitudinal Inpatient Electronic Health Records

Ruichen Rong[1,#], Donghan M. Yang[1,#], Zifan Gu[1,#], Hongyin Lai[1], Tanna Nelson[2], Tony Keller[2], Clark Walker[2], Kevin W. Jin[1], Catherine Chen[3], Eric D. Peterson[3], Ferdinand Velasco[2], Guanghua Xiao[1,4,5], Yang Xie[1,4,5]

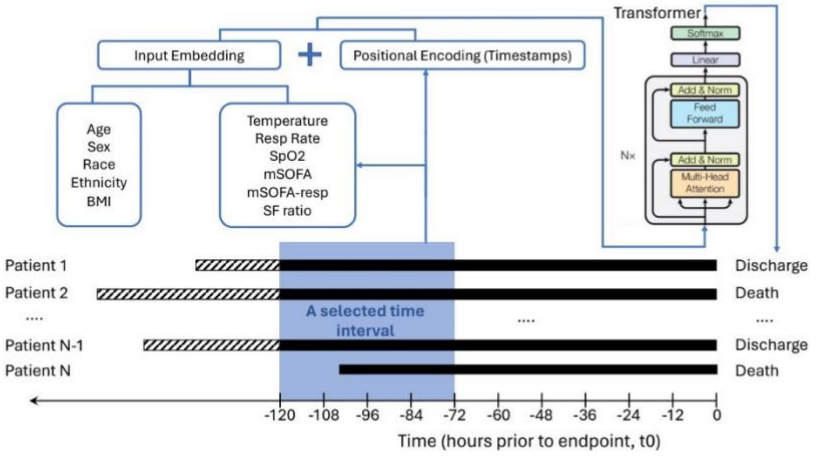
[1]Quantitative Biomedical Research Center, Peter O'Donnell Jr. School of Public Health, UT Southwestern, Dallas, TX; [2]Texas Health Resources, Arlington, TX; [3]Department of Internal Medicine, UT Southwestern; [4]Department of Bioinformatics, UT Southwestern; [5]Simmons Comprehensive Cancer Center, UT Southwestern
[#]Contributed equally as first authors

Objective: Recent advances in deep learning show significant potential in analyzing long- range, time-series electronic health records (EHR) data for clinical outcome prediction. We aim to develop a Transformer-based, Encounter-level Clinical Outcome (TECO) model to predict intensive care unit (ICU) outcomes using inpatient EHR data.

Materials and Methods: We developed and validated TECO using various baseline and time-dependent variables from 2579 hospitalized COVID-19 patients to predict ICU outcomes (death vs. discharge). We then externally validated TECO in two non-COVID-19 cohorts available from the Medical Information Mart for Intensive Care (MIMIC)-IV: an ARDS cohort (n=2799) and a Sepsis cohort (n=6622).

Results: In the COVID-19 cohort, TECO achieved a higher area under the receiver operating characteristic (AUROC; 0.88–0.98) compared to Epic Deterioration Index, random forest, and XGBoost. In the two external MIMIC cohorts, TECO outperformed other models with an AUROC ranging from 0.67 to 0.89 depending on the data intake time interval. Additionally, TECO identified clinically interpretable features correlated with the outcome.

Conclusions and Impacts: TECO outperformed proprietary methods as well as conventional machine learning models in predicting ICU outcomes among COVID and non- COVID populations. Further validation is needed, but TECO could be powerful in providing negative outcome guidance for various diseases in the inpatient setting.



Poster #2621 – Towards A Personalized Nutrition Assistant for Dietitians and Patients

Hari Vennelakanti, PT, D.P.T., CLWT[1]; Yasbanoo Moayedi, M.D., M.H.Sc.[2,3]; Steven G. Hershman, Ph.D., M.S., M.S.E.[4]

[1] Therapy Services, UT Southwestern, Dallas, TX; [2]Ted Rogers Centre for Heart Research, Ajmera Transplant Centre, University of Toronto, Toronto, Canada; [3]Ajmera Transplant Centre, University of Toronto, Toronto, Canada; [4]School of Information, The University of Texas at Austin, Austin, TX

Introduction: In the United States, poor diet contributes to 20% of healthcare costs related to heart disease, stroke, and diabetes. While dietitians guide dietary changes, their limited availability and the complexity of personal dietary needs make sustained change challenging. This qualitative study evaluates whether the latest large language models (LLMs), particularly the Gemini models, can support dietitians or directly assist patients by providing personalized nutritional advice.

Methods: We prompted these models, including Gemini 1.5 and 1.0 with and without reference documents, to generate meal plans and dietary suggestions across various medical conditions.

Results: Even with zero-shot prompting, Gemini models can create meal plans for a variety of medical conditions. Gemini 1.5, but not 1.0, is particularly effective at giving culturally tailored recipe recommendations for ethnic groups like the Cree and Zuni people, with recipes for culturally associated foods such as pemmican and pozole (though the model used the Anglicized spelling "posole"). These models can offer practical advice for dining out or modifying recipes to suit medical diets. However, some recommendations may need review by medical nutritionists and chefs due to potential inaccuracies.

Conclusion: Our findings suggest that LLMs, especially newer versions like Gemini 1.5, have potential to augment dietetic practice with features like culturally tailored plans and actionable dietary advice. Future research will focus on prioritizing and validating these features through a quantitative study with certified dietitians evaluating the models' recommendations for accuracy and applicability.

Poster #2623 - Distant Metastasis Prediction of Head and Neck Cancers Using Graph Neural Networks

Erich Schmitz, Jing Wang

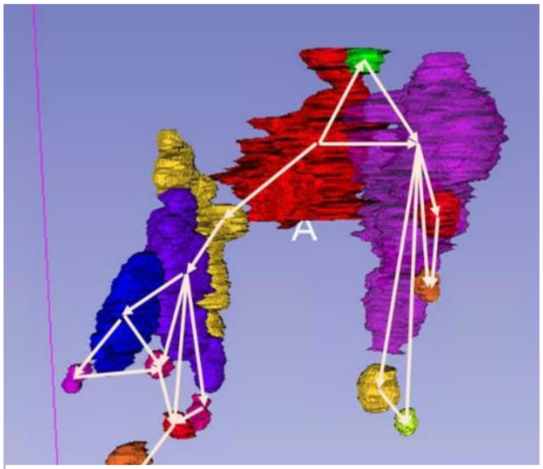
Medical Artificial Intelligence and Automation Lab, Department of Radiation Oncology, UT Southwestern Medical Center

Purpose: The incidence of metastatic cancer is often a fatal prognosis for cancer patients. The evaluation of such a risk is an important part of the clinical decision-making process. To take advantage of available imaging and clinical information machine learning is commonly used in predictive tasks and recently more attention has been given to graphs and graph neural networks and their ability to exploit information about the interconnectivity of biological systems. In head and neck cancer (HNC), it is common for the cancer to spread to local lymph nodes (LNs), and graphs have been used to model the connections between the primary tumor and those involved LNs. For this study we built a model that incorporates both extracted imaging features and connections between primary tumors and their involved LNs in the form of a graph to predict the occurrence of distant metastasis.

Methods: Using CT scans and accompanying masks of gross tumor volumes of the primary tumor (GTVp) and involved LNs (GTVns), directed graphs are made to model the potential cancer spread from the GTVp to the GTVns. Each region of interest (ROI) is treated as a node in the graph while the connections between ROIs are determined using hierarchical clustering based on the physical distance from the GTVp. Imaging features are extracted from each ROI using a Residual Network (ResNet) and used as node features in the graphs. The graphs are used to train a residual gated Graph Convolutional Network (GCN) to predict the risk of distant metastasis within 2 years. The performance of the GCN is compared with that of a ResNet trained only on the ROI of the GTVp.

Results: The residual-gated GCN achieved an area under the Receiver Operating Characteristic curve (ROC-AUC) of 0.732 on the validation dataset using nested 5-fold Cross Validation. This is compared to a ROC-AUC of 0.622 using the ResNet and GTVp ROI only.

Conclusion: By incorporating LN information and their connectivity with the primary tumor into a graph we achieved an improved performance in predicting distant metastasis compared to a simple ResNet using only primary tumor information.



Example of a primary tumor (red, center) and its connections to the involved LNs

Poster #2625 - Efficient Individual Treatments Effect Estimation Using Causal Transformers in Federated Learning Settings

Disha Makhija[1], Joydeep Ghosh[1,2], Yejin Kim[3]

[1]The University of Texas at Austin, Austin, TX; [2]Dell Medical School, Austin, TX; [3]University of Texas Health at Houston, Houston, TX

Introduction: The use of Artificial Intelligence (AI) systems to estimate the individualized treatment effects (ITE) of interventions have facilitated large-scale decision-making in several healthcare applications. The conventional estimation methods often require acquisition of substantial amounts of costly data per considered intervention. In this work, we present a novel framework for collaborative learning of heterogeneous ITE estimators across distributed institutions (such as hospitals) via Federated Learning, thus allowing training on a large and diverse dataset without the need to share any sensitive health data information. The proposed method is flexible to handle diverse patient populations and non-identical patient measurements (co-variates) across different sites and allows for the estimation of treatment effects for different treatments being administered across these sites. This method can also be readily adapted to predict the effects of new and unseen treatments.

Method: The proposed framework involves learning a common representation space for the co-variates across the participating institutions and using local multi-task learning to predict the outcomes for interventions being administered locally at each institution, as depicted in Figure 1(b). The representation learning module is based on a transformer architecture -- a core building block of GenAI models -- but uses a novel variant that considers causality. The flexibility of the framework allows it to accommodate non-identical feature sets (co-variates) and diverse treatments across sites (Figure 1(a)).

Results and Conclusions: We tested our method by collecting data from 3 randomized clinical trials for intracerebral hemorrhage (ICH): ATACH2 (NCT01176565), MISTIE3 (NCT01827046) and ERICH (NCT01202864), across 3 hospital locations, where each location has a set of patient-level pre-treatment measurements (such as demographics and clinical presentation of the ICH), a binary treatment variable, and the outcomes measured as the modified Rankin Score (mRS) representing patient's severity. We observe that our method outperforms current SOTA methods as measured by the factual RMSE and Average Treatment Effect on the Treated (ATT) with about 60% reduction in RMSE and 72% reduction in ATT. Also, a model interpretability approach based on examining the attention heads learned by the transformer model reveals that these heads effectively capture essential disease-related information, thereby attesting to the meaningfulness and explainability of the learned representations.

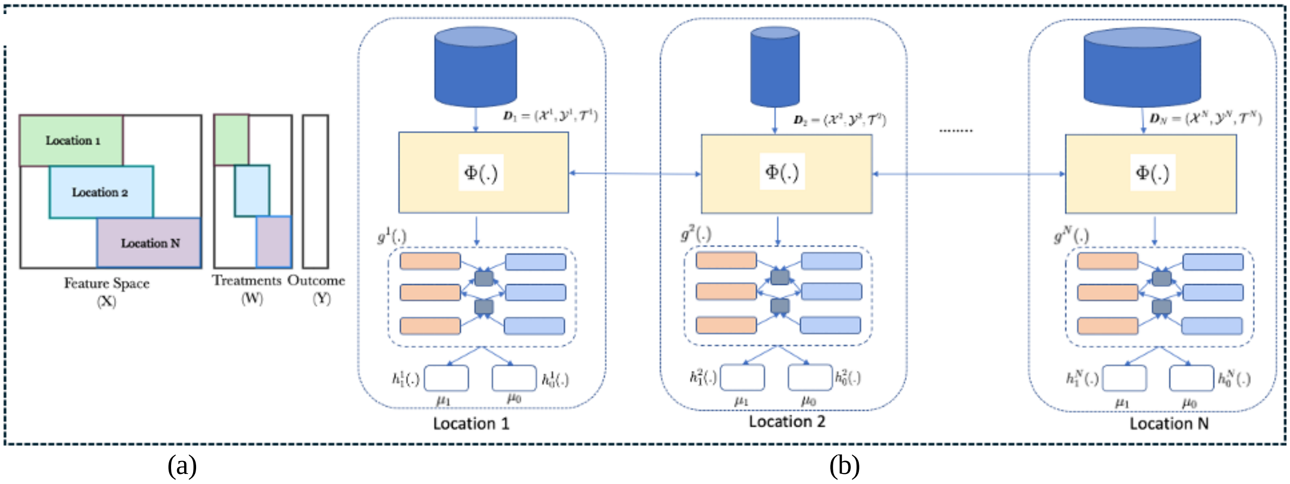


Figure 1 : (a) an overview of the setting with 3 locations, (b) a schematic diagram of the Federated Learning solution.

Poster #2626 - Utilizing Large Language Models (LLMs) to Enhance Clinical Trial Matching

Jacob Beattie, B.Sc.[1,2]; Sarah Neufeld, M.B.A.[2]; Daniel Yang, M.D.[1,2]; Christian Chukwuma, B.Sc.[2]; Ahmed Gul, B.Sc.[2]; Neil Desai, M.D.[2]; Steve Jiang, Ph.D.[1,2]; Michael Dohopolski, M.D.[1,2]

[1]Medical Artificial Intelligence and Automation Lab, UT Southwestern Medical Center, Dallas, TX;
[2]Department of Radiation Oncology, UT Southwestern

Purpose: To utilize Large Language Models (LLMs) to improve clinical trial screening efficiency and accuracy while maintaining clinician-friendly implementation.

Methods: We collected records from 35 patients enrolled in a phase II clinical trial and 39 patients seen by the same physician group but not enrolled. Both groups were manually screened for the same phase II trial. 10 patients from each group were reserved for prompt fine-tuning, while the rest were used for testing. Patient electronic health records were embedded in a vector database to facilitate relevant information retrieval. GPT-3.5 Turbo and GPT-4 APIs were then used to process the prompts and relevant documents. We tested four different prompting styles, investigating the impact of structured/unstructured response requirements, human-made expert guidance, and fully automated reasoning based on LLM-generated reasoning structures. The results were compared to ground truth eligibility statuses. Errors were manually reviewed and documented.

Results: Among the enrolled group, GPT-3.5 Turbo achieved the best performance when eliciting structured output using human-made expert guidance, with an accuracy of 94% and sensitivity of 94%. Among the non-enrolled group, GPT-4 achieved the best performance using the same approach, with an accuracy of 88%, sensitivity of 89%, and specificity of 83%. Notably, fully automated approaches exhibited similar performance among the comparison group with an accuracy of 88%. Incorrect classifications were due to insufficient document retrieval, incorrect temporal reasoning, or incorrect understanding of criteria.

Conclusion: These findings highlight the significant potential of LLMs in the clinical trial matching process, particularly the promise of fully automated systems. Implementing automated screening across multiple clinical trials with minimal programming knowledge could greatly reduce research staff workload and improve clinical trial enrollment.

Poster #2628 - Mechanism Interpretable Drug-Gene Interaction(MIDI) Transformer for Cancer Drug Effect Prediction

Tingyi Wanyan, Jiwoong Kim, Ling Cai, Rongqing Yuan, Ruheng Wang, Ruichen Rong, Peifeng Ruan, Tao Wang, Xiaowei Zhan, Guanghua Xiao, Yang Xie

UT Southwestern Medical Center, Dallas, TX

Background: Cancer Drug discovery using genetic information is under development. Precisely locating drug atoms on explaining the targeting effect is crucial in precision medicine since it helps understand the drug's mechanism of action. To the current date, a lot of data has been collected regarding drug response against cancer cell lines, as well as many models that predict the drug response based on genomic information. However, to our knowledge, none of the data-driven techniques propose to detect the targeting mechanism of small drug molecules against gene expression pathways.

Objective: The objective of this project is to develop an original model to predict cancer drug effect, and provide detailed interpretations of the mechanism of action on specific cancer drugs, as well as detect the important gene signaling pathways that the drug is targeting.

Methods: In this work, we propose MIDI(Mechanism Interpretable Drug-Gene Interaction) transformer to delve into the targeting relation between drug molecular against gene expression pathways, we show that purely based on a data-driven approach, the attention mechanism in our model could capture the important binding effect of small molecular towards gene pathways. We incorporate the two most advanced architectures (Geneformer and Graphformer) to model both genetic input and drug input. we provide both theoretical derivation and experiment results to show the information flow regarding the attention mechanism.

Results: Figure 1 shows the drug self-attention heat map output from our model, it shows the important atom binding site of a specific drug. We demonstrate that with the Interpretation mechanism, our model presents much higher prediction performance than the other state-of-the-art drug response prediction models. We show that our self-attention heat map can focus mostly on the binding site of drug small molecular structures to the targeted protein region.

Conclusions: Target identification and mechanism of action in chemical biology is important to drug discovery. Based on our knowledge, Currently, no technique provides the functionality of discovering the binding effect between drug compounds with oncogene signaling pathways. As well as detecting the crucial binding atoms within a certain drug.

Our model possesses great potential to open a way of generating precise medicine given patient gene expression values on the tumor region. We are developing and modifying more profound multi-modality foundation models based on the current model**.**

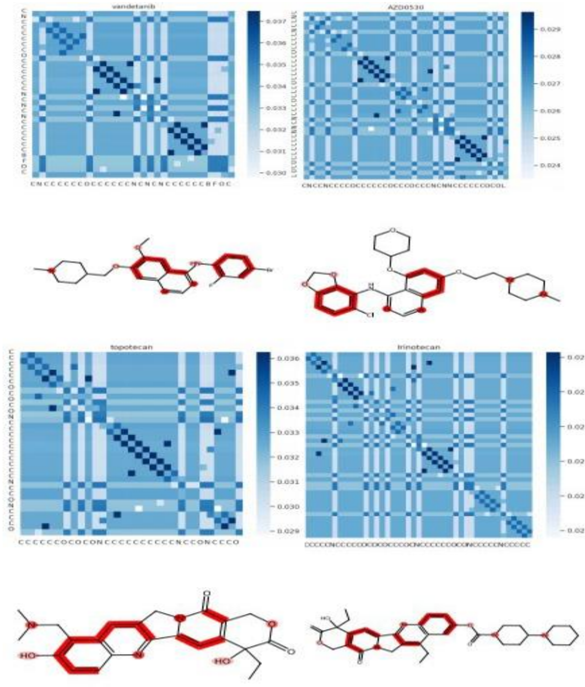


Figure 1. Drug self-attention heat map

Poster #2630 - Assessing Disease Severity in Cutaneous Lupus Patients Using Natural Language Processing

Kuroush Nezafati, M.D.[1,#]; Laura Wang, B.A.[2,#]; Ruichen Rong, Ph.D.[1,#]; Andrew Park, M.D.[2]; Jane Zhu, M.D.[2]; Guanghua Xiao, Ph.D.[1]; Yang Xie, Ph.D.[1]; Donghan M. Yang, Ph.D.[1,*]; Benjamin F. Chong, M.D.[2,*]

[1]Quantitative Biomedical Research Center, Peter O'Donell Jr. School of Public Health, UT Southwestern, Dallas, TX;
[2]Department of Dermatology UT Southwestern
[#]Contributed equally; *Corresponding authors

Background: Cutaneous lupus erythematous (CLE) is an autoimmune skin disorder that manifests as inflammatory cutaneous lesions in photosensitive areas. The Cutaneous Lupus Erythematosus Disease Area and Severity Index (CLASI) is a clinically validated tool for measuring disease activity and damage in Cutaneous Lupus Erythematosus (CLE) patients. A viable strategy for mining of physical examination (PE) notes of CLE patients can substantially expand the applications of CLASI in research and patient care. Recent advancements in natural language processing (NLP) provide an unprecedented opportunity to effectively mine free-text medical notes for this purpose.

