Temerty Centre for AI Research
and Education in Medicine
UNIVERSITY OF TORONTO

T-CAIREM AI in Medicine Conference

Ideas to Impact

October 12 - 13, 2023

CONFERENCE PROGRAM

T-CAIREM CONFERENCE
AI in MEDICINE

Welcome Note

Welcome to the T-CAIREM AI in medicine conference. We're delighted to host this event and bring together some of the most dynamic speakers and innovative scientists and clinicians in the field today.

The genesis of the Temerty Centre for AI Research and Education in Medicine (T-CAIREM), which launched in 2020, stemmed from the need to bring together Canada's Health AI community. Canada is a world leader in both AI and medicine with leading scientists, clinicians, and research institutes in each of the respective fields – we boast some of the very top medicine, computer science, engineering, and statistics programs in the world. We felt the thoughtful intersection of the fields would be transformative not only to advance innovative research, but change clinical practice and impact patient outcomes in a transformative manner.

Sensing this need, T-CAIREM was established to bring together seemingly disparate fields that now clearly have common goals – advance groundbreaking research and education for massive improvements in health outcomes for our local, national, and global communities. The participants and attendees at this conference are proof of the incredible talent and collaborative spirit that Canada has to offer the world.

Our team hopes that over the next two days that you enjoy yourself, learn some things you didn't know before, and meet potential collaborators. Thanks so much for being part of this conference and for your interest in transforming health through AI.

Best Wishes,

Muhammad Mamdani PharmD, MA, MPH ("he, him")
Vice President - Data Science and Advanced Analytics, Unity Health Toronto
Odette Chair in Advanced Analytics
Faculty Affiliate – Vector Institute
Director - University of Toronto Temerty Centre for Artificial Intelligence Research and Education in Medicine (T-CAIREM)
Professor - University of Toronto

GENERAL INFORMATION

Faculty Disclosure

It is the policy of the University of Toronto, Temerty Faculty of Medicine, Continuing Professional Development to ensure balance, independence, objectivity, and scientific rigor in all its individually accredited or jointly accredited educational programs. All speakers, moderators, facilitators, authors and scientific planning committee members participating in University of Toronto accredited programs, are required to disclose to the program audience any real or apparent conflict(s) of interest that may have a direct bearing on the subject matter of the continuing education program. This pertains but is not limited to relationships within the last FIVE (5) years with for-profit organizations, not-for-profit and public sector sponsors and donors, biomedical device manufacturers, or other corporations whose products or services are related to the subject matter of the presentation. The intent of this policy is not to prevent a speaker with a potential conflict of interest from making a presentation. It is merely intended that any potential conflict of interest should be identified openly so that the listeners may form their own judgements about the presentation with the full disclosure of facts. It remains for the audience to determine whether the speaker's outside interests may reflect a possible bias in either the exposition or the conclusions presented.

Session Polling and Q&A

Visit slido.com and enter the code **AIMED** (not case-sensitive) or scan the QR code below to participate in polling questions and to submit your question during each session. The moderator will review the questions and ask the speaker during Q&A.

Wi-Fi Internet Access

Network: AI in Medicine
Password: tcariem23

Full Conference Program

You can download the digital version of the conference program which includes the submitted abstracts from the link in your conference reminder email or from the conference home page at tcairem-conference.ca

SPEAKERS

Keynote Speakers

Leo Celi
Principal Research Scientist, Massachusetts Institute of Technology

Colleen Flood
Dean of Queen's University Faculty of Law

Alistair Johnson
T-CAIREM Infrastructure Co-lead, Assistant Professor, Department of Biostatistics, University of Toronto

Senthil Nachimuthu
Nightingale Open Science, Center for Applied AI, University of Chicago Booth School of Business

Invited Speakers and Moderators

Mamatha Bhat
T-CAIREM Partnerships & Engagement Lead, Co-Lead of Transplant AI initiative (TAI), Ajmera Transplant Centre, UHN

Bo Wang
Lead Artificial Intelligence Scientist, Peter Munk Cardiac Centre and the Techna Institute, UHN