Objective: To develop a deep learning-based NLP algorithm for estimating CLASI activity and damage status using pre-existing PE notes in an EHR system.

Methods: PE notes containing clinical exam descriptions of CLE lesions were collected from patients enrolled in the UTSW CLE Registry. A total of 124 PE notes were included for 50 CLE patients (median age 46 years, IQR 37-55 years; 86% female). The cohort was split into a training (n=24; 89 notes) and test set (n=26; 35 notes). For each exam, skin disease assessments were also performed by a dermatologist (BFC) with subsequent CLASI activity and damage scores labeled by clinically trained personnel serving as the ground truth. A BERT (Bidirectional Encoder Representations from Transformers)-based model was trained to extract these CLASI-related entities and attributes, based on which CLASI scores were calculated (Figure 1). We evaluated the performance of the NLP algorithm in estimating 1) individual CLASI activity and damage scores using linear correlation and 2) dichotomized CLE severity levels (mild vs. moderate/severe) based on estimated activity and damage scores using confusion matrix, accuracy, precision, recall, and F1.

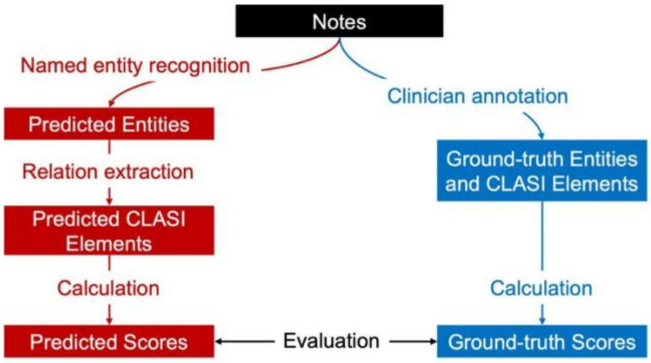


***Figure 1**. NLP model development workflow.*

Results: In predicting CLASI scores, the model results had a correlation of 0.79 and 0.82 with the ground truth for activity and damage, respectively, in the training set, and a correlation of 0.64 and 0.88 in the test set. The algorithm estimated CLASI-based severity levels at an accuracy of 0.91 (activity) and 0.89 (damage) in the test set.

Conclusions: Using the routine free-text PE notes, our NLP algorithm can assess CLE disease severity, offering significant potential in efficiently generating new data for CLE research and care.

Poster #2631 - Encoding Patient-Specific Dosimetric Goals as Three-Dimensional Inputs for Tailored Radiotherapy Dose Prediction

Austen Matthew Maniscalco; Hui-Ju Wang; Steve Jiang, Ph.D.; Dan Nguyen, Ph.D.

Medical Artificial Intelligence and Automation Laboratory, Department of Radiation Oncology, UT Southwestern Medical Center, Dallas, TX

Background: Creating radiation therapy treatment plans is a complex process, tailored to each patient's unique disease and anatomy. Physicians generate unique dosimetric directives based on patient-specific considerations, which dictate the dosimetric objectives for planning target volumes (PTVs) and organs at risk (OARs). A treatment plan is then manually crafted that achieving all requested objectives while optimally maximizing dose to the PTVs and minimizing dose to the OARs. However, determining the optimal treatment plan is often subjective due to conflicting dosimetric tradeoffs between segmented volumes.

Purpose: Deep learning models are used for dose prediction, taking a patient's computed tomography (CT) image and PTVs/OARs as input to predict a dose distribution based on anatomical features. However, these deep learning models lack stylistic output and may blends styles learned from the training data. We aim to develop a method that encodes dosimetric objectives in a 3D representation for input to a convolutional neural network (CNN), thereby training a model that can predict stylistic, patient-specific dose distributions. By enabling clinicians to modify dosimetric objectives and

re-input them into the model, they can assess various dose prediction tradeoffs. We anticipate our physician-directed (PD) model will show statistically significant differences in dose prediction performance over a baseline (BL) traditional model.

Methods: Our dataset includes 149 head and neck radiotherapy patients, divided into training, validation, and test subsets of 70%, 10% and 20%, respectively. Dosimetric goals were encoded as follows: Boolean OAR arrays were generated, dosimetric objectives identified sub-volumes of OARs to be spared, and relevant sub-volume voxels were assigned to objective value(s). For instance, if no more than 20% of the right lung can receive 20 Gy or more, then 80% of the right lung voxels were assigned a value of 20 Gy. We trained the BL dose prediction model using CT, PTV, a 3D map of Euclidean distance from PTV, and OARs as input data. The PD dose prediction model was trained using dosimetric goal arrays instead of OARs. Both models output a dose prediction, and loss is measured by its difference from the original plan's dose distribution.

Results: We predicted dose distributions for each test patient using the BL and PD models, and evaluated their difference from the original plans using mean absolute percent error (MAPE). The BL model yielded 2.21% MAPE, whereas the PD model yielded 1.85% (p=0.0072). We also qualitatively evaluated differences in dose prediction when a single organ's dosimetric goals were modified, setting the left parotid mean objective to 500 cGy, 2000 cGy, and 7500 cGy (Figure 1).

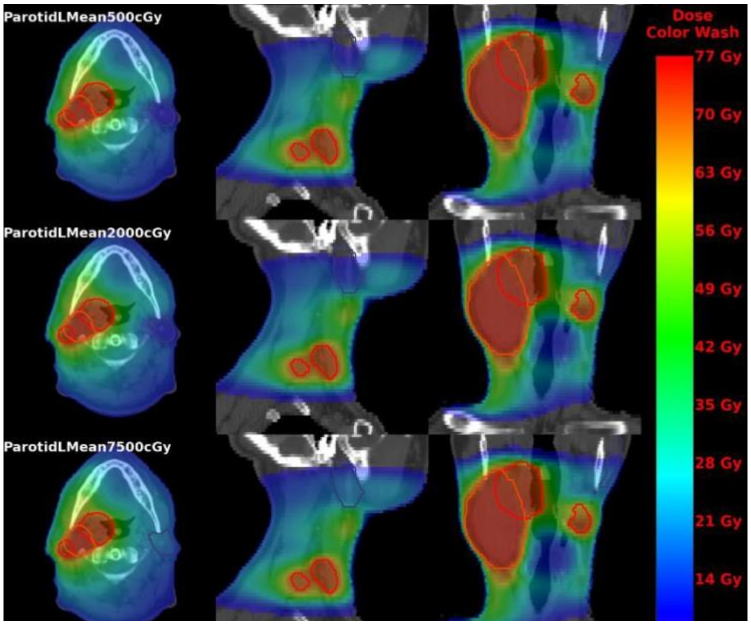


Figure 1

Conclusion: Our approach is innovative for its utilization of stylistic input and its removal of OAR-specific channel dependence, which may allow for disease-site-agnostic dose prediction in future works. The PD model demonstrated statistically significant differences in dose prediction and may offer greater flexibility than a traditional dose prediction model, increasing the personalization in radiotherapy dose prediction.

Poster #2632 - Opportunistic Screening for Osteoporosis Using Artificial Intelligence in Chest CT Scans: Validation Against DEXA Scans

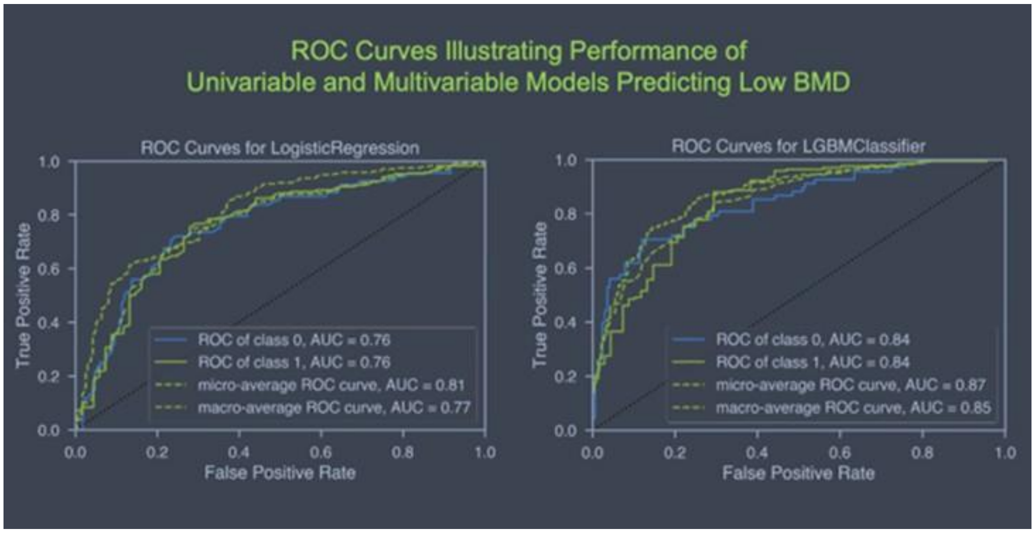
Angelo Scanio, MS2; Christopher Fan, MS2; Michael Long, M.S.; Roderick McColl, Ph.D.; Orhan K. Oz, M.D., Ph.D.; Ronald M. Peshock, M.D.; Fernando Kay, M.D., Ph.D.

Background: Osteoporosis, characterized by low bone mineral density (BMD), is a prevalent yet often undiagnosed condition, leading to significant morbidity and mortality. Dual-energy X-ray absorptiometry (DXA) is the current gold standard for measurement of BMD. Leveraging existing chest CT scans, this study proposes an opportunistic approach using an Artificial Intelligence tool, (AI Rad Companion, Siemens), along with a machine learning algorithm to predict low BMD as defined by DXA scans.

Methods: A total of 781 patients underwent a non-contrast chest CT and a DXA scan within a one-year period between Sept. 9, 2022, and Dec. 29, 2023. Utilizing Hounsfield units of the trabecular bone from chest CT scans and vertebral body height collected from AI-Rad companion, we trained a Light Gradient Boosting Machine Learning model (LightGBM) to predict low BMD from multiple covariates. A Logistic Regression (LR) model was derived using only Hounsfield units at the most caudal thoracic spine vertebra for comparison, which is the standard approach for opportunistic screening of low BMD on chest CT.

Results: In the unseen test cohort (30% of the sample), the LightGBM model achieved an area under the receiver operating characteristics curve of 0.84, in comparison to the LR model of 0.76 ($p<0.001$). The LightGBM model demonstrated higher accuracy and sensitivity in predicting low BMD, 79% and 87% respectively, compared to the chest CT radiological reports, 36% and 12%, respectively. However, the AI model reported a lower specificity of 59% compared to 94%.

Conclusion: Quantitative metrics extracted from chest CT scans using AI can be integrated with machine learning algorithms to predict low BMD. The LightGBM machine learning model demonstrated robust performance in detecting low BMD, outperforming the sensitivity of radiological reports. These findings advocate for a broader adoption of AI models in opportunistic screening for low BMD among patients undergoing chest CTs.

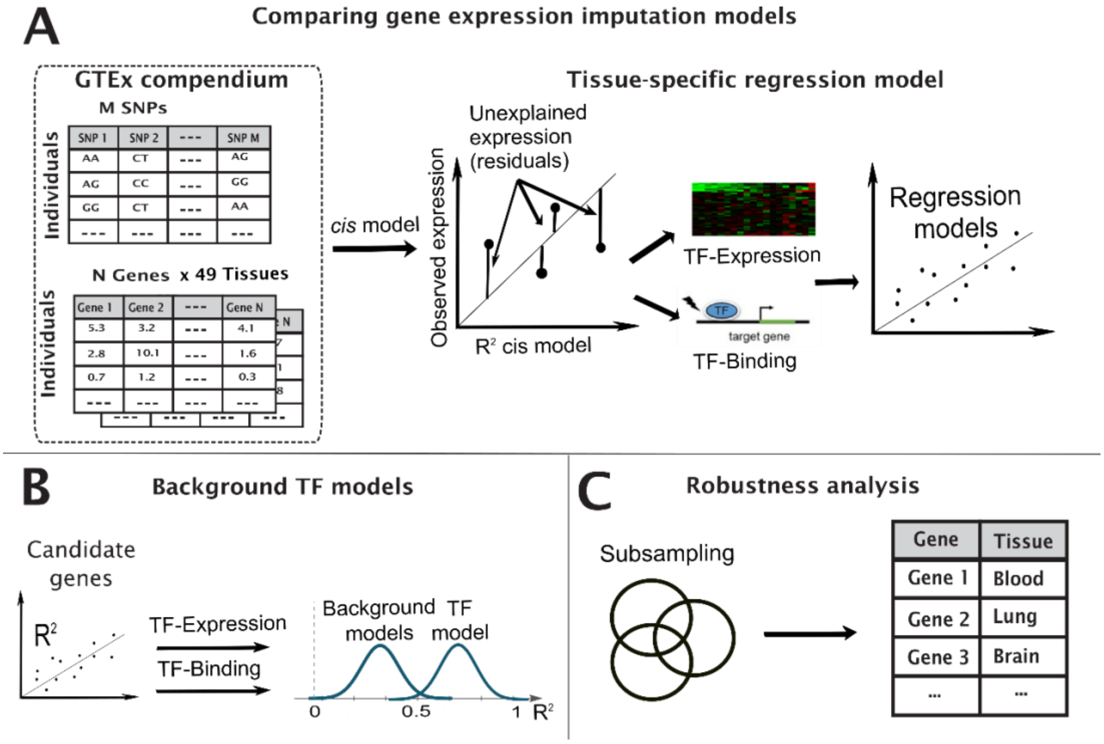


Poster #2633 -Tissue-Specific Atlas of Trans-Models for Gene Regulation Elucidates Complex Regulation Patterns

Robert D’Agostino, Assaf Gottlieb

Center for Precision Health, McWilliams School of Biomedical informatics, University of Texas Health Science Center at Houston, Houston, TX

Background: Deciphering gene regulation is essential for understanding the underlying mechanisms of healthy and disease states. While the regulatory networks formed by transcription factors (TFs) and their target genes has been mostly studied with relation to *cis* effects such as in TF binding sites, we focused on *trans* effects of TFs on the expression of their transcribed genes and their potential mechanisms.
Methods: We developed hypothesis-driven computational models to identify second-tier regulation by variability in TFs. Our approach considered two potential mechanisms – the combinatorial regulation by the expression of the TFs, and by genetic variants within the TF.
Results: We provide a comprehensive tissue-specific atlas, spanning 49 tissues of TF variations affecting gene expression. We demonstrate that similarity between tissues based on our discovered genes corresponds to other types of tissue similarity. The genes affected by complex TF regulation, and their modelled TFs, were highly enriched for pharmacogenomic functions, while the TFs themselves were also enriched in several cancer and metabolic pathways. Additionally, genes that appear in multiple clusters are enriched for regulation of immune system while tissue clusters include cluster-specific genes that are enriched for biological functions and diseases previously associated with the tissues forming the cluster. Finally, our atlas exposes multilevel regulation across multiple tissues, where TFs regulate other TFs through the two tested mechanisms.
Conclusions: Our tissue-specific atlas provides hierarchical tissue-specific *trans* genetic regulations that can be further studied for association with human phenotypes.



SNP 1	SNP 2	---	SNP M
AA	CT	---	AG
AG	CC	---	GG
GG	CT	---	AA
---	---	---	---

Gene 1	Gene 2	---	Gene N
5.3	3.2	---	4.1
2.8	10.1	---	1.6
0.7	1.2	---	0.3
---	---	---	---

Gene	Tissue
Gene 1	Blood
Gene 2	Lung
Gene 3	Brain
...	...

Poster #2637 - Making Your Points in Under a Minute: Utilization of Natural Language Processing to Produce Concise Bilingual Summaries of Common Orthopaedic Conditions

Richard Samade, M.D., Ph.D. [1,2]; Robert C. Weinschenk, M.D. [1,2]

[1]Department of Orthopaedic Surgery, UT Southwestern Medical Center, Dallas, TX; [2]Department of Biomedical Engineering, UT Southwestern

Introduction: Cognitive research has shown that sustained optimal attention spans in adults is approximately 1 minute in duration. Natural language processing (NLP) models that utilize transformer neural network architecture are currently available that summarize and translate text. Our question focused on whether NLP models could produce concise and accurate summaries of orthopaedic conditions for Spanish-speaking patients.

Methods: English language patient handouts were selected (https://orthoinfo.aaos.org/en/diseases--conditions) by the authors (both practicing faculty orthopaedic surgeons). A custom Python script was written and executed on a MacBook® laptop with 8 GB of RAM and CPU processor. Handout text was summarized with the BART Large NLP Model (Meta Platforms). The FLAN T5 Large NLP Model (Alphabet) was used to translate the English summary into Spanish. The authors reviewed the texts of the handouts and summaries to ensure that appropriate information was given. To standardize the spoken length, audio files of the summaries were generated with the MMS: English Text-to-Speech Model (Meta Platforms).

Results: The NLP models overall produced meaningful and appropriate content summaries for the 21 conditions evaluated. Handouts had word lengths of 2171 ± 706 words. Subsequent English and Spanish summaries had word lengths of 50 ± 8 words and 53 ± 7 words, respectively. Assuming a seamless consecutive dialogue of the English and then Spanish summaries, 95% of all summaries could be spoken in 53 seconds.

Conclusion: Readily available NLP models can produce accurate and concise bilingual summaries of orthopaedic conditions. Future research will prospectively assess surgeon and patient feedback on these NLP models.

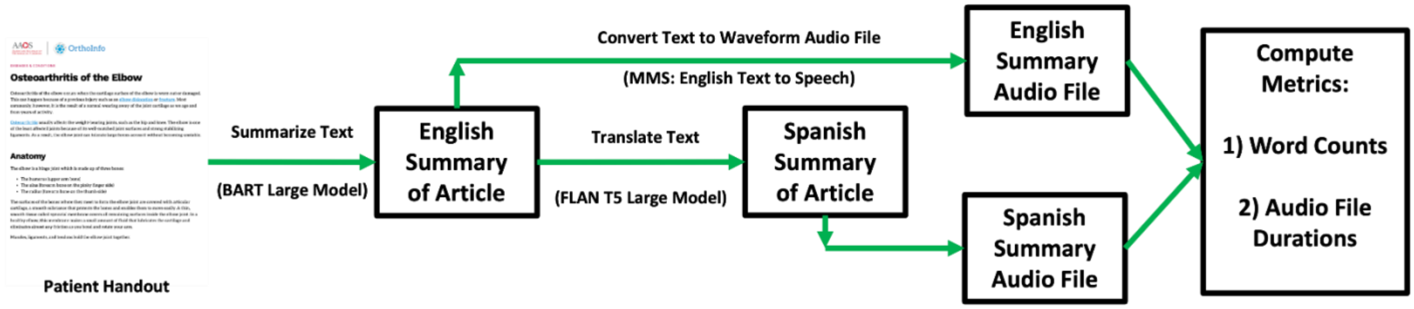

AAOS
OrthoInfo
Osteoarthritis of the Elbow
Anatomy
Patient Handout
Summarize Text
(BART Large Model)
English Summary of Article
Convert Text to Waveform Audio File
(MMS: English Text to Speech)
Translate Text
(FLAN T5 Large Model)
Spanish Summary of Article
English Summary Audio File
Spanish Summary Audio File
Compute Metrics:
1) Word Counts
2) Audio File Durations

Figure 1: *Workflow for text processing and computing metrics.*

Poster #2640 - Identification and Stratification of Myositis Subtypes Using Machine Learning

Peifeng Ruan[1], Amit Rupani[2], Salman Bhai[3,4]

[1]Department of Public Health, UT Southwestern Medical Center; [2]Cleveland Clinic Foundation; [3]Department of Neurology, UT Southwestern; [4]Neuromuscular Center, Institute for Exercise and Environmental Medicine

Background: Myositis is an inflammatory disorder aRecting muscle with multiorgan involvement and increased cancer risk. Though myositis subtypes are recognized, including dermatomyositis (DM), immune mediated necrotizing myopathy (IMNM), inclusion body myositis (IBM), clinical presentations are heterogeneous, treatments are non-specific, and pathophysiologic mechanisms are unclear. We clustered myositis subtypes by molecular characteristics to better stratify patients and identify potential therapeutic targets.

Methods: We obtained and merged raw bulk transcriptomic data from the NIH Gene Expression Omnibus (GSE: 220915, 143323, 102138) and performed molecular clustering. We created a similarity network and denoised the data through a network enhancement algorithm. We then implemented an enhanced spectral clustering method to identify molecular subtypes. Biological characteristics of the molecular subtypes were evaluated via linear regression diRerential expression and canonical pathways.

Results: We demonstrated 7 distinct subtypes across 157 myositis samples. Subtypes 1-4 were primarily DM. Subtype 5 was dominated by IMNM with a specific clinical antibody subtype (HMGCR). Subtype 6 was a mix of DM and IMNM, and Subtype 7 was a mix of IBM and IMNM. We identified unique signaling in Subtype 6 compared to Subtype 7 related to upregulation of Myc targets and oxidative phosphorylation (OXPHOS). Subtype 5 showed unique signaling in JAK/STAT pathways relative to Subtype 6.

Discussion: Given the rarity of myositis, we demonstrated feasibility of combining datasets and using machine learning to produce clinically relevant subtypes with unique molecular features. Myc target upregulation is associated with cancer risk and may help stratify patients for cancer screening. OXPHOS pathway upregulation in DM may suggest mitochondrial dysfunction and portend a worse prognosis. Identification of JAK/STAT signaling in a particular subtype of IMNM opens a potential novel targeted therapeutic option.

Poster #2641 - Exploring Pre-Trained Language Models for Vocabulary Alignment in the UMLS

Xubing Hao[1], Rashmie Abeysinghe[2], Jay Shi[3], Licong Cui[1,*]

[1]McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX; [2]Department of Neurology, McGovern Medical School, The University of Texas Health Science Center at Houston; [3]Intermountain Healthcare, Denver, CO
*Corresponding author

Objective: The Unified Medical Language System (UMLS) Metathesaurus is a biomedical vocabulary integration system created by the US National Library of Medicine. Given that the UMLS Metathesaurus incorporates millions of terms from more than 180 source vocabularies, integrating and aligning these terms is challenging. This study aims to investigate the efficacy of Pre-trained Language Models (PLMs) for vocabulary alignment in the UMLS Metathesaurus.

Methods: We frame the research question as two Natural Language Processing (NLP) tasks: Text Classification and Text Generation. We fine-tune four open-source cutting-edge PLMs including BERT and RoBERTa, GPT-2, and BLOOM. We devise two distinct input/prompt configurations for each task setting, with one comprising solely the names of the two terms while the other provides additional information about their source vocabularies and parent terms. Our evaluation metrics include precision, recall, and F1 score. Furthermore, we delve into the proficiency of PLMs in detecting missing synonymous terms within the UMLS Metathesaurus. We extract a randomly chosen subset of suggested potentially missing synonymous terms for manual evaluation by a domain expert with experience in clinical terminology assessment.

Results: Experiments show that the best model is RoBERTa achieving a precision, recall, and F1 score of 0.965, 0.940, and 0.952 respectively. Incorporation of contextual information in the inputs improves the model performance in the Text Classification task, albeit with a limited impact on the Text Generation task. In addition, domain expert evaluation of 100 randomly selected suggested potentially missing synonymous terms generated by the best model revealed that 78 of them as valid synonymous terms.

Conclusions: In this study, we evaluated the efficacy of PLMs including BERT, RoBERTa, GPT-2, and BLOOM for UMLS vocabulary alignment. Results indicated the promise of PLMs in facilitating vocabulary alignment in the UMLS Metathesaurus.

Poster #2642 - Prior Frequency-Guided Diffusion Model for Limited-Angle (LA) CBCT Reconstruction (PFGDM)

Jiacheng Xie, Hua-Chieh Shao, Yunxiang Li, You Zhang

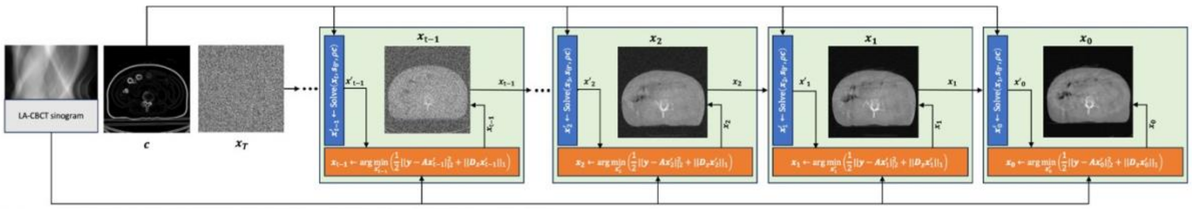
Medical Artificial Intelligence and Automation (MAIA) Laboratory, Department of Radiation Oncology, UT Southwestern Medical Center, Dallas, TX

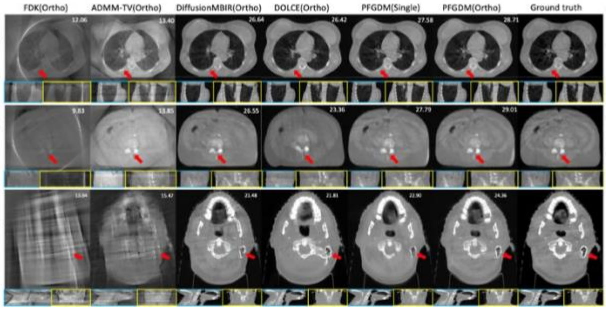
Purpose: Reconstructing CBCTs from limited-angle acquisitions (LA-CBCT) is highly desired for improved imaging efficiency, dose reduction, and better mechanical clearance. We developed a diffusion model-based framework, prior frequency-guided diffusion model (PFGDM), for robust and structure- preserving LA-CBCT reconstruction.

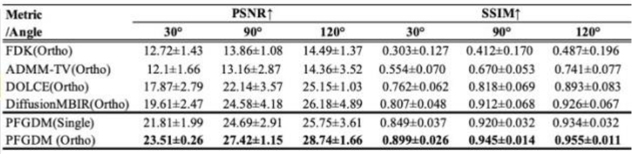
Methods: Diffusion models can generate data/images by reversing a data-noising process through learned data distributions; and can be incorporated as a denoiser/regularizer in LA-CBCT reconstruction. In PFGDM, we introduced patient-specific prior CT information further to enhance the regularization potential of population-based diffusion models. Specifically, we developed a flexible conditioning scheme for PFGDM, by conditioning the reverse diffusion process with prior high-frequency CT information, to preserve anatomical details in LA-CBCTs. The controllable scheme allows the condition to be flexibly dropped during the reconstruction, enabling both similar and differing CT/CBCT anatomies to be reconstructed. The diffusion model of PFGDM was trained on 3000 CBCT slices from 30 patients of three anatomical sites: head-and-neck, lung, and pelvis. The high-frequency information from training CBCTs was extracted by a Sobel filter as prior information and randomly introduced into the reverse diffusion process to train an anatomy-preserving, conditioning-enabled model. During test-time LA-CBCT reconstruction, we incorporated the diffusion model into an alternating-direction-method-of-multipliers (ADMM) framework as a regularizer and used Sobel-filtered prior CT information to introduce prior knowledge and condition the diffusion model. FPGDM was tested on a separate set of 12 patients (4 for each anatomical site), and evaluated via metrics including peak-signal-to-noise-ratio (PSNR) and structure-similarity-index-measure (SSIM).

Results: PFGDM outperformed all traditional and diffusion model-based methods. As shown in Fig. 1, PFGDM exhibits superior performance across all metrics.

Conclusion: PFGDM reconstructs high-quality LA-CBCTs under very-limited gantry angles, allowing faster and more flexible CBCT scans with dose reductions.


LA-CBCT sinogram
c
x_T
x_{t-1}
x_2
x_1
x_0
x'_{t-1}
x'_2
x'_1
x'_0
$x'_{t-1} \leftarrow \text{Solve}(x_t, \theta_\theta, pc)$
$x_{t-1} \leftarrow \arg\min_{x'_{t-1}} \left(\frac{1}{2}\|y - Ax'_{t-1}\|_2^2 + \|D_z x'_{t-1}\|_1\right)$
$x_2 \leftarrow \arg\min_{x'_2} \left(\frac{1}{2}\|y - Ax'_2\|_2^2 + \|D_z x'_2\|_1\right)$
$x_1 \leftarrow \arg\min_{x'_1} \left(\frac{1}{2}\|y - Ax'_1\|_2^2 + \|D_z x'_1\|_1\right)$
$x_0 \leftarrow \arg\min_{x'_0} \left(\frac{1}{2}\|y - Ax'_0\|_2^2 + \|D_z x'_0\|_1\right)$





Metric /Angle	PSNR↑ 30°	PSNR↑ 90°	PSNR↑ 120°	SSIM↑ 30°	SSIM↑ 90°	SSIM↑ 120°
FDK(Ortho)	12.72±1.43	13.86±1.08	14.49±1.37	0.303±0.127	0.412±0.170	0.487±0.196
ADMM-TV(Ortho)	12.1±1.66	13.16±2.87	14.36±3.52	0.554±0.070	0.670±0.053	0.741±0.077
DOLCE(Ortho)	17.87±2.79	22.14±3.57	25.15±1.03	0.762±0.062	0.818±0.069	0.893±0.083
DiffusionMBIR(Ortho)	19.61±2.47	24.58±4.18	26.18±4.89	0.807±0.048	0.912±0.068	0.926±0.067
PFGDM(Single)	21.81±1.99	24.69±2.91	25.75±3.61	0.849±0.037	0.920±0.032	0.934±0.032
PFGDM (Ortho)	**23.51±0.26**	**27.42±1.15**	**28.74±1.66**	**0.899±0.026**	**0.945±0.014**	**0.955±0.011**

Fig. 1. PFGDM framework and experiment results. The limited-angle projections were simulated with either orthogonal-view (ortho) acquisitions or single-view (single) acquisitions. With the high-frequency prior CT condition and the flexible condition-dropping mechanism, PFGDM reconstructs the anatomical structures to match well with the 'ground-truth', even for the 30-degree scan angle scenario. Quantitatively, PFGDM consistently outperforms other methods under all limited-angle scenarios.

Poster #2643 - AI Prognostics in Trauma: Predicting Outcomes via an Automated, Real-Time Machine Learning Algorithm

K. Gopal, M.Ed.; K. Diercks; M. Cheng; A. Bain; C. Hirschkorn; V. Chowdhry; D. Sanders; A. Starr; C. Park, M.D., M.P.H.

Department of Surgery, UT Southwestern Medical Center

Objective: To explore PTIM predicative capabilities on patient outcomes and resource utilization

Summary Background Data: Machine learning algorithms are abundant, but most remain retrospective and external to the electronic medical record, with less real-time implementation. The Parkland Trauma Index of Mortality ("PTIM") is a validated tool that can help predict 48-hour mortality on an hourly basis. Prior work with this novel automated machine learning ("ML") algorithm, PTIM includes perceived utility in communication, resuscitation, and predicting mortality. We hypothesized that PTIM score may be associated with resource utilization and clinical outcomes.

Methods: A retrospective single-center study was conducted between December 2020 to August 2021 at a Level 1 Trauma center and included patients who met criteria for level 1 and level 2 activations. The effects of PTIM score on outcome variables were analyzed using simple linear and logistic regression.

Results: There were a total of 171 patients, with a mean PTIM score of .26 [IQR=.37]. Blunt trauma was more common than penetrating. Higher PTIM max score was not associated with decreased time to OR, while ICU Days significantly increased by 9.66 days with one unit of PTIM Max score increase (P<0.001). The number of days on a ventilator and days in the hospital also significantly increased with a one unit increase in PTIM max score, by 7.24 days (p<0.001) and 16.68 days (p<0.001) respectively (Figure 2a). The amount of blood transfused was found to have increased by 3.70 units with a 1 unit increase in PTIM score (p<0.002) (Figure 2b). The odds of shock, pneumonia, and mortality were also found to have increased with an increase in PTIM max score (p=0.022, p<0.001, p<0.003).

Conclusion: PTIM is a validated machine learning algorithm that can help predict risk of mortality in real-time and anticipate clinically significant outcomes and utilization such as transfusion, ICU days and ventilator days. Next steps include external validation at other sites to increase generalizability and explore other uses of this real-time mortality prediction model.

Poster #2645 - AI-Assisted Patient Query Resolution for Enhanced Healthcare Efficiency and Personalized Care

Safia Khan, M.D.

Department of Family and Community Medicine, UT Southwestern Medical Center

Abstract: The increasing demand for efficient and personalized healthcare has necessitated the exploration of innovative solutions. This abstract presents an AI-assisted system designed to streamline patient query resolution within the MyChart platform, significantly reducing the workload of nurses and doctors while ensuring patient data privacy and HIPAA compliance.

Our proposed solution leverages an on-premises Language Model (LLM) that securely accesses patient files, maintaining the highest standards of data protection. Utilizing advanced natural language processing techniques, the AI system analyzes patient queries and their corresponding health histories to provide tailored and context-aware responses.

The AI system employs a Retrieval-Augmented Generation (RAG) approach, enabling it to access and extract relevant information from trusted and university-approved resources, such as UpToDate by Wolters Kluwer. By combining patient-specific data with evidence-based medical knowledge, the AI generates comprehensive and personalized answers to patient queries.

To ensure the accuracy and appropriateness of the AI-generated responses, a review and approval process is implemented. Healthcare providers receive the proposed answers for evaluation, allowing them to make necessary modifications or provide additional insights. Once approved, the responses are seamlessly delivered to the patients via MyChart, accompanied by any supplementary actions or recommendations deemed necessary.

This AI-assisted solution offers numerous benefits, including:

1. Increased efficiency: By automating the initial stages of patient query resolution, healthcare providers can focus on more complex and critical tasks, optimizing their time and resources, while reducing documentation burnout.

2. Personalized care: The AI system considers each patient's unique health history, ensuring that the provided answers are tailored to their specific needs and circumstances.

3. Evidence-based responses: Leveraging trusted medical resources, the AI generates responses grounded in the latest scientific evidence, promoting high-quality care.

4. Enhanced patient satisfaction: Timely and personalized responses to patient queries improve communication and foster a stronger patient-provider relationship.

5. Scalability: The AI system can handle a large volume of patient queries simultaneously, accommodating the growing demands of healthcare organizations.

By integrating AI technology with the MyChart platform, this solution revolutionizes patient query management, ultimately improving healthcare efficiency, personalization, and patient satisfaction. The proposed system represents a significant step forward in the application of AI in health system operations, showcasing the potential for transformative change in the healthcare industry.

Poster #2646 - Improving the Readability of Patient Education Materials on Traumatic Injuries Using ChatGPT

Bahaa Succar[1], Andrew Bain[1], Kaustubh Gopal[2], Justin Kosley[2], Linda Dultz[1], Caroline Park[1], Ryan P. Dumas[1]

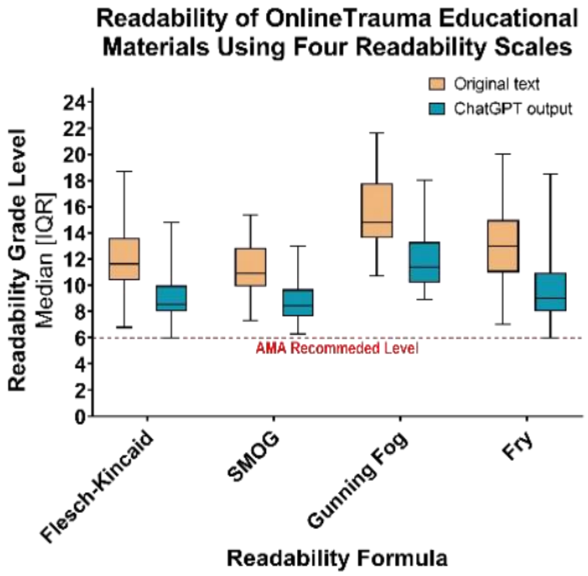
[1]Department of Trauma, Burn, and Acute Care Surgery, UT Southwestern Medical Center; [2]UT Southwestern Medical School

Introduction: Online health information has become a crucial resource for patients. Research indicates that online trauma surgery materials remain significantly more complex than the American Medical Association (AMA) recommended reading level. The emergence of large language models such as ChatGPT might provide a solution to this issue. We hypothesized that current reading grades of popular public-facing websites remain higher than AMA recommendations, and that ChatGPT could convert these materials on traumatic injuries into accessible texts with improved readability levels.

Methods: Text from all patient education webpages across national trauma organizations' websites was collected. Each webpage provides information regarding common types of traumatic injuries and measures for injury prevention and control. All text except complete sentences was excluded. ChatGPT 3.5 was tasked to improve the readability of the original texts. Four readability measures (SMOG, Gunning Fog, Flesch-Kincaid, and Fry) were used to assess the readability level required to fully comprehend both the original texts and ChatGPT outputs. Mean and median readability scores were calculated for each measure across webpages to compare the ChatGPT outputs with the original texts.

Results: A total of 51 webpages underwent readability analysis. The mean readability score of original texts across the four measures was 13, corresponding to a grade-level of a college freshman. Zero out of 51 webpages adhered to the AMA recommendation of presenting patient education materials at a grade 6 reading level. Upon rewriting the webpage content, ChatGPT significantly improved the mean readability across all scoring measures to a grade level of 9 ($p<0.001$), the equivalent to a high school freshman. Moreover, after ChatGPT revision, 4 out of 51 webpages (8%) met the AMA recommended grade level on at least one of the readability scales.

Conclusion: ChatGPT can help reduce the reading level observed in trauma-related patient education materials. Future endeavors should further explore the capacity of large language models to improve the readability levels of online patient-targeted health information. This is essential to ensure equitable access to these resources, particularly among marginalized groups with low health readability.


Readability of OnlineTrauma Educational Materials Using Four Readability Scales
Original text
ChatGPT output
Readability Grade Level
Median [IQR]
24
22
20
18
16
14
12
10
8
6
4
2
0
AMA Recommeded Level
Flesch-Kincaid
SMOG
Gunning Fog
Fry
Readability Formula

Poster #2647 - Ethical Considerations for the Use of Artificial Intelligence in Mental Health Care

Matthew Hutnyan, B.S.