Amol Verma
Temerty Professor of AI Research and Education in Medicine, University of Toronto

Michaël Chassé
Medical Specialist and Principal Scientist, Centre hospitalier de l'Université de Montréal (CHUM)

Devin Singh
T-CAIREM Research Co-lead, Assistant Professor, Faculty of Medicine and Department of Computer Science, University of Toronto

Benjamin Haibe-Kains
T-CAIREM Infrastructure Co-Lead, Senior Scientist, Princess Margaret Cancer Centre

Anna Goldenberg
T-CAIREM Research Co-Lead, Varma Family Chair in Biomedical Informatics and Artificial Intelligence, Hospital for Sick Children

Alejandro Berlin
Clinician-Scientist, Radiation Oncologist
Princess Margaret Cancer Centre, UHN

Ian Stedman
Assistant Professor, Canadian Public Law and Governance in the School of Public Policy and Administration, York University

Mara Lederman
Co-Founder and COO of Signal 1, Professor of Strategic Management, Rotman School of Management, University of Toronto

Dave Anderson
Senior Instructor, Department of Biochemistry & Molecular Biology, University of Calgary

DAY 1 PROGRAM

Thursday, October 12, 2023

Time	Topic		Room
7:30	On-Site Registration & Breakfast		Foyer
8:30	**Chair's Opening Address + Triva Game** *Muhammad Mamdani*		Ballroom
9:00	**Big Data, Big Bias: The road to hell is paved with good intentions** *Leo Celi* Moderated by *Muhammad Mamdani*		Ballroom
10:00	Coffee and Poster Viewing (P01 - P30)		Foyer
10:30	**Panel Discussion: Emerging Areas in AI Affecting Changes in Care** Moderated by *Mamatha Bhat* Generative AI - *Bo Wang* Risk Prediction Models - *Amol Verma* AI Driving Automation - *Michaël Chassé*		Ballroom
12:00	Lunch and Poster Viewing (P01 - P30)		Foyer
1:00	**Debate: Would You Trust Modern AI to Make Decisions About Your Health?** Moderated by *Devin Singh*		Ballroom
	Pro *Benjamin Haibe-Kains* *Shilpa Raju*	**Con** *Leo Celi* *Maggie Keresteci*	
2:00	Coffee and Poster Viewing (P01 - P30)		Foyer

Time	Topic	Room
2:30	**Trials and Tribulations of Deploying AI** Moderated by *Anna Goldenberg* Academic/Hospital Perspective - *Alejandro Berlin* Ethical Considerations - *Ian Stedman* Start-up Perspective - *Mara Lederman*	Ballroom
3:30	**The Regulating of AI in Healthcare** *Colleen Flood* Moderated by *Muhammad Mamdani*	Ballroom
4:30	**Closing Remarks** *Muhammad Mamdani*	Ballroom
5:00	Networking Reception and Poster Viewing (P01 - P30)	Foyer

Friday, October 13, 2023

Time	Topic	Location
7:30	On-Site Registration & Breakfast	Foyer
8:30	**Chair's Opening Address & Recap of Day 1** *Muhammad Mamdani*	Ballroom
8:45	**Health Data Nexus** *Alistair Johnson* **Democratizing Medical Data and Computing for Research and Education – Nightingale Open Science** *Senthil Nachimuthu* Moderated by *Muhammad Mamdani*	Ballroom
10:00	Coffee and Poster Viewing (P31 - P60)	Foyer
10:30	**Adjudicated 10-Min Oral Abstract Presentations**	
	Paediatrics & Critical Care Abstracts: OR01, OR02, OR03, OR04, OR05, OR06 Moderated by *Gemma Posthill*	Ballroom
	Oncology, Medical Imaging & Genomics Abstracts: OR07, OR08, OR09, OR10, OR11, OR12 Moderated by *Abhishek Moturu*	Kingsway
	Mental Health & Gastroenterology Abstracts: OR13, OR14, OR15, OR16, OR17, OR18 Moderated by *Sujay Nagaraj*	Caledon
	Signal Processing, Core Machine Learning & Deployment Abstracts: OR19, OR20, OR21, OR22, OR23, OR24 Moderated by *Jethro Kwong*	Oakville
12:00	Lunch and Poster Viewing (P31 - P60)	Foyer