Division of Psychology, Department of Psychiatry, UT Southwestern Medical Center

Abstract: The proliferation of artificial intelligence (AI) technologies and their growing capabilities has inspired hope and fear across many industries and professions, including among those who provide mental health care (MHC). AI technologies are increasingly leveraged in the delivery of health care services traditionally delivered solely by humans and, although AI is not widely integrated into the everyday delivery of mental health services, many applications are developing. Importantly, AI is poorly understood by many, and faces ethical concerns among mental health professionals, making the use of it in the delivery of care controversial. It is crucial that MHC professionals understand the current and future uses of AI systems, be prepared to effectively implement them, and engage in thoughtful discourse regarding the ethical and responsible development, implementation, and regulation of AI. While much has been written about emerging AI technologies and the use of AI in practice, relatively little text has been dedicated primarily to the discussion of ethical considerations. To fill this gap, a brief overview of current and potential future uses of AI in the delivery of mental health services and detailed discussion of ethical considerations associated with the adoption of AI is provided. An ethical analysis of AI in MHC was conducted using the framework of the American Psychological Association’s Code of Ethics (2017) and other biomedical ethics literature (e.g., Beauchamp and Childress’ principles). Five primary domains of ethical challenges emerged: harm and nonmaleficence, autonomy and informed consent, fidelity and responsibility, privacy and confidentiality, and bias and justice. Specifically, concerns related to the capabilities of machines (e.g., ability to perform basic functions, potential for malfunction), freedom of choice and the ability to define the value proposition for oneself, trust and maintenance of the therapeutic relationship, data privacy and limits to confidentiality, and access to and quality of care were identified. Based on these challenges, a set of recommendations is provided, including, but not limited to, increasing opportunities for education about AI for MHC professionals and trainees and developing guidelines for the use of AI in MHC.

Poster #2648 - Harnessing the AI Revolution: Application of Large Language Models (LLMs) for Genotype Imputation

Devansh Pandey[1], Vagheesh M. Narasimhan[1, 2]

[1]Department of Integrative Biology, The University of Texas at Austin; [2]Department of Statistics and Data Science, The University of Texas at Austin

Abstract: Genotype imputation involves predicting missing genetic variants using data from observed variants from large reference panels like the UK Biobank (UKB). Traditional methods, such as those based on Hidden Markov Models (HMMs), provide a strong statistical approach for deducing unobserved haplotypes. However, these methods face challenges including slow imputation speeds when dealing with genome-wide data across multiple individuals and the limited diversity in the reference panels that train these HMMs.

Large Language Models (LLMs), which are increasingly used in the natural language processing domain for tasks such as machine translation, sentiment analysis, and text generation, are now being considered for applications in genomics. These models have proven to be effective in handling various tasks involving sequential data. By adopting LLMs for genotype imputation, we aim to overcome the limitations of HMMs and enhance the efficiency and inclusivity of population genetic analysis.

In this study, we utilized haplotype sequence data from the UKB, focusing on chromosome 20, to pretrain a Generatively Pretrained Transformer (GPT-2) model. We started by extracting genotype data and splitting it into two equal-length haplotype strings that span the entire chromosome. These strings were then divided into sections of 101 SNPs each, matching the average linkage disequilibrium block length in the data. The GPT-2 model was trained to predict the 101st SNP from the first 100 SNPs in each section, masking the target SNP during training. Through this pre-training, the model learned complex patterns such as allele frequency distributions and haplotype structures. It achieved a 94.5% accuracy rate in predicting masked SNPs on a validation dataset. Following this successful pre-training phase, we aim to fine-tune the model for genotype imputation using varied reference panels representing diverse ancestries, enhancing the model's ability to interpret genetic variations across different populations.

This study demonstrates the potential of LLMs in population genetics, suggesting a path forward for developing models tailored to genomic data rather than language, thereby expanding the applicability of LLM methodologies in population genetics research.

Poster #2650 - LLM-Powered Summarization and Analysis of Patients Transcripts

Terence Lim, Ph.D.[1]; Carlos Mery, M.D., M.P.H.[2]; Andrew Well, M.D., M.P.H.[2,3]; Ying Ding, Ph.D.[2,4]

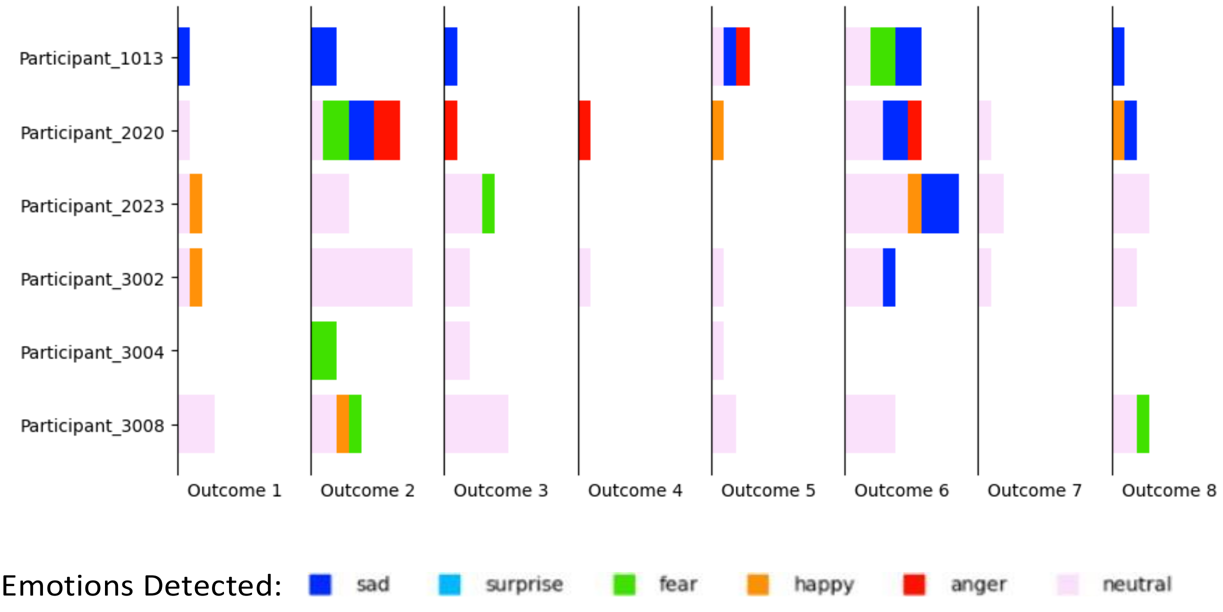
[1]College of Natural Sciences, The University of Texas at Austin; [2]Dell Medical School, The University of Texas at Austin; [3]McCombs School of Business, The University of Texas at Austin; [4]School of Information, The University of Texas at Austin

Abstract: With the increasing number of surgeries in the United States coupled with shorter hospital stays, patients and their families now bear the responsibility of managing their own recovery. This presents a critical challenge for healthcare professionals, who must comprehend the nuances of postoperative recovery, including identifying medical delivery gaps and assessing patient outcomes. Semi-structured interviews, conducted individually or in focus groups, are commonly used in the post-operative recovery process to gather descriptive data. However, the manual transcription and analysis of these lengthy interviews can consume hundreds of hours for health professionals.

We introduce a novel solution leveraging large language model (LLM) technology to analyze interview transcripts, and uncover challenges and emotions expressed by patients in their recovery journey. We also incorporate NLP methods for gathering and attributing sentences based on their semantic meanings, which we apply to reducing lengthy inputs and linking to evidence supporting the topics identified. By harnessing the cognitive reasoning functions of LLMs, we may eliminate manual coding and reduce analysis time, response delay, professional burnout, and clinical costs.

From a pilot study involving interview transcripts from four focus groups on post-pediatric open-heart surgeries, our methodology extracted meaningful outcomes, and detected the emotions and engagement level of participants across outcome topics. We aim to systematically assess these automated outputs against health professionals' results, to meet the relevant quality requirements.

Figure 1. Participants Engagement and Emotion



Outcome Topics:

1. Clear And Accurate Information For Diagnosis And Treatment

2. Understanding And Coping With A Medical Condition
3. Continuing With Normal Activities Despite A Medical Condition
4. Better Healthcare Quality And Insurance Coverage
5. Gaining Knowledge And Perspective Through Research And Consulting
6. Instilling Healthy Habits In Child And Preparing For Parenthood
7. Making Independent Decisions And Handling Sensitive Information
8. Managing Anxiety And Finding Support

Poster #2651 - Machine Learning for Early Prediction of Electrographic Seizures in Newborns

Srinivas Kota[1], Yaser Elnakeib[2], Lina Chalak[1]

[1]Department of Pediatrics, UT Southwestern Medical Center, Dallas, TX; [2]Clinical Informatics Center, UT Southwestern

Introduction: Seizures occur in 1.8 to 3.5 per 1,000 live births, with hypoxic ischemic encephalopathy (HIE) being the primary cause in newborns. Higher burden of electrographic seizures is linked to poorer outcomes like death or disability in neonates. Early seizure prediction and prompt intervention could therefore significantly improve these outcomes.

Methods: This study utilized a publicly available dataset of neonatal (n = 79) scalp electroencephalogram (EEG) recordings with continuous seizure annotations by three experts (Stevenson et al. 2019). A fourth-order, zero-phase Butterworth bandpass filter was applied to EEG data from 18 bipolar electrodes. These filtered signals were then used to calculate features for classification, including absolute and relative spectral power for various frequency bands (slow delta, fast delta, delta, theta, alpha, beta, and total power), as well as spectral edge frequency. Features affected by outliers (Webb et al. 2021) were imputed using a median filter based on the previous 10 minutes of data. EEG features were segmented into 15-minute intervals, with each segment analyzed to predict the likelihood of a seizure occurring in the following three minutes. A two-layer LSTM network (90% training, 10% testing) with dropout layers to prevent overfitting with a sigmoid activation function was used for seizure prediction. Training utilized a binary cross- entropy loss and balanced accuracy to address class imbalance.

Results: The model's architecture was validated by monitoring balanced accuracy and loss over 20 epochs. Performance metrics on the test subset achieved an accuracy of 88.69%, sensitivity of 72.55%, specificity of 95.73%, ROC AUC of 93.67%, and balanced accuracy of 84.14%.

Conclusion: Our results highlight the model's ability to discriminate between seizures and non-seizure events, establishing a robust framework for electrographic seizure prediction. Early and accurate seizure prediction would enable prompt intervention, minimizing seizure burden and potentially improving both short-term and long-term outcomes for newborns with seizures.

Poster #2653 - Site-Agnostic 3D Dose Prediction Through Deep Learning for Brain, H&N, and Prostate Cancer Patients

Hui-Ju Wang, Billing Wang, Junjie Wu, Austen Maniscalco, David Sher, Mu-Han Lin, Steve Jiang, Dan Nguyen

Department of Radiation Oncology, UT Southwestern Medical Center, Dallas, TX

Objective: Artificial intelligence has made significant progress in radiation therapy treatment planning, particularly in dose predictors for various types of cancer. However, current site-specific models face limitations in predicting doses for multiple treatment sites, influenced by factors like organ structures, tumor complexity, and multi-constraint planning objectives. To address this challenge, our team proposes a site-agnostic deep learning-based model incorporating dosimetric goals and utilizing various treatment site data. We aim to enhance the model's adaptability and generalizability, ultimately improving overall planning efficiency.

Methods: This study examines disease datasets treated with VMAT, including 111 cases of prostate cancer, 142 cases of head and neck (H&N) cancer, and 104 cases of brain cancer, which are divided into training (70%), validation (10%), and testing (20%) for model development and evaluation. Our proposed U-net-based model outputs the dose prediction distribution by integrating the input of computed tomography scans, planning target volume, the organ at risk(OAR), and a dosimetric goal array that consolidates spatial data from multiple OAR arrays with dosimetric context through treatment planning objectives. We also trained a baseline model from scratch only with a single-site disease.

Results: The proposed model can predict dose for multiple disease sites (Fig. 1) and shows improvement compared to the single-site model, demonstrating a reduction in mean absolute percent error for patients' brain cases (3.12 vs. 3.74%), prostate cases (1.44 vs. 1.66%), and H&N cases (3.22 vs. 3.14%).

Conclusion: The proposed site-agnostic model can be adapted to different treatment sites efficiently while maintaining clinically minimal differences between the clinical and predicted plans.

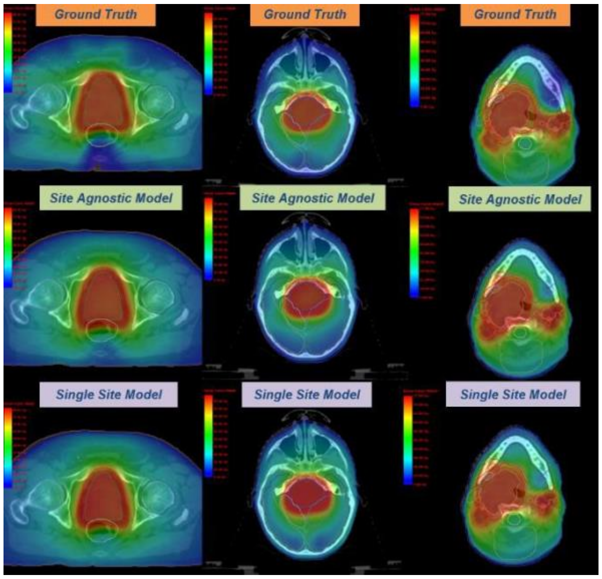


Fig 1. Dose wash for testing patients in three disease cases with ground truth, site-agnostic model, and single-site model

Poster #2654 - ChatGPT and Large Language Models (LLMs) Awareness and Use. A Prospective Cross-Sectional Survey of Medical Students

Conner Ganjavi[1,2,3], Michael Eppler[1,2,3], Devon O'Brien[3], Lorenzo Storino Ramacciotti[1,2], Muhammad Shabbeer Ghauri[4], Issac Anderson[5], Jae Choi[6], Darby Dwyer[7], Claudia Stephens[8], Victoria Shi[9], Madeline Ebert[10], Michaela Derby[11], Bayan Yazdi[12], Andre Abreu[1,2,3], Giovanni E. Cacciamani[1,2,3]

[1]USC Institute of Urology and Catherine and Joseph Aresty Department of Urology, Keck School of Medicine, University of Southern California, Los Angeles, CA; [2]AI Center at USC Urology, USC Institute of Urology, University of Southern California; [3]Keck School of Medicine, University of Southern California; [4]California University of Science and Medicine, Colton, CA; [5]Wayne State University School of Medicine, Detroit, MI; [6]UT Southwestern Medical School, Dallas, TX; [7]Texas A&M School of Medicine, Bryan, TX; [8]Frederick P. Whiddon College of Medicine, University of South Alabama, Mobile, AL; [9]University of Missouri-Kansas City School of Medicine, Kansas City, MO; [10]Medical College of Wisconsin, Milwaukee, WI; [11]Sanford School of Medicine, University of South Dakota, Vermillion, SD; [12]Stritch School of Medicine, Loyola University Chicago, Maywood, IL

Background/ Purpose/ Goal/ Hypothesis: Generative-AI (GAI) models like ChatGPT are becoming widely discussed and utilized tools in medical education. It can be used to assist with studying for exams, shown capable of passing the USMLE. However, there have been concerns expressed regarding fair and ethical use. To evaluate the views of medical students, we conducted a nationwide survey focused on students' current views of and utilization of GAI technology.

Methods: The survey was available for 14 days and was distributed via social media and via student collaborators who serve as members of the Organization of Student Representatives (OSR), the student delegation to the AAMC. Narrative and chi-squared analysis were used in this study.

Results: Overall, 415 students from 28 medical schools completed the survey. The vast majority (96%) of respondents had heard of ChatGPT and 52% had used it for medical school coursework. Among those who used ChatGPT, the most common use in pre-clerkship and clerkship phase was asking for explanations of medical concepts (100%) and assisting with diagnosis/treatment plans (73.7%), respectively. The most common use in academic research was for proof reading and grammar edits (59.6%). Respondents recognized the potential limitations of ChatGPT, including inaccurate responses (53.4%), patient privacy (87.4%), and plagiarism (81.9%). Students recognized the importance of regulations to ensure proper use of this novel technology (91.7%)

Conclusions: This nationwide survey of US medical students demonstrates that learners are already integrating GAI technology like ChatGPT into their study resources and support tools. Further, students recognize the limitations of ChatGPT and the ethical concerns related to its unregulated use. Understanding the views of students is essential to crafting workable instructional courses, guidelines, and regulations that ensure the safe, productive use of generative-AI in medical school.

Poster #2656 - Find Physical Exam Period in OSCEs

Shinyoung Kang, B.S.[1]; Michael J. Holcomb, M.S.[1]; Sol Vedovato, M.S.[1]; David Hein, M.S.[1]; Ameer H. Shakur, Ph.D.[1]; Thomas O. Dalton, M.D.[2]; Krystle K. Campbell, D.H.A.[3]; Gaudenz Danuser, Ph.D.[1]; Daniel J. Scott, M.D.[3,4]; Andrew R. Jamieson, Ph.D.[1]

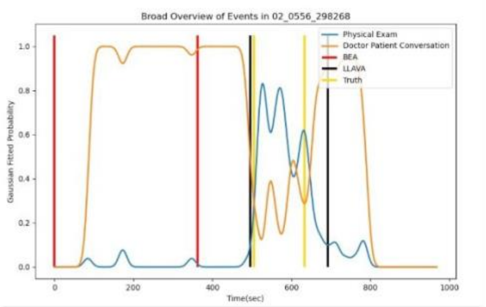
[1]Lyda Hill Department of Bioinformatics, UT Southwestern Medical Center; [2]Department of Internal Medicine, UT Southwestern; [3]Simulation Center, UT Southwestern; [4]Department of Surgery, UT Southwestern

Abstract: Medical schools use Objective Structured Clinical Exams (OSCEs) to assess students' clinical skills. OSCEs are standardized simulation-based exams that test history-gathering, communication, physical examination, procedures, documentation, and clinical reasoning skills to mirror real patient encounters in clinical settings. The assessment of these encounters can be summarized mainly as obtaining medical information, performing physical exams, and providing a summarized diagnosis to the patient. We present a way to find the physical exam period during OSCEs by using multimodal models to obtain descriptions of the videos and using large language models to make decisions.

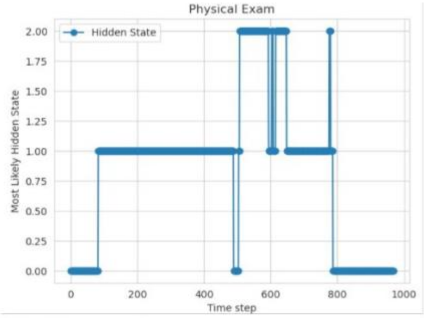
In phase one, we build a pipeline for reading video on one frame per second and preprocess individual frames using You Only Look Once version 8 (YOLOv8) human pose detector to find the number of people present in the frame and the bounding box of the people to 672 x 672 pixels. If two people are detected in the frame, we obtain patient-doctor interaction using a Large Language-and-Vision Assistant (LLAVA).

In phase two, we use GPT3.5 Generative Pre-trained Transformer 3.5 to match the patient-doctor interaction description to either "No doctor present," "Doctor is having a conversation with the patient/taking notes," or "Doctor is performing a physical exam."

In phase three, we use the first twenty videos of each station to create a transition and emission matrix for Viterbi decoding. Using Viterbi decoding, we obtain the physical exam period within the video and calculate the recall, precision, and Intersection of Union from the hand-labeled physical exam time.



BEA: Distance Method (Doctor to patient distance)
LLAVA: PD Rolling Method (How many times did the model think it was a physical exam over 10 second period
Truth: Hand labeled data, watching through the video



State 2: Physical exam
State 1: Conversation with the patient
State 0: No doctor is present in the frame

Poster #2657 - RAG vs. Zero-Shot? Leveraging LLMs for Transcript-Based Automated Grading of Medical Student Exams

Ameer H. Shakur, Ph.D.[1]; Michael J. Holcomb, M.S.[1]; Sol Vedovato, M.S.[1]; David Hein, M.S.[1]; Shinyoung Kang, B.S.[1]; Thomas O. Dalton, M.D.[2]; Krystle K. Campbell, D.H.A.[3]; Gaudenz Danuser, Ph.D.[1]; Daniel J. Scott, M.D.[3.4]; Andrew R. Jamieson, Ph.D.[1]

[1]Lyda Hill Department of Bioinformatics, UT Southwestern Medical Center; [2]Department of Internal Medicine, UT Southwestern; [3]Simulation Center, UT Southwestern; [4]Department of Surgery, UT Southwestern

Abstract: The evaluation of medical students' exams is a critical component of their education and professional development. Traditional methods of assessment, such as manual grading of simulated examination videos, can be time-consuming and subject to human bias. Advances in automatic speech recognition (ASR) and natural language processing (NLP) have unlocked opportunities for a finer grained evaluation of these medical encounters. ASR systems are approaching human-like adeptness at accurately transcribing conversations. Meanwhile, large language models (LLMs) have shown impressive capabilities at tasks like sentiment analysis, text understanding, text summarization, reasoning etc. However, LLM's are not without their faults – particularly in terms of the high costs of operations, and tendency to hallucinate responses. Retrieval augmented generation (RAG) is one solution that has been effective at dealing with these drawbacks in real world applications.

In this work, we (1) develop an ASR system that accurately transcribes medical examination encounter videos (2) use state-of-the-art LLMs to evaluate the students' performance in these encounters and (3) investigate the impact of RAG techniques on improving the system's performance in terms of cost reduction and hallucination mitigation.

To ensure the effectiveness and accuracy of our model, we evaluate the level of agreement between our model and the judgement of trained human evaluators. In collaboration with UT Southwestern Simulation Center, student encounters with specialized patients as recorded during the Objective Structured Clinical Examinations (OSCE) are used to evaluate our system.

This study will contribute to the evolving role of AI in medical education and pave the way creating a more efficient, and equitable system for training future medical professionals, and enhance their clinical readiness towards meeting the needs of their patients.

Poster #2658 - Synthesizing High-Resolution IDH-Specific Brain Tumor MRIs Using Latent Diffusion Models

Nghi C.D. Truong[1], Chandan Ganesh Bangalore Yogananda[1], Benjamin C. Wagner[1], James M. Holcomb[1], Divya Reddy[1], Niloufar Saadat[1], Sadeem Lohdi[1], Kimmo J. Hatanpaa[2], Toral R. Patel[3], Baowei Fei[1,4], Matthew D. Lee[5], Rajan Jain[5], Richard J. Bruce[6], Marco C. Pinho[1], Ananth J. Madhuranthakam[1], Joseph A. Maldjian[1]

[1]Department of Radiology, UT Southwestern Medical Center; [2]Department of Pathology, UT Southwestern; [3]Department of Neurological Surgery, UT Southwestern; [4]Department of Bioengineering, UT at Dallas; [5]Department of Radiology, NYU Grossman School of Medicine; [6]Department of Radiology, University of Wisconsin-Madison

Purpose: Isocitrate dehydrogenase (IDH) mutations are pivotal prognostic indicators in gliomas. This study introduces a generative AI model designed to produce diverse and high-quality MRI images specific to IDH mutations.

Methods: We propose MedStable, a two-stage framework for generating high-resolution (HR), 3D multi-contrast brain tumor MRI images. MedStable employs two latent diffusion probabilistic models (LDPM) [1] operating in sequential phases. The first LDPM generates low-resolution (LR) multi-contrast 3D scans with tumor mask and IDH status as conditions. The second LDPM upscales the LR MRI samples to HR images. We trained MedStable using three datasets: The Cancer Genome Atlas dataset [2], the UCSF Preoperative Diffuse Glioma MRI dataset [3], and an internal dataset from UT Southwestern Medical Center. We generated 3000 synthetic multi-contrast 3D samples (1500 IDH-mutated, 1500 IDH-wildtype) using MedStable, which were then used to train an IDH prediction model. This model was subsequently tested on real data consisting of 224 mutated and 1,002 wild-type cases from the University of Pennsylvania glioblastoma dataset, the Erasmus Glioma Database [4], and two held-out internal datasets.

Results: Fig. 1 showcases two examples of HR synthetic images produced by MedStable. The conditional tumor mask (depicted by the red contour) remained consistent for generating both mutated (first row) and wildtype (second row) samples. Furthermore, the IDH prediction model, trained on synthetic images and tested on real data, exhibited remarkable performance metrics with 93.6% accuracy, 87.5% sensitivity, 94.9% specificity, and 0.92 AUC.

Conclusion: MedStable demonstrates the capability to generate highly realistic 3D multi-contrast MRI data with whole-tumor mask and IDH mutation status as conditions. This advancement holds significant promise for application in various image-based classification tasks, particularly in scenarios characterized by scarce or imbalanced data.

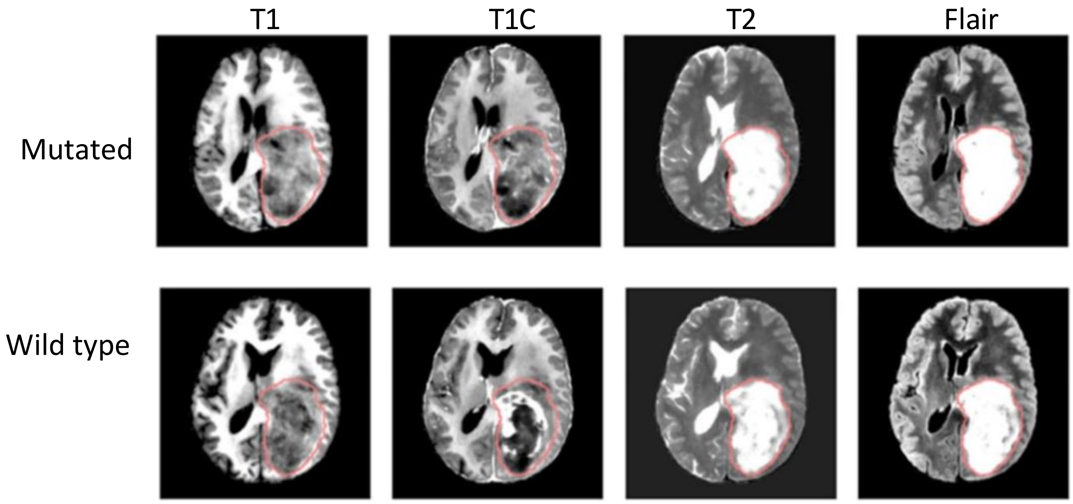


Fig 1. HR multi-contrast MRI generated by MedStable using the same tumor mask (red contour).
First row: IDH mutated example. Second row: IDH wild-type example.

References:

1. Rombach, R., et al., *High-Resolution Image Synthesis with Latent Diffusion Models.* 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022: p. 10674-10685.
2. Ceccarelli, M., et al., *Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma.* Cell, 2016. **164**(3): p. 550–563-550–563.
3. Calabrese, E., et al., *The University of California San Francisco Preoperative Diffuse Glioma MRI Dataset.* Radiology: Artificial Intelligence, 2022. **4**(6): p. e220058-e220058.
4. van der Voort, S.R., et al., *The Erasmus Glioma Database (EGD): Structural MRI Scans, WHO 2016 Subtypes, and Segmentations of 774 Patients with Glioma.* Data in Brief, 2021. **37**: p. 107191-107191.

Poster #2659 - Prediction of High-Sensitivity Cardiac Troponin T (hs-cTnT) Elevation Using Cardiac Substructure Dose in Lung Cancer Radiotherapy

Xinru Chen[1,2], Xiaodong Zhang[1,2], Ting Xu[1], Yan Chu[1], Yao Zhao[1], Radhe Mohan[1,2], Joshua Niedzielski[1,2], Sanjay Shete[1,2], Laurence Court[1,2], Zhongxing Liao[1], Jinzhong Yang[1,2]

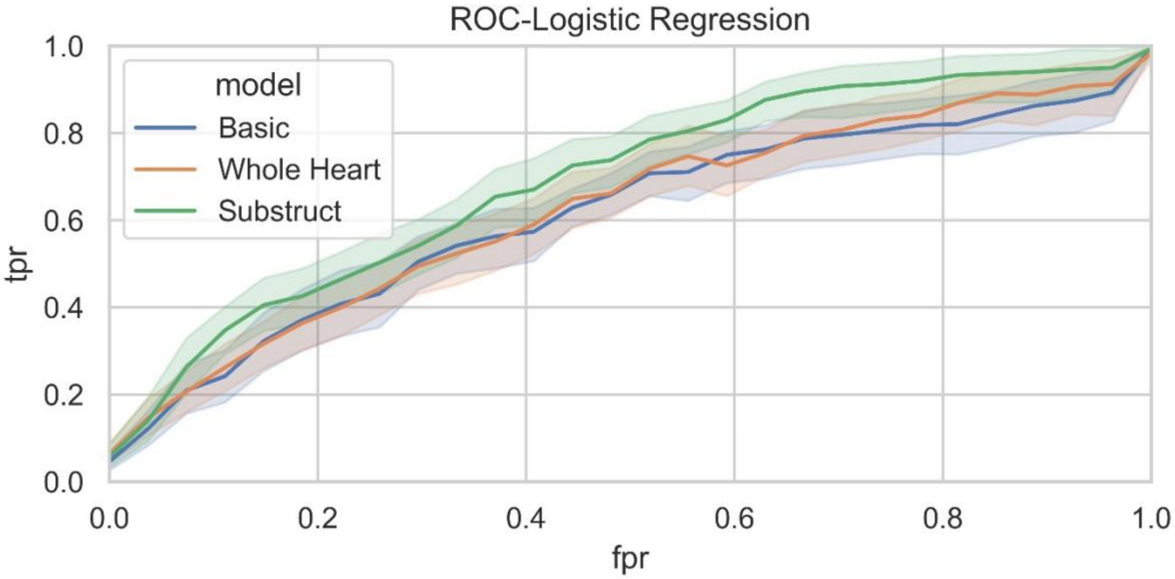
[1]The University of Texas MD Anderson Cancer Center, Houston, TX; [2]UT Health Houston Graduate School of Biomedical Sciences, MD Anderson Cancer Center

Purpose: Radiation-induced cardiotoxicity poses a significant concern for lung cancer patients undergoing radiotherapy. This study aims to assess the predictive power of cardiac substructure dose for hs-cTnT elevation, a crucial biomarker for early detection of cardiac adverse events.

Methods: A cohort of 160 Non-Small Cell Lung Cancer (NSCLC) patients was used in this analysis. An in-house trained nnU-Net auto-segmentation model was used to delineate 19 cardiac substructures from the planning CT of each patient, including the whole heart, 4 chambers, 6 great vessels, 4 cardiac valves, and 4 coronary arteries. The impact of substructure dose-volume histogram (DVH) metrics on the hs-cTnT elevation was examined. A 100-iteration cross-validation with a split of 60%/40% were used to reduce the influence of random data split. Three logistic regression classification models were built: (1) a basic model, considering only clinicopathological factors; (2) a whole-heart model, integrating clinicopathological factors and whole-heart DVH metrics; (3) a substructure model, involving clinicopathological factors, whole-heart, and cardiac substructure DVH metrics. Model performance was assessed using area under the curve (AUC) and sensitivity.

Results: The incidence of hs-cTnT elevation was 31.9% among the 160 patients. The substructure model exhibited superior performance (median AUC: 0.73 [0.65, 0.74] (95% CI), median sensitivity: 0.69 [0.54, 0.69]),outperforming the basic model (AUC: 0.66 [0.60, 0.69], sensitivity: 0.62 [0.54, 0.63]) and the whole-heart mode (AUC: 0.64 [0.60, 0.69], accuracy: 0.62 [0.53, 0.66]). In contrast to whole-heart dose, coronary artery DVH metrics, in particular the left anterior descending coronary artery, showed greater feature importance in predicting hs-cTnT elevation, suggesting potential myocardial ischemia due to radiation-induced damage to coronary arteries in patients with hs-cTnT elevation.

Conclusion: We performed a comprehensive evaluation of using hs-cTnT as a surrogate to predict cardiactoxicity in lung cancer radiotherapy. Our findings underscored the superior predictive power of cardiacsubstructure DVH metrics.



Poster #2661 - Developing Artificial Intelligence Models for Medical Student Suturing and Knot-Tying Video-Based Assessment and Coaching

Madhuri B Nagaraj, M.D.[1,2]; Alexis Desir, M.D.[1]; Sofia Garces Palacios, M.D.[1]; Babak Namazi, Ph.D.[1]; Shekharmadhav Khairnar, M.S.[1]; Huu Phong Nguyen, Ph.D.[1]; Ganesh Sankaranarayanan, Ph.D.[1]; Daniel J. Scott, M.D.[1,2]

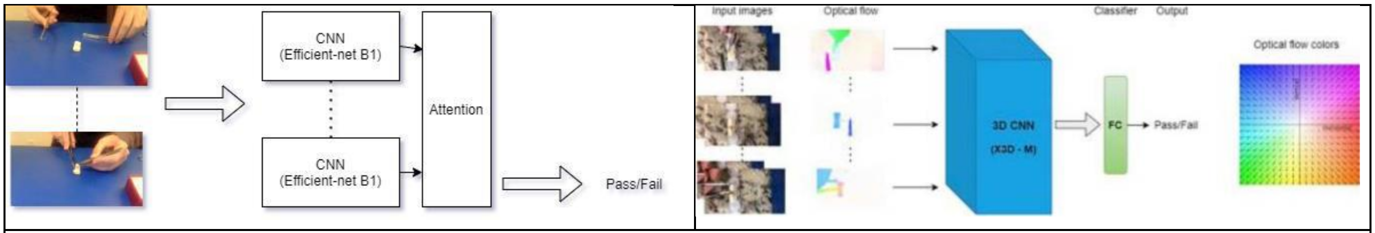
[1]Artificial Intelligence and Medical Simulation (AIMS) Lab, Department of Surgery, UT Southwestern Medical Center; [2]Simulation Center, UT Southwestern

Background: Early introduction and distributed learning have been shown to improve student comfort with basic requisite suturing skills. The need for more frequent and directed feedback, however, remains an enduring concern for both remote and in-person training. A previous in-person curriculum for our second-year medical students transitioning to clerkships was adapted to an at-home video-based assessment model due to the social distancing implications of COVID-19. We aimed to develop an Artificial Intelligence (AI) model to perform video-based assessment.

Methods: Second-year medical students were asked to submit a video of a simple interrupted knot on a penrose drain with instrument tying technique after self-training to proficiency. Proficiency was defined as performing the task under two minutes with no critical errors. All the videos were first manually rated with a pass-fail rating and then subsequently underwent task segmentation. We developed and trained two AI models based on convolutional neural networks to identify errors (instrument holding and knot-tying) and provide automated ratings.

Results: A total of 229 medical student videos were reviewed (150 pass, 79 fail). Of those who failed, the critical error distribution was 15 knot-tying, 47 instrument-holding, and 17 multiple. A total of 216 videos were used to train the models after excluding the low-quality videos. A k-fold cross-validation (k = 10) was used. The accuracy of the instrument holding model was 89% with an F-1 score of 74%. For the knot-tying model, the accuracy was 91% with an F-1 score of 54%.

Conclusions: Medical students require assessment and directed feedback to better acquire surgical skill, but this is often time-consuming and inadequately done. AI techniques can instead be employed to perform automated surgical video analysis. Future work will optimize the current model to identify discrete errors in order to supplement video-based rating with specific feedback.



(Let) Instrument holding error detection model (Right) Knot-tying error detection model

Poster #2662 - MLGAN: a Meta-Learning Based Generative Adversarial Network Adapter for Rare Disease Differentiation Tasks

Rui Li, Andrew Wen, Hongfang Liu

School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX

Abstract: Rare disease diagnosis is very challenging due to the rarity and lack of scientific knowledge. Comparing with tradi- tional diagnosis prediction task, rare disease detection has the unique challenges: (1) the prevalence is low; (2) The label of the negative data may contain noise. Due to the difficulty associated with rare disease detection, regarding patients without a rare disease diagnosis as negative samples introduces label noise. For each patient, the clinical record *x* can be viewed as a sequence of encounters, where each encounter record may contain multiple diagnosis codes. The label *y* is binary indicating whether a patient is diagnosed with the specific rare disease we are aiming to differentiate. The objective is to design a module that can directly adapt existing diagnosis prediction models to rare disease detection task.

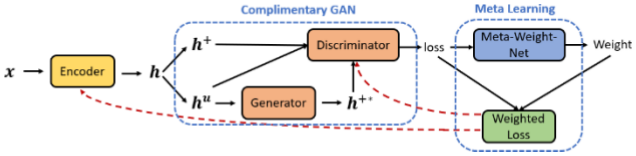


Figure 1: Framework for MLGAN.

We propose the module MLGAN that can be used to adapt existing diagnosis prediction methods to perform better on the rare disease detection task. Figure 1 shows the overview of the proposed MLGAN framework, consisting of two major components: (1) the complementary GAN component that generates the synthetic embedding of positive samples, and (2) the meta learning component that uses the Meta-Weight-Net (MW-Net) to impose weights on both real and synthetic sample loss. After a well-trained complementary GAN is obtained, the generator is used to generate the embedding of the synthetic positive samples to augment the positive data, and the discriminator is used as the classifier to classify whether the patient has a certain rare disease. For the complementary GAN, the real data *x* is fed into the Encoder (*Enc*), which refers to any patient embedding model, and we obtain the patient embeddings *h*. *h* contains the embeddings of positive samples h^+ and the embeddings of unlabeled samples h^u. h^u is fed into the generator of the complementary GAN, denoted as *G*, and we obtain the embedding of the synthetic positive samples $h^{+*} = G(h^u, \vartheta_g)$, ϑ_g are the parameters of the generator. Both h^+ and h^{+*} are fed into the discriminator. The discriminator tries to better classify the real positive sample h^+ and the synthetic positive sample h^{+*}. In order to enhance the robustness, MW-Net automatically assigns different weight to real data and synthetic data, which uses a multi-layer perceptron (MLP) mapping training loss to sample weight, and then iterating between weight recalculation and classifier updates.

We leverage data from Mayo Clinic Enterprise Data Warehouse. Four specific rare diseases were chosen as our focus, including idiopathic pulmonary fibrosis (IPF), mastocytosis (MAS), hypereosinophilic syndrome (HES) and rare kidney stones (RKS). For each disease, a list of possible differential diagnoses were first determined. The case cohort and control cohort were constructed, in which all patient diagnoses in the two years preceding the initial relevant disease diagnosis was retrieved, grouped temporally at an encounter level. We select three base models that are widely- used for any diagnosis prediction task, including GRU, Dipole and HiTANet, and we compare the performance with and without our module. We also consider two methods that are specially designed for the rare disease detection task, including medGAN and CONAN. We compared the performance of all baselines and the performance of three base models combined with our MLGAN module. The result shows that after combination with MLGAN, significantly improved performance is observed for all base models across all datasets, and the performance surpass the models designed specially for rare diseases.

Poster #2666 - One Working Model Is All you Need to Implement Surgical Tool Segmentation and Tracking on Any Procedural Video Using Segment Anything Model (SAM)

Shekharmadhav Khairnar, M.S.; Alexis Desir, M.D.; Sofia Garces Palacios M.D.; Huu Phong Nguyen, Ph.D.; Carla Holcomb, M.D.; Daniel J. Scott, M.D.; Ganesh Sankaranarayanan, Ph.D.

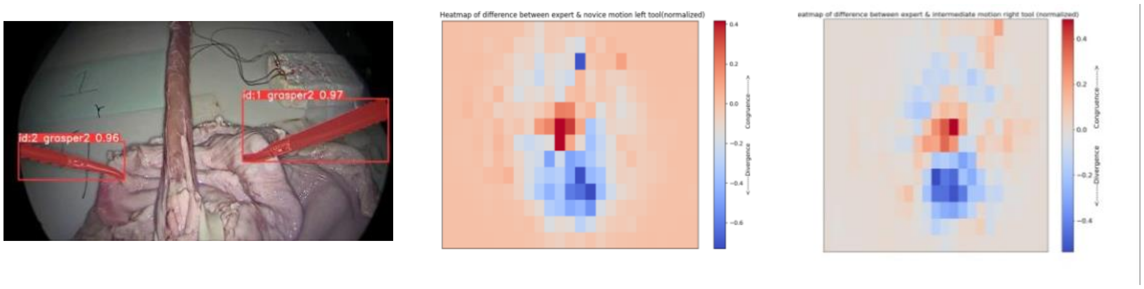
Artificial Intelligence and Medical Simulation (AIMS) Lab, Department of Surgery, UT Southwestern Medical Center, Dallas, TX

Introduction: Surgical tool tracking is vital for activity detection and performance assessment. AI based tool tracking requires annotating many images which is very labor intensive. Moreover, models trained using one data may not transfer well to others. Segment Anything Model (SAM) is an AI model that can segment any objects in an image without the need for additional training. The goal of this work is to develop a training pipeline that eliminates the need for manual annotation of images for surgical tool detection on any laparoscopic surgical procedures by using SAM.

Method: 500 images were randomly extracted from each video recording of a previously conducted study in which participants from varying experience levels performed laparoscopic Nissen fundoplication on an explanted porcine stomach model. The extracted images were then fed to a YOLOv8 model which was pretrained on Cholec80 dataset to detect and track laparoscopic surgical tools, generate bounding box prompts. The images along with their prompts were then provided to SAM to generate accurate segmentation masks. The resulting images and the masks were then split into test, training, and validation sets to train a new YOLOv8 model for the detection of tools in every frame of the videos. Performance of the surgical tool detection model was evaluated using mean average precision (MAP) for the range of intersection over union (IoU) thresholds and the F1 score. Heat maps of tool motions were generated and analyzed for differences.

Results: A total 64 videos were collected from 38 participants of which 19 (expert = 2, intermediate = 11, novice = 6) videos were used in this study. A total of 996 images were used for training of which 36 were background images, 306 for testing and 136 for validation. The mean average precision for IoU thresholds from 50% to 95% was 0.952. The F1 score for single class detection was 0.99. A bin size of 75 by 75 pixels was used to generate motion heatmaps for all the videos.
Visual inspection of difference heatmaps of normalized motion densities show a clear difference in concentration of motion both between experts and intermediates and novices.

Conclusion: By using SAM and a pretrained tool detection model, we demonstrated the generation of training images without the need for manual annotation. The new model trained using this approach showed outstanding performance and this approach can be used for tool detection and tracking in any surgical videos.



(left) Fundoplication simulation with tool tracking. (middle) Difference in motion density between expert and novice. (right) Difference in motion density between expert and intermediate.

Poster #2669 - Model Distillation for Extraction of Clinical Entities from Pathology Reports

David Hein[1], Alana Christie[2], Lindsay Cowell[2], Payal Kapur[3], Andrew Jamieson[1]

[1]Lyda Hill Department of Bioinformatics, UT Southwestern Medical Center; [2]Peter O'Donnell Jr. School of Public Health, UT Southwestern; [3]Harold C. Simmons Comprehensive Cancer Center, UT Southwestern

Background: This project aimed to extract and standardize clinical entities from unstructured renal cell carcinoma (RCC) pathology reports. Challenges included dealing with synonyms, contradictions between medical history and current test results, and the inflexibility of a preexisting regex-based tool. The goal was to develop an efficient, adaptable model that can train somewhat unsupervised. State-of-the-art models (SOTA) are powerful but expensive. A promising solution is to use small, open-source models like LLaMA 2 with parameter-efficient fine-tuning techniques, along with methods like "distillation step-by-step" to collect pseudo- labels from SOTA models for training.

Implementation: First we created a JSON schema containing items, labels, and item-specific instructions, which allowed for easy addition of new labels and quick modification of instructions. The next step involved using GPT-4 to segment relevant text based on specific and general instructions, while also providing a rationale for its decisions. The LLM then provided a JSON string formatted 'rationale' and label for each segment.

Python functions were developed to process the JSON string output as tabular data and store each LLM node's output for future training. Prompts were iteratively improved via manual review to enhance instruction following. QLoRA (Quantized Low-Rank Adaptation) was used to train the local model, a LLaMA 2 14B with 4-bit quantization and 16 rank 8 LoRA layers.

Results & Next Steps: Manual review of SOTA model outputs on histology of 360 specimens found only 5 misclassification (98.6% accuracy), providing encouragement that the pseudo-label process is performing well. Review of the LLaMA model output before and after fine-tuning on only 25% of generated rationales/labels found start empirical improvements in quality, with quantification of its performance and training on more data as our next steps.

Poster #2672 - The Influence of Deep-Brain Stimulation on Prefrontal Cortex Activity in Parkinson's Disease and Essential Tremor: Insights from an fNIRS Investigation

Mónica Lozano, B.S.[1,2]; Gabriel A. de Erausquin, M.D., Ph.D., M.Sc.[3]; Igor Zwir, Ph.D.[1,2,4]

[1]UTRGV School of Medicine, [2]UTRGV Institute of Neuroscience, [3]UT Health San Antonio, [4]Washington University School of Medicine in St. Louis

Abstract: Parkinson's Disease (PD) and Essential tremor (ET) are neurological movement disorders characterized by tremors, rigidity, and bradykinesia. Deep brain stimulation (DBS) has emerged as a promising therapeutic approach for managing PD and ET symptoms by modulating brain activity and improving motor function. However, the precise mechanisms underlying its efficacy remain incompletely understood.

In this study, we investigated cortical activation patterns in 14 PD/ET subjects with implanted DBS devices using functional near-infrared spectroscopy (fNIRS). Our experimental design included six sessions of inter- and intra-subject comparison before and after DBS stimulation. fNIRS recordings were conducted using the NIRx NIRScout X system, capturing data at 6.25 Hz with four different stimulation settings (C+0-, C+1-, C+2-, C+3-). Optodes were placed over the prefrontal cortex following the EEG 10/20 system. Data preprocessing involved filtering and trimming to 3 minutes per participant using NIRx Satori analysis software, with specific parameters applied for optimal data quality. Changes in oxyhemoglobin (HbO) and deoxyhemoglobin (HbR) concentrations were extracted and co-clustered to assess measurement robustness. Clustering analysis using Ward's method with half-square Euclidean distance categorized subjects within each response-time and co-clustered across times. The coincidence rate between domains (co-clustering) was assessed using hypergeometric statistics (Fisher's Exact Test) to establish patient trajectories.

Our analysis revealed distinct trajectories of cortical activation-response patterns across different DBS conditions, indicating differential neural responses to varying DBS parameters. These trajectories form an equifinality and multifinality temporal network, where similar patterns are shared before and after treatment or different patterns converge into a single post-treatment state. Integrating results with clinical data enables identification of patients best responding to person-centered treatments.

Overall, our findings contribute to understanding the neural mechanisms underlying DBS efficacy in PD/ET subjects. By elucidating cortical activation patterns associated with different DBS settings, our results offer insights into individualized treatment optimization strategies. Further research is warranted to explore the functional implications of these activation patterns and refine treatment strategies for neurodegenerative movement disorders.

Poster #2674 - Enhancing Radiological Efficiency and Accuracy: The Development and Implementation of an NLP-Driven Protocoling Algorithm

Dogan Polat, Ron Peshock, Yee Ng, Paulo Kuriki

UT Southwestern Medical Center

Objective: The aim is to develop a Natural Language Processing (NLP) algorithm for automating the protocoling process, thereby enhancing the efficiency and accuracy in radiological practices.

Methods: This is an IRB-approved, HIPAA-compliant study, encompassing 138,737 MRI and CT studies over 600 protocols, spanning 01/01/2015-11/1/2023. The DebertaV3 model was used with 80:10:10 partition for training, validation, and testing. Model was trained using reason for exam, imaging procedure and the International Classification of Diseases (ICD) codes. The model generated the top three most suitable protocols based on the input, and their performance was measured using a weighted F1 score.Additionally, a web interface was developed, enabling users to enter modality, body part, clinical information, and contrast contraindications to receive protocol recommendations.

Results/Findings: In the test cohort of 13,556 studies, model achieved accuracy levels of 82%, 97%, and 99% when top 1, top 2 and top 3 recommendations were assessed, respectively. When considering only the top recommended protocol, Deberta's accuracy ranged from 47% in MR cardiac imaging(lowest) to 94% in CT cardiac studies(highest). When the top three recommendations were considered, performance ranged from 87% in CT musculoskeletal imaging to 100% in various CT protocols including cardiac, chest, ENT, neuro, and spine.

Conclusion: We developed an NLP algorithm that significantly improves the protocoling process. The web interface emphasizes practicality and enhances user experience. Future research will explore data augmentation using LLM to combat class imbalance, compare performance with other architectures, and aim to integrate this algorithm into our current protocol workflow in the Electronic Health Record system.

Specialty	Count	Top 1	Top 2	Top 3
CT Cardiac	226	0.95	1.00	1.00
CT Neuro	1680	0.93	1.00	1.00
CT Spine	428	0.92	0.98	1.00
MR MSK	189	0.91	0.96	0.98
MR breast	157	0.90	0.98	1.00
CT ENT	665	0.90	0.99	1.00
CT Body	3183	0.83	0.97	0.99
Total	13556	0.82	0.96	0.98
CT Chest	2839	0.82	0.99	1.00
MR Spine	1075	0.80	0.98	0.99
MR Body	1000	0.79	0.95	0.98
MR Neuro	1873	0.71	0.90	0.95
CT MSK	128	0.50	0.63	0.70
MR Cardiac	113	0.50	0.70	0.78

Table 1: Algorithms performance by subspecialty

Poster #2676 - Chain-of-Thought Artificial Intelligence Agent for Visual Overlay Decision Making in COSCE Videos of Facial Exams

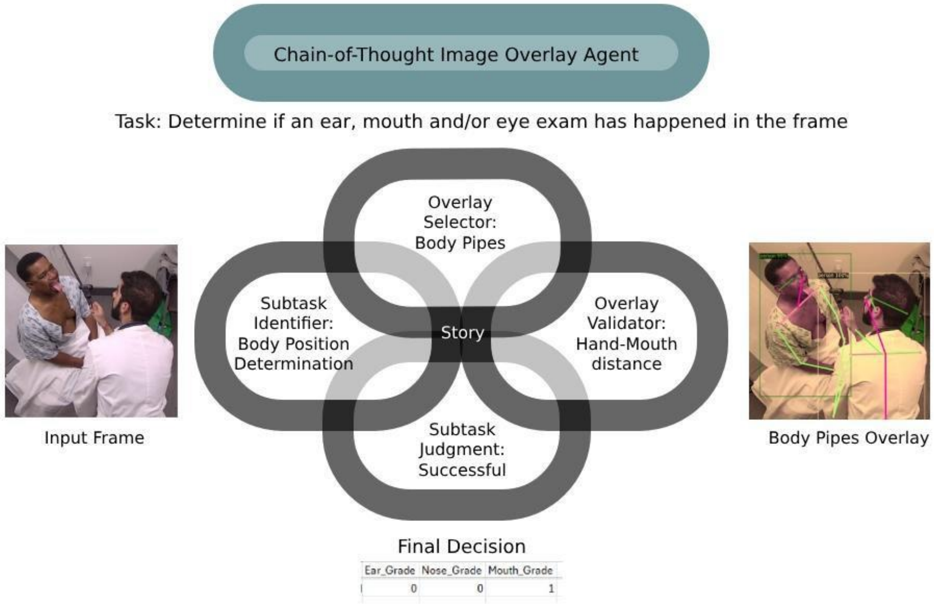
Sol Vedovato[1], Shinyoung Kang[1], Michael J. Holcomb[1], Ameer Hamza Shakur[1], Krystle K. Campbell[2], Daniel J. Scott[2,3], Thomas O. Dalton[4], Gaudenz Danuser[4], Andrew R. Jamieson[1]

[1]Lyda Hill Department of Bioinformatics, UT Southwestern Medical Center; [2]Simulation Center, UT Southwestern; [3]Department of Surgery, UT Southwestern; [4]Department of Internal Medicine, UT Southwestern

Abstract: In Objective Structured Clinical Encounter (OSCE) examinations, an accurate assessment of specific parts of the physical exams is crucial. To enhance this process, we developed a sequential decision-making artificial intelligence (AI) system capable of applying and verifying visual overlays on video frames. These overlays aid in identifying tools, positions, and gestures involved in facial exams such as ear, nose, and mouth examinations.

Our approach involves a chain-of-thought AI agent navigating through fixed prompts to identify specific physical exams depicted in video frames. The agent decides which visual overlays including scaffolding coordinates, hand detection, body piping, and facial landmarks are applied and then validates its findings against the original frames to prevent false detections. Continuing the chain, the agent iteratively assesses the sufficiency of visual information and applies additional overlays as needed.

After analyzing individual frames, the AI agent aggregates information to determine the facial exams depicted in each specific video. These determinations are compared against grading rubrics used in the original exams. This methodology was tested on over 200 video excerpts from a 2019 OSCE station, with comparisons made to the grades assigned by standardized patients and evaluators.


Chain-of-Thought Image Overlay Agent
Task: Determine if an ear, mouth and/or eye exam has happened in the frame
Overlay Selector: Body Pipes
Subtask Identifier: Body Position Determination
Story
Overlay Validator: Hand-Mouth distance
Subtask Judgment: Successful
Input Frame
Body Pipes Overlay
Final Decision

Ear_Grade	Nose_Grade	Mouth_Grade
0	0	1

Poster #2677 - Using Robotics-Aided Upper Limb Evaluation to Predict Early-Stage Parkinson's Disease in Clinical Practice: A Machine Learning Case Study

Daniel Salinas, Diego Rojano, Marysol Cabello, Tomas Gomez, Chris Cavazos, Nawaz Khan Abdul Hack, Ramu Vadukapuram, Igor Zwir, Kelsey Baker

Neuro and Behavioral Health Integrated Service Unit, Division of Neuroscience, University of Texas Rio Grande Valley School of Medicine, Harlingen, TX

Background: Parkinson's Disease (PD) is characterized by both motor and non-motor symptoms, and its diagnosis primarily relies on clinical presentation. There is a growing need for diagnostic tools to identify the early signs of PD in a variety of populations, motor impairments often manifested as tremors with weakness in the upper extremities. In the observed population, there are often varying degrees of PD severity of unequal naturally occurring proportions of Hoehn and Yahr (HY) level ratings. Our study focuses on using machine learning techniques with hand robotic evaluation to predict the PD clinical stage. Specifically, we have sought to evaluate how upper limb motor function and muscle strength performance variables obtained on the InMotion 2 Arm can be utilized to identify the PD patient and Healthy Controls.

Methods: Eight individuals with PD and 39 Healthy individuals participated in the study. Participants completed two to four arm evaluations on the InMotion2 Arm robot using both Left and Right hands. Data preprocessing and predictive modeling were analyzed using R. Due to sample size, data imbalance, and overfitting issues, the data was randomly split to 50% training and 50% testing and used three repeated 10-fold cross-validations. Training data was balanced using the Synthetic Minority Over-sampling Technique (SMOTE) with the minority class being PD Condition. Models included Random Forest, Neural Networks, Naive Bayes, Support Vector Machines with Linear Kernel (SVM), and Penalized Multinomial Logistic Regression.

Results: Performance metrics suggests that a hyperparameter-tuned eXtreme Gradient Boosting model outperformed other algorithms, achieving an overall model accuracy of 92.9%, interclass balanced accuracy of 73.1%, Kappa of 46.2%, average weighted AUC of 86.6%, and a model learning time of 0.05 seconds. The top variables included: hold distance compass, hold distance, smoothness, displacement compass, reach error, and path error.

Discussion: Our results suggest that the InMotion2 hand evaluation may be used to predict PD patient upper limb function in a PD population. Future work will seek further participant recruitment and the development of ensemble algorithms with gait analysis.

Poster #2679 - Enhancing Clinical Decision-Making: MRI-Based Deep Learning for Confidence-Informed IDH Prediction in Gliomas

Chandan Ganesh[1], Nghi Truong[1], Benjamin Wagner[1], Divya Reddy[1], James Holcomb[1], Niloufar Saadat[1], Kimmo Hatanpaa[1], Toral Patel[1], Baowei Fei[2], Matthew Lee[3], Rajan Jain[3], Richard Bruce[4], Marco Pinho[1], Ananth Madhuranthakam[1], Joseph Maldjian[1]

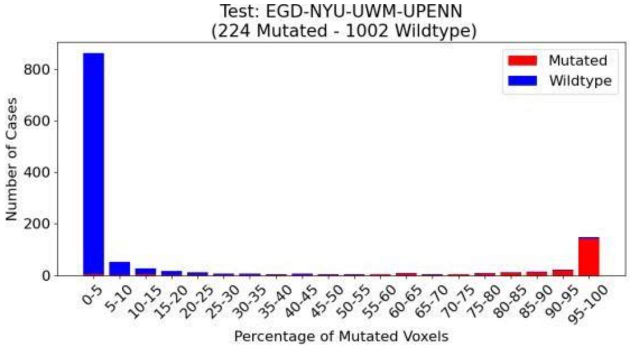
[1]UT Southwestern Medical Center, [2]University of Texas at Dallas, [3]NYU Grossman School of Medicine, [4]University of Wisconsin School of Medicine

Objective: This study aims to enhance the clinical application of deep learning networks (DLNs) for predicting IDH mutation status in gliomas. We address the gap between DLN predictions and their clinical relevance by developing a framework to integrate Confidence-Scores (CS) with traditional metrics like accuracy, sensitivity, and specificity.

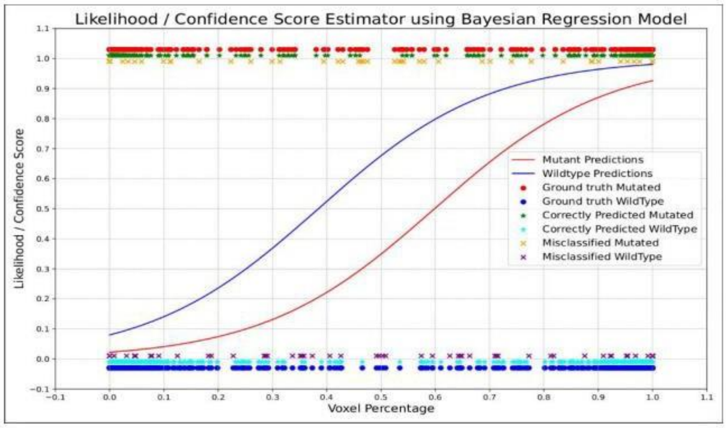
Materials and Methods: We developed an MRI-based-DLN (*MC-net*) for non-invasive prediction of IDH using a UNet. It was implemented for a voxel-wise dual-class-segmentation of the whole-tumor & IDH-prediction. *MC-net* was trained on 1082 cases (TCIA+UTSW+UCSF) and evaluated on 1236 cases (NYU+UWM+EGD+UPENN). We employed a Bayesian logistic regression (BLR) model to calculate confidence scores (CS) for each prediction, using predicted voxel percentages and ground truth from the test data. *MC-net+BLR* was further tested and validated on 196 additional UTSW cases.

Results: *MC-net* achieved an excellent test-accuracy of 96.4%. Its voxel-wise prediction showed a distribution of most cases classified with very low (0-5%) or very high (95-100%) percentages of mutated-voxels (Fig 1). This bimodal distribution suggests that *MC-net* is highly effective in predicting IDH-status. *MC-net+BLR* achieved an accuracy of 94.4% on additional test-cases providing CS higher than the predicted accuracy (Fig 1).

Conclusion: We developed a DL-framework to enhance clinical decision making using DLN-predictions. The CS represents a tangible-measure of the DLN's reliability in predicting IDH-status. Providing clinicians with the IDH status and a confidence score, this framework supports a more nuanced interpretation of diagnostic results and informed treatment planning. This approach has the potential to enhance patient outcomes significantly by enabling confidence-informed therapeutic strategies for gliomas.



Distribution of percent mutant voxels on test dataset. The distribution shows very low (0-5%) or very high (95-100%) percentages of predicted mutated voxels, suggesting a clear differentiation between mutated and wildtype gliomas.



Regression curves to estimate Confidence Scores.

Poster #2681 - Evaluation of an Echo-Based Artificial Intelligence Algorithm for Detecting HFpEF

K. Yaros, M. Segar, V. Subramanian, A. Chandra, T. Koshy, R. Upton, A. Akerman, A. Pandey

Background: Diagnosis of Heart failure with preserved Ejection Fraction (HFpEF) is challenging and often requires invasive or non-invasive assessment of LV filling pressure. Recent FDA-cleared echo-based AI HFpEF model employs a 3-dimensional convolutional neural network to detect HFpEF using a single 4-chamber clip from a resting echocardiogram. However, the external validation of this algorithm is limited.

Purpose: To evaluate the diagnostic and prognostic performance of the AI model in a cohort of HFpEF patients and matched controls.

Methods: Cases were clinically adjudicated based on the history, signs, and symptoms, LVEF (>45%), and evidence of elevated filling pressures by resting (PCWP > 15 mm Hg) or exercise invasive hemodynamics (PCWP > 25 mm Hg) or echocardiogram (E/e' >14).Controls were age, sex, and BMI matched subjects without heart failure and a normal echocardiogram. The performance of the model was evaluated using ROC curves. The association of the AI-HFpEF phenotype with elevated PCWP and peak exercise oxygen uptake (VO2peak) was assessed using multivariable logistic and linear regression models adjusted for age, sex, race, BMI, and comorbidities (diabetes, hypertension, kidney disease, atrial fibrillation). Based on Youden's index, probability threshold >0.75 was the optimal cutoff for detecting HFpEF (sensitivity 0.85, accuracy 0.74, and specificity 0.66).

Results: Among patients with clinically adjudicated HFpEF and matched controls (N = 122 each), the AI algorithm outperformed HF2PEF score in identification of adjudicated and hemodynamically confirmed HFpEF (AUROC: 0.75 for each vs 0.69 and 0.70, resp). In the HFpEF cohort, a higher AI-based probability of HFpEF was associated with lower VO2peak (b [95% CI] per 5% higher probability: -0.11 [-0.21 to -0.01, P-value: 0.03] and greater odds of elevated PCWP (Odds ratio [95% CI] per 5% higher probability: 1.07 [1.01 – 1.15, P-value: 0.04] at rest or exercise adjusted for confounders.

Conclusion: The model demonstrated excellent sensitivity and discrimination in identifying clinical HFpEF and prognostic utility for disease severity categorization.

Poster #2683 - Optimizing AI Thresholds for the Detection of Actionable Coronary Artery Calcium in Chest CT Imaging: A Validation Study Against Human Scoring

Maya Wiessman, M.D.; James Wyatt Miller, B.S.; Ann Marie Navar, M.D., Ph.D.; Parag Joshi, M.D.; Travis Browning, M.D.; Arzu Canan, M.D.; Suhny Abbara, M.D.; Ronald M. Peshock, M.D.; Fernando Kay, M.D., Ph.D.

UT Southwestern Medical Center, Dallas, TX

Objectives: We implemented an AI software to quantify coronary artery calcium volume (AI-CACvol) in chest CT scans at UT Southwestern in September 2022. Our goal was to retrospectively establish optimal detection thresholds for actionable CAC and validate their performance in a post- implementation population. AI results were compared against qualitative human CAC scoring (H-CACqual).

Methods: Eligible patients had undergone a non-contrast chest CT and a cardiac CT with Agatston CAC scoring (CACscore, standard of reference). CACvol was calculated using commercially available AI software (AI Rad Companion, Siemens), and H-CACqual involved subjective CAC scoring by a board-certified radiologist following 2016 SCCT/STR guidelines.

Results: The derivation cohort included 333 patients (59% female, median age 62 years, range 28–92 years). Median AI-CACvol was 21.0 mm³ (IQR: 193.0 mm³); median CACscore was 30.7 AU (IQR: 250.0 AU). The Pearson's correlation coefficient for log-transformed AI-CACvol and CACscore was 0.91 (95% CI: 0.88–0.92). H-CACqual showed significant association with CACscore, Spearman's correlation coefficient 0.90 (95% CI: 0.88–0.92). ROC curve analysis for detecting CACscore ≥ 100 AU showed an AUC of 0.97 for AI-CACvol vs. 0.92 for H-CACqual ($p < 0.001$). Optimal thresholds were 70.1 mm³ for AI-CACvol and "moderate" or greater for H-CACqual. In the prospective cohort of 39 patients (75% female, median age 67 years, range 46–80 years), median CACscore was 43 AU (IQR: 215 AU). Using the optimized threshold, AI-CACvol achieved 100% sensitivity and 88% specificity for detecting higher-risk patients, compared with 71% sensitivity and 100% specificity for H-CACqual.

Conclusion: AI-CACvol correlates strongly with the Agatston method and could be a precise, accessible metric for opportunistic cardiovascular risk screening, offering high-performance thresholds that might inform future automated best practice alerts.

Poster #2684 - Automated Cell Quantification and Intensity-Based Analysis Reveal Glioblastoma Multiforme Cell Heterogeneity in High-Throughput Microwell Devices

Fnu Bilal[1], Jesung Moon[4], Smita Rindhe[1,2,3], Elizabeth A. Maher[1,2,3,4], Robert M. Bachoo[1,2,3,4]

[1]Department of Internal Medicine, UT Southwestern Medical Center, Dallas, TX; [2]Simmons Comprehensive Cancer Center, UT Southwestern; [3]Annette G. Strauss Center for Neuro-Oncology, UT Southwestern; [4]Department of Neurology, UT Southwestern

Abstract: Glioblastoma (GBM), the most common adult primary malignant brain cancer is lethal despite multimodality aggressive therapies. GBM therapeutic resistance is, in part, driven by high intra- tumoral heterogeneity, which enables survival of drug resistant persister cell populations to drive treatment resistant disease. While there have been significant advances in single-cell genomic technologies, to date there is no ex vivo platform for single-cell phenotype analysis. We have designed and fabricated high throughput (HT), high content (HC) microwell devices that enable tracking and retrieval of thousands of individual cells. The cells are confined in microwells that approximate the size of single cells (15um x15um) that physically restrict cellular expansion required for symmetric division, or in larger (50x50um) microwells that permit single cell clonal expansion. The ability to track thousands of individual non-dividing dormant cells and cells capable of clonal expansion is critical for identifying drugs that can kill treatment resistant persister cells. To quantify thousands of individual GBM cells in the size restricted microwells (dormant) or those undergoing clonogenic growth we have developed an intensity-based automated pipeline using nuclear florescent labeling (H2B or DAPI) of an established GBM cell line (U251) and a primary human GBM cell line. In the first approach, we quantify cell numbers between days 1 and 5 using an automated pipeline integrating machine learning (ilastik) and image processing (MATLAB). Ilastik tunes a pixel classifier based on textural features from annotated exemplars, while MATLAB performs postprocessing, including thresholding and labeling, to extract single-cell statistics. A novel cell-well assignment strategy using Euclidean distances enables robust tracking of individual cells and their progeny. Domain knowledge integration refines the detected regions, providing accurate cell counts. The second approach focuses on intensity-based analysis using Imaris software to assess clonogenic potential on day 5. By comparing fluorescence intensity between days 1 and 5 we can identify and characterize the clonal expansion of GBM cells. With these complementary analytic tools, the variability in proliferative capacity and self-renewal properties of individual GBM cells can be quantified. Notably, we employ these two parallel approaches not only to gain a more comprehensive understanding of GBM cell heterogeneity but also to validate the results of each approach against the other. Together, these approaches will be used to derive insights into GBM biology and begin to identify novel therapeutic targets.

Poster #2685 -Improving Referrals in Pediatric Subspecialties: Exploring the Use of Standardized Electronic Referrals and Artificial Intelligence

Kyra L. Andre, B.S.N., RN, CPN

The University of Texas at Tyler, School of Nursing

Abstract: Health care advancements continue to improve viability for extremely premature infants and extend the lifespan for children born with congenital anomalies. Additionally, interventions such as vaccinations have minimized or eliminated deadly pediatric illness and more children are living to adulthood with chronic conditions. The pediatric subspecialty workforce is struggling to meet the demand for increased referrals. A standardized referral process is needed to minimize unnecessary referrals and delays in care. The purpose of this research was to review the literature for insight into the current use of standardized referrals and artificial intelligence (AI) systems to determine if the evidence supports use in the pediatric specialty environment.

A targeted keyword search of available literature in CINAHL and PubMed was performed to identify current use of both electronic standardized referrals and AI. The articles identified were then read in full and the data extrapolated was synthesized to identity the effectiveness and benefit of the systems utilized.

Use of standardized electronic referral systems can reduce the number of unnecessary referrals and improve provider-to-provider communications in recommending care in place of the patient being seen by the specialist. Artificial Intelligence has shown promise in referral reviews in determining the appropriateness of a referral and in assisting PCPs in identifying when a referral is indicated for their patients.

Poster #2686 - Multi-Level Contrastive Learning for Protein-Ligand Binding Residue Prediction

Ruheng Wang, Tingyi Wanyan, Xiaowei Zhan
Quantitative Biomedical Research Center, Peter O’Donell Jr. School of Public Health, UT Southwestern, Dallas, TX

Background: Protein-molecule interactions play a crucial role in various biological functions, with their accurate prediction being pivotal for drug discovery and design processes. Traditional methods for predicting protein-molecule interactions are limited. Some can only predict interactions with a specific molecule, restricting their applicability, while others aim for multiple types but fail to effectively utilize information across different interactions, leading to increased complexity and inefficiency.

Objective: The objective of this project is to develop an original model to improve the prediction of multiple molecule-protein interactions and the identification of potential molecule-binding residues.

Methods: In this work, we propose a novel deep learning model named MucLiPred and a dual contrastive learning mechanism. As in Figure 1, we proposed two novel contrastive learning paradigms at residue and type levels, training the discriminative representation of samples. The residue-level contrastive learning hones in on distinguishing binding from non-binding residues with precision, shedding light on nuanced local interactions. In contrast, the type-level contrastive learning delves into the overarching context of molecule types (such as DNA or RNA), ensuring that representations of identical molecule types gravitate closer in the representational space and bolstering the model's proficiency in discerning interaction motifs, enhancing the model's ability to recognize global interaction patterns.

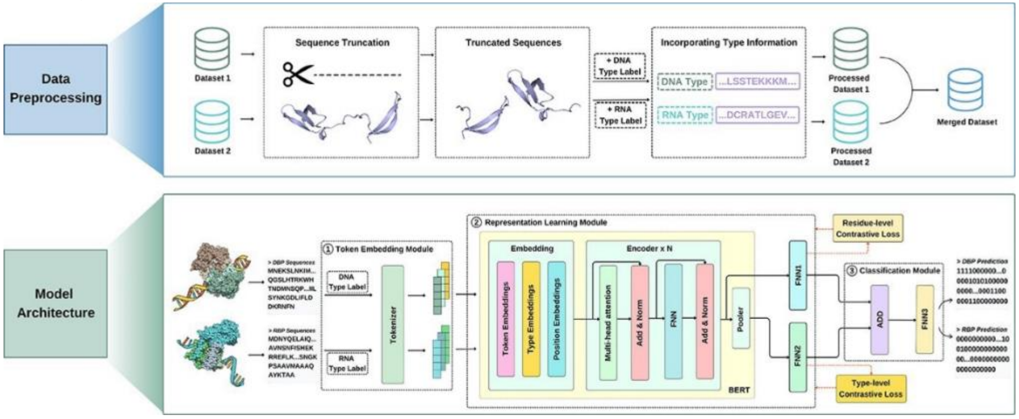


Figure 1. *The workflow and framework of the proposed MucLiPred.*

Results: Empirical findings underscore MucLiPred's dominance over existing models, highlighting its robustness and unparalleled prediction accuracy. We demonstrate that after introducing residue-level contrastive learning module, it is capable of adaptively learning more discriminative and high-quality representations of the binding residues for our model, fully utilizing the majority of samples in the imbalanced dataset. The type-level contrastive learning module allows the model to pay attention to the category information of the sequences, forming different feature representations for different categories, thereby enabling the model to predict the binding residues of different types of molecules simultaneously.

Conclusions: This approach culminates in nuanced multi-molecule predictions, unraveling relationships between various molecule types, and fortifying the potential for precise protein-molecule interaction predictions. The integration of dual contrastive learning techniques amplifies its capability to detect potential molecule-binding residues with precision. Optimizing the model to separate representational and classification tasks improved performance, establishing MucLiPred as an innovative tool for protein-molecule interaction prediction and setting a new precedent for future research in this field.

Poster #2687 - Binary Classification of Chest X-Rays: An Analysis of Accuracies Using Several Deep Learning Methods with Varying Sample Sizes and Compositions

Rao P. Pokala

The University of Texas at Austin

Abstract: Reading Chest X-rays (CXR) is a critical part of diagnosis by radiologists, pulmonologists, and physicians. CXR can be visually complex with subtle abnormalities in distinguishing anatomical structures, leading to potential errors in diagnosis. It is further complicated by overlapping organs and tissues, or the presence of artifacts. Interpreting information from multiple sources, including the appearance of lung fields, cardiac silhouette, mediastinum, pleura, and bony structures increases the complexity. Certain pathologies might have similar contrast levels to surrounding tissues, hindering detection. In addition, cognitive biases, information overload, and variability in image quality can complicate the diagnosis. This paper analyzes CXR provided with labels of "Normal" and "Pneumonia", obtained from Kaggle, that has 1,342 normal CXR (25%), and 3,944 pneumonia CXR (75%). This paper uses several deep learning methods to analyze the accuracies with training and validation sets of varying sizes with compositions for 25%, 40%, 50%, 60%, 75% of normal CXR and the rest with pneumonia. The results are presented with recommendations which indicate that for balanced (50% normal and 50% pneumonia CXR), TensorFlow based Convolutional Neural Network (CNN) with just two layers using a kernel size of 3x3 provide the best results, as evidenced by the Area Under the Receiver Operating Characteristic (AUROC) and the Area Under Precision-Recall Curve (AUPRC) measures. For extremely imbalanced (25% normal, 75% pneumonia, or 75% normal, and 25% pneumonia CXR) sample sizes, the higher the number of CNN layers of four or more with batch normalizations and dropouts provide the best results. Overall, the accuracies for training reach almost 100% with 50 epochs, and validation accuracies also reach 99.7% for both these deep learning methods.

Poster #2691 - Synthetic Lies: Characterizing AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions

Jiawei Zhou[1,2], Munmun De Choudhury[1]

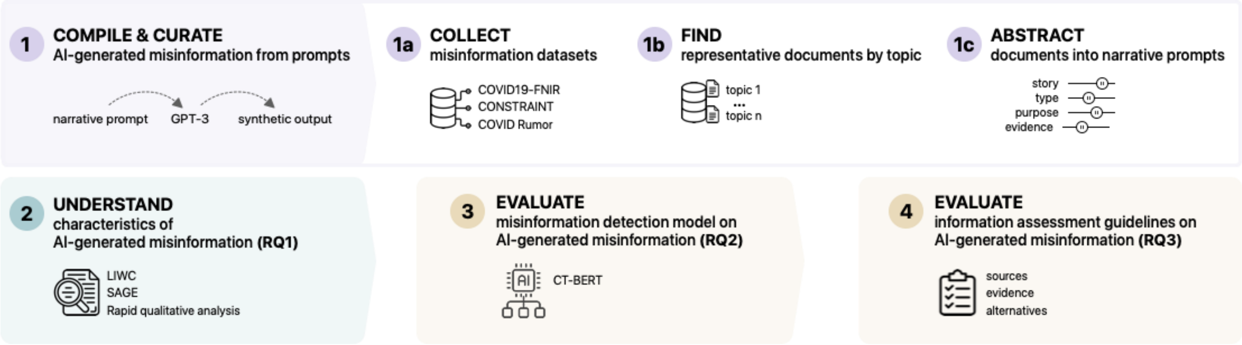
[1]School of Interactive Computing, Georgia Institute of Technology; [2]Center for Translational AI Excellence and Applications in Medicine, UTHealth Houston

Introduction: Large Language Models (LLMs) — machine learning algorithms that can recognize, predict, and generate human languages on the basis of large sets of human-written content — are now widely used in producing human-like texts. They have the ability to create high-volume human-like texts and can be used to generate persuasive misinformation with a scalability we have never seen before. However, the risks remain under-explored.

With off-the-shelf LLMs becoming more accessible to the general public, we respond to this critical gap by 1) examining the differences between AI-generated and human-created misinformation, and 2) assessing the extent to which pre-existing solutions are applicable to AI-generated misinformation.

Method: We compiled a dataset of human-created misinformation from existing work and extracted the most representative documents. Guided by Narrative Theory, we abstracted representative documents into *"narrative prompts"* that captured the core narrative elements for GPT3 to output AI-Misinfo.

With the paired human and AI misinformation, we first examined the characteristics of AI-generated misinformation through text analysis and rapid qualitative analysis. Next, we evaluated two common and pre-existing misinformation solutions: misinformation detection models and information assessment guidelines developed by journalists.



1 COMPILE & CURATE
AI-generated misinformation from prompts
narrative prompt → GPT-3 → synthetic output

1a COLLECT
misinformation datasets
COVID19-FNIR
CONSTRAINT
COVID Rumor

1b FIND
representative documents by topic
topic 1
...
topic n

1c ABSTRACT
documents into narrative prompts
story
type
purpose
evidence

2 UNDERSTAND
characteristics of
AI-generated misinformation **(RQ1)**
LIWC
SAGE
Rapid qualitative analysis

3 EVALUATE
misinformation detection model on
AI-generated misinformation **(RQ2)**
CT-BERT

4 EVALUATE
information assessment guidelines on
AI-generated misinformation **(RQ3)**
sources
evidence
alternatives

Results: Our results suggest significant linguistic differences in AI-generated misinformation as it had more emotions and cognitive processing expressions than human creations. We also observed that AI- generated misinformation tended to enhance details, communicate uncertainties, draw conclusions, and simulate personal tones.

We discovered existing detection models had performance degradation when classifying AI-generated misinformation as opposed to human creations. Similarly, information assessment guidelines had questionable applicability, as AI-generated misinformation was more likely to mimic criteria in credibility, transparency, and comprehensiveness.

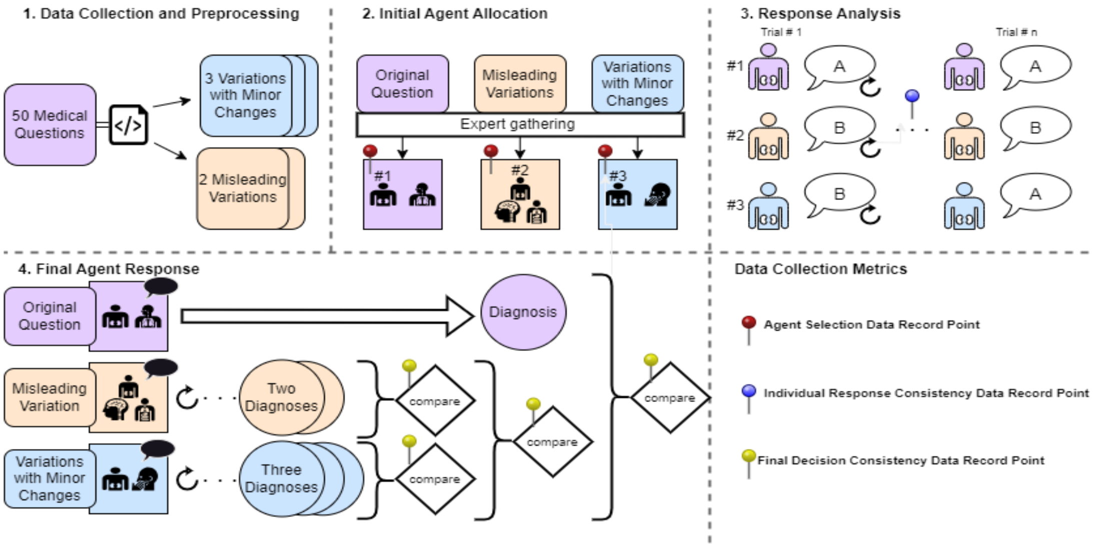
Conclusion: We found SOTA LLM not only failed to identify misinformation in prompts but also amplified it via hallucinated evidence, emotional appeal, and fabricated details. Current algorithmic and human solutions have questionable applicability in handling the new form of misinformation.

Poster #2692 – Evaluating Medagent Consistency In Medical Diagnoses Through Iterative Testing

Ayman Mahfuz, Rohit Akash Rajendran, Seheon Chang, Conan Sum, Ying Ding

University of Texas at Austin

Abstract: The transformative potential of Large Language Models (LLMs) in medical diagnostics is profound, offering unparalleled access to domain-specific knowledge and reasoning capabilities. Building upon the foundational work of "MEDAGENTS: Large Language Models as Collaborators for Zero-shot Medical Reasoning." In an attempt to improve the accuracy of diagnosis that MedAgents framework provides, by reducing mis-retrieval of domain knowledge through the Autogen library, we discovered that the MedAgents framework chooses different agents and corresponding responses on different attempts.This project pivots from accuracy to consistency as a metric of reliability and performance in medical diagnosis. Based on this issue, this study proposes a detailed exploration into the consistency of LLM-based medical agents (MedAgents) in providing diagnoses. To rigorously assess the consistency of MedAgents in diagnosing medical conditions through repetitive testing of identical prompts and their variations, thereby identifying patterns in agent selection and response stability. The methodology involves the selection of 50 medical questions from an existing dataset for testing the MedAgent system. The focus lies on assessing the consistency of expert selection, individual agent responses, and final agent responses to prompts. Figure below represents the overview of this study's experimental design. Additionally, variations to prompts are introduced to gauge MedAgent consistency under different conditions. These variations include three minor changes maintaining the medical context and two variations with potentially misleading details. The analysis aims to evaluate response consistency across repeated tests, frequency and diversity of agent selection, and the impact of question variations, particularly the influence of misleading information on agent responses and diagnostic outcomes.Expected outcomes entail understanding LLM consistency and identifying strengths and vulnerabilities. This study's significance lies in advancing AI in healthcare and guiding future enhancements through studying its consistency.



*Keywords: Large Language Models(LLMs), MedAgents, Medical Diagnostics, Agent Selection, AI Consiste*ncy

Poster #2694 - ISABR-SELECT: A Clinical-Radiomics Model for Personalized Immunotherapy in Early-Stage NSCLC

Maliazurina B. Saad, Ph.D.[1,†]; Eman Showkatian, M.S.[1,†]; Qasem Al-Tashi, Ph.D.[1]; Muhammad Aminu, Ph.D.[1]; Xu Xinyan, M.D.[1,2]; Mohamed Qayati Mohamed, M.D.[1]; Morteza Salehjahromi, Ph.D.[1]; Sheeba J. Sujit, Ph.D.[1]; Steven H. Lin, M.D.[2]; Zhongxing Liao, M.D.[2]; Saumil Gandhi, M.D.[2]; David Qian, M.D.[2]; David Jaffray, Ph.D.[1,3]; Caroline Chung, M.D.[2,3]; Natalie Vokes, M.D.[4]; Jianjun Zhang, M.D.[4]; John V. Heymach, M.D.[4,‡]; Jia Wu, Ph.D.[1,3,4,‡]; Joe Y. Chang, M.D.[2,‡]

[1]Department of Imaging Physics, The University of Texas MD Anderson Cancer Center, Houston, TX; [2]Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center; [3]Institute of Data Science in Oncology, The University of Texas MD Anderson Cancer Center; [4]Department of Thoracic/Head and Neck Medical Oncology, The University of Texas MD Anderson Cancer Center
† Contributed equally; ‡Co-senior authors

Background: Our recent Phase II randomized I-SABR trial demonstrated the improvement of event-free survival (EFS) of combining SABR with immune checkpoint inhibitor in early-stage NSCLC compared to SABR alone. However, not every patient benefit from adding immunotherapy. As a part of the secondary endpoints, we reported here a subsequent biomarker study by leveraging clinical-radiomics and machine learning to identify patients who potentially benefit from adding immunotherapy.

Methods: 141 early-stage NSCLC patients enrolled in I-SABR trial were analyzed and divided into discovery (n=101) and validation cohorts (n=40). We developed I-SABR-SELECT framework to model heterogeneous treatment effects and inform patient selection for combining immunotherapy with SABR. Radiomics patterns of the tumor/ peritumoral and lung region, as well as tumor-surrounding angiogenesis network were profiled. Radiomics features were harmonized, qualified, and integrated with clinical factors for downstream selection to mitigate model overfitting. The relationships between patient characteristics and treatment outcome were model separately for I-SABR and SABR arms using Random survival forest. Counterfactual reasoning was implemented to assess the individualized treatment effects and optimize selection. Evaluation was performed separately on I-SABR discovery and validation cohorts, as well as an independent external STARS trial of SABR treated early-stage NSCLC.

Results: Overall, the model recommended 46 out of 141 patients enrolled to I-SABR to switch treatments including 34 out of 75 in the SABR arm and 12 out of 66 in the I-SABR arm. Patients treated according to model's recommendation achieved significantly improvement of EFS in both arms during model discovery and validation. Stratified by this recommendation, SABR patients adding immunotherapy showed an EFS 1.1 to 1.6 times longer than those without immunotherapy. Notably, patients who were treated according to I-SABR-SELECT recommendation exhibited a significant increased EFS with HR = 22.8 ($p<0.001$), when compared to matched counterparts who did not treat as recommended. Conversely, when the model suggested SABR monotherapy, no statistical survival difference was observed between patients who were treated according to or against recommendation ($p=0.29$). In the benefit stratum by the model, the average immunotherapy effect was over two-fold greater than in the randomized trial. In addition to worse performance status, a less complex angiogenesis network and larger tumors were associated with more benefits from combining immunotherapy and SABR.

Conclusions: I-SABR-SELECT provides individualized approach to guide who can benefit from combination of immunotherapy and SABR in early-stage NSCLC.

Poster #2695 – Deep Learning Signature from Pretreatment Chest CT Associated with Immunotherapy Benefit in EGFR/ALK-Negative NSCLC

Maliazurina B. Saad, Ph.D.[1,†]; Lingzhi Hong, M.D.[1,2,†]; Muhammad Aminu, Ph.D.[1,†]; Natalie I. Vokes, M.D.[2,3,†]; Pingjun Chen, Ph.D.[1]; Morteza Salehjahromi, Ph.D.[1]; Kang Qin, M.D.[2]; Sheeba J. Sujit, Ph.D.[1]; Xuetao Lu, Ph.D.[4]; Elliana Young, M.S.[5]; Qasem Al-Tashi, Ph.D.[1]; Rizwan Qureshi, Ph.D.[1]; Carol C. Wu, M.D.[6]; Brett W. Carter, M.D.[6]; Steven H. Lin, M.D.[7]; Percy P. Lee, M.D.[1,7]; Saumil Gandhi, M.D.[7]; Joe Y. Chang, M.D.[7]; Ruijiang Li, Ph.D.[9]; Michael F. Gensheimer, M.D.[9]; Heather A. Wakelee, M.D.[10,11]; Joel W. Neal, M.D.[10,11]; Hyun-Sung Lee, M.D.[12]; Chao Cheng, Ph.D.[13]; Vamsidhar Velcheti, M.D.[14]; Yanyan Lou, M.D.[15]; Milena Petranovic, M.D.[16]; Waree Rinsurongkawong, Ph.D.[4]; Xiuning Le, M.D.[2]; Vadeerat Rinsurongkawong, Ph.D.[4]; Amy Spelman, Ph.D.[2]; Yasir Y. Elamin, M.D.[2]; Marcelo V. Negrao, M.D.[2]; Ferdinandos Skoulidis, M.D.[2]; Carl M. Gay, M.D.[2]; Tina Cascone, M.D.[2]; Mara B. Antonoff, M.D.[17]; Boris Sepesi, M.D.[17]; Jeff Lewis, B.S.[4]; Ignacio I. Wistuba, M.D.[18]; John D. Hazle, Ph.D.[1]; Caroline Chung, M.D.[7,8]; David Jaffray, Ph.D.[1,19]; Don L. Gibbons, M.D.[2]; Ara Vaporciyan, M.D.[17]; J. Jack Lee, Ph.D.[4]; John V. Heymach, M.D.[2,‡]; Jianjun Zhang, M.D.[2,3,‡]; Jia Wu, Ph.D.[1,2,‡]

[1]Department of Imaging Physics, The University of Texas MD Anderson Cancer Center, Houston, TX; [2]Department of Thoracic/Head and Neck Medical Oncology, The University of Texas MD Anderson Cancer Center; [3]Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center; [4]Department of Biostatistics, The University of Texas MD Anderson Cancer Center; [5]Department of Enterprise Data Engineering & Analytics, The University of Texas MD Anderson Cancer Center; [6]Department of Thoracic Imaging, The University of Texas MD Anderson Cancer Center; [7]Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center; [8]Department of Neuroradiology, The University of Texas MD Anderson Cancer Center; [9]Department of Radiation Oncology, Stanford University School of Medicine, Palo Alto, CA; [10]Department of Medicine, Division of Oncology, Stanford University School of Medicine, Stanford, CA; [11]Stanford Cancer Institute, Stanford, CA; [12]Systems Onco-Immunology Laboratory, David J. Sugarbaker Division of Thoracic Surgery, Michael E. DeBakey Department of Surgery, Baylor College of Medicine, Houston, TX; [13]Institute for Clinical and Translational Research, Baylor College of Medicine, Houston, TX; [14]Department of Hematology and Oncology, New York University Langone Health, New York, NY; [15]Division of Hematology and Oncology, Mayo Clinic, Jacksonville, FL; [16]Department of Radiology, Massachusetts General Hospital, Boston, MA; [17]Department of Thoracic and Cardiovascular Surgery, The University of Texas MD Anderson Cancer Center; [18]Department of Translational Molecular Pathology, The University of Texas MD Anderson Cancer Center; [19]Department of Radiation Physics, The University of Texas MD Anderson Cancer Center

Background: Immune checkpoint inhibitors (ICIs) has revolutionized treatment landscape for non-small cell lung cancer patients (NSCLC). Despite remarkable shift in duration of responses seen with ICIs, only 20-30% patients experience such benefit. We aimed to investigate the application of deep learning on chest CT scans to derive an imaging signature of response to ICI and evaluate its added value in the clinical context.

Methods: 976 patients with metastatic, EGFR/ALK wild-type NSCLC treated with ICIs at MD Anderson and Stanford were enrolled from Jan 1, 2014, to Feb 29, 2020. An ensemble deep learning model was built and tested on pretreatment CT scan (Deep-CT) to predict overall survival (OS) and progression-free survival (PFS) following ICIs therapy. We also evaluated the added predictive value of the Deep-CT model in the context of existing clinicopathological and radiological metrics.

Results: Deep-CT demonstrated robust stratification of patient survival on MD Anderson testing set, which was further validated in the external Stanford set. The performance of the Deep-CT model remained significant on subgroup analyses stratified by PD-L1, histology, age, sex, and race. In univariate analysis, Deep-CT outperformed the conventional risk factors, including histology, smoking status, and PD-L1 expression, and remained an independent predictor after multivariate adjustment. Integrating the Deep-CT model with conventional risk factors demonstrated significantly improved prediction performance, with overall survival C- index increases from 0·70 (clinical model) to 0·75 (composite model) during testing. On the other hand, the deep learning risk scores correlated with some radiomics features, but radiomics alone could not reach the performance level of deep learning, indicating that the deep learning model effectively captured additional imaging patterns beyond known radiomics features.

Conclusions: We have developed and validated a deep learning signature associated with OS and PFS in ICI-treated NSCLC patients which appears to be independent of and superior to known clinicopathological risk factors, bringing the goal of precision immunotherapy for patients with NSCLC closer.

Poster #2696 - CoCo-ST: Comparing and Contrasting Spatial Transcriptomics Data Sets Using Graph Contrastive Learning

Muhammad Aminu[1,10], Bo Zhu[2,10], Natalie Vokes[2,10], Hong Chen[2], Lingzhi Hong[2], Jianrong Li[9], Junya Fujimoto[8], Yuqui Yang[12], Tao Wang[12], Bo Wang[13], Alissa Poteete[2], Monique B. Nilsson[2], Xiuning Le[2], Cascone Tina[2], David Jaffray[3,7], Nick Navin[5], Lauren A. Byers[2], Don Gibbons[2], John Heymach[2], Ken Chen[6], Chao Cheng[9], Jianjun Zhang[2,11], Jia Wu[1,2,7,11]*

[1]Department of Imaging Physics, The University of Texas MD Anderson Cancer Center, Houston, TX; [2]Department of Thoracic/Head and Neck Medical Oncology, The University of Texas MD Anderson Cancer Center; [3]Office of the Chief Technology and Digital Officer, The University of Texas MD Anderson Cancer Center; [5]Department of Systems Biology, The University of Texas MD Anderson Cancer Center, Houston, TX; [6]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX; [7]Institute for Data Science in Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX; [8]Clinical Research Center, Hiroshima University, Hiroshima, Japan; [9]Department of Medicine, Institution of Clinical and Translational Research, Baylor College of Medicine, Houston, TX; [12]Department of Public Health, UT Southwestern Medical Center, Dallas, TX; [13]Department of Medical Biophysics, University of Toronto, Ontario
[10]Contributed equally; [11]Co-senior authors; *Corresponding author

Abstract: Traditional feature dimension reduction methods have been widely used to uncover biological patterns or structures within individual spatial transcriptomics data. However, these methods are designed to yield feature representations that emphasize patterns or structures with dominant high variance, such as the normal tissue spatial pattern in a precancer setting. Consequently, they may inadvertently overlook patterns of interest that are potentially masked by these high- variance structures. Herein we present our graph contrastive feature representation method called CoCo-ST (Comparing and Contrasting Spatial Transcriptomics) to overcome this limitation. By incorporating a background data set representing normal tissue, this approach enhances the identification of interesting patterns in a target data set representing precancerous tissue. Simultaneously, it mitigates the influence of dominant common patterns shared by the background and target data sets. This enables discerning biologically relevant features crucial for capturing tissue-specific patterns, a capability we showcased through the analysis of serial mouse precancerous lung tissue samples.

Poster #2699 - Synthetic PET from CT Improves Diagnosis and Prognosis for Lung Cancer

Morteza Salehjahromi[1,†], Tatiana V Karpinets[2,†], Sheeba J. Sujit[1], Mohamed Qayati[1], Pingjun Chen[1], Muhammad Aminu[1], Maliazurina B. Saad[1], Rukhmini Bandyopadhyay[1], Lingzhi Hong[1,9], Ajay Sheshadri[3], Julie Lin[3], Mara B. Antonoff[4], Boris Sepesi[4], Edwin J. Ostrin[5], Iakovos Toumazis[6], Peng Huang[7], Chao Cheng[8], Tina Cascone[9], Natalie Vokes[9], Carmen Behrens[9], Jeffrey H. Siewerdsen[1,15], John D. Hazle[1], Joe Chang[10], Jianhua Zhang[2], Yang Lu[11], Myrna C. B. Godoy[12], Caroline Chung[10,15], David Jaffray[1,15], Ignacio Wistuba[13], J. Jack Lee[14], Ara A. Vaporciyan[4], Don L. Gibbons[9], Gregory Gladish[12], John V. Heymach[9], Carol C. Wu[12, ‡], Jianjun Zhang[2,9,16,17, ‡], Jia Wu[1,9,15,‡]

[1]Department of Imaging Physics, MD Anderson Cancer Center, Houston, TX; [2]Department of Genomic Medicine, MD Anderson Cancer Center; [3]Department of Pulmonary Medicine, MD Anderson Cancer Center; [4]Department of Thoracic and Cardiovascular Surgery, MD Anderson Cancer Center; [5]Department of General Internal Medicine, MD Anderson Cancer Center; [6]Department of Health Services Research, MD Anderson Cancer Center; [7]Department of Oncology, The Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins; [8]Institute for Clinical and Translational Research, Baylor College of Medicine, Houston, TX; [9]Department of Thoracic/Head and Neck Medical Oncology, MD Anderson Cancer Center; [10]Department of Radiation Oncology, MD Anderson Cancer Center; [11]Department of Nuclear Medicine, MD Anderson Cancer Center; [12]Department of Thoracic Imaging, MD Anderson Cancer Center; [13]Department of Translational Molecular Pathology, MD Anderson Cancer Center; [14]Department of Biostatistics, MD Anderson Cancer Center; [15]Institute for Data Science in Oncology, MD Anderson Cancer Center; [16]Lung Cancer Genomics Program, MD Anderson Cancer Center; [17]Lung Cancer Interception Program, MD Anderson Cancer Center
†Contribute equally; ‡Co-senior authors

Abstract: F-Fluorodeoxyglucose positron emission tomography (FDG-PET) and computed tomography (CT) are indispensable components in modern medicine. Although PET can provide additional diagnostic value, it is costly and not universally accessible, particularly in low-income countries. Lately, deep learning has reinvigorated the landscape of medical image synthesis, but CT-to-PET conversion is understudied, especially in lung cancer. To bridge this gap, we have developed a conditional generative adversarial network pipeline that can produce PET from diagnostic CT scans based on multicenter multimodal lung cancer datasets (n = 1478). Importantly, the fidelity of synthetic PET images was validated at imaging, biological, and clinical levels. At the imaging signal level, equivalent imaging quality and tumor contrast between the synthetic PET scans and ground-truth FDG-PET scans have been confirmed by experienced radiologists' adjudication as well as by metrics at both pixel and structure levels. Radiogenomics further confirms the dysregulated cancer hallmark pathways of synthetic PET are consistent to ground-truth PET. We also demonstrate the clinical value of the synthetic PET in improving lung cancer diagnosis, staging, risk prediction,and prognosis. Taken together, this proof-of-concept study testifies the feasibility of applying deep learning to obtain the high-fidelity PET translated from CT.