Time	Topic	Location
1:00	**Adjudicated 3-Min Oral Abstract Presentations**	
	Paediatrics, Neurology & Genomics Abstracts: OR25, OR26, OR27, OR28, OR29, OR30, OR31, OR32, OR33, OR34 Moderated by *Gemma Posthill*	Ballroom
	Cardiology, Vision Learning, Medical Imaging & Deployment Abstracts: OR35, OR36, OR37, OR38, OR39, OR40, OR41, OR42, OR43 Moderated by *Konrad Samsel*	Kingsway
	Oncology, Core Machine Learning, Large Language Models & AI Education Abstracts: OR45, OR46, OR47, OR48, OR49, OR50, OR51, OR52, OR53, OR54 Moderated by *Sujay Nagaraj*	Caledon
	Critical Care & Natural Language Processing Abstracts: OR55, OR56, OR57, OR58, OR59, OR60, OR61, OR62, OR63, OR64 Moderated by *Jethro Kwong*	Oakville
	AI in Medicine Education Session *Andrew Austin, Michael Balas, Pamela Molina, Tiam Feridooni* Moderated by *David Anderson*	Ballroom
3:00	Coffee, and Poster Viewing (P31 - P60)	Foyer
3:30	**Shark Tank Pitch Competition**	Ballroom
4:30	**Closing Remarks and Awards** *Muhammad Mamdani*	Ballroom
5:00	**Conference Adjourns**	

Friday, October 13 - 10:30 AM-12:00 PM

Abstract #	Abstract Title	Location
OR01	Artificial intelligence-based decision support predicts requirement for neurosurgical intervention in acute traumatic brain injury *Christopher Smith*	Ballroom
OR02	Deep Learning for prediction of Neurodevelopmental Impairment in Preterm Babies from Cranial Ultrasound exam *Alessandro Guida*	Ballroom
OR03	Improving Pediatric Low-Grade Neuroepithelial Tumors Molecular Subtype Identification Using a Novel AUC Loss Function for Convolutional Neural Networks *Khashayar Namdar*	Ballroom
OR04	Machine Learning for the Prediction of Massive Transfusion in Trauma *Anton Nikouline*	Ballroom
OR05	NLP and Machine Learning Pipeline to Automate Extraction of Clinical Injury Data *Alper Celik, Nicholas Singh*	Ballroom
OR06	The State of Artificial Intelligence in Pediatric Surgery: A Systematic Review *Mo Elahmedi*	Ballroom
OR07	A Machine Learning Approach to Processing and Interpreting Ex Vivo Lung Radiographs Predicts Transplant Outcomes *Bonnie T. Chao*	Kingsway
OR08	Biological Pattern Discovery in Glioma Stem Cell Spatial Organization using Computer Vision and Transcriptomics *Shamini Ayyadhury*	Kingsway
OR09	Designing a Scalable Pipeline for ML Ops to Expedite AI Research and Deployment at Princess Margaret Cancer Centre *Benjamin Grant*	Kingsway
OR10	Exploring deep-learning to accurately classify breast tissue morphology using Wide-field OCT. *Ali Yassine, Yanir Levy*	Kingsway

ORAL PRESENTATION INDEX

Abstract #	Abstract Title	Location
OR11	Identifying nucleosome positioning features based on Deep Residual Networks *Yosef Masoudi-Sobhanzadeh*	Kingsway
OR12	Zero-Shot Medical Image Captioning with Frozen Vision Transformers and Large Language Models *David Li*	Kingsway
OR13	A Motivational-Interviewing Chatbot with Generative Reflections for Increasing Readiness to Quit Among Smokers *Jonathan Rose*	Caledon
OR14	Biopsychosocial Characterization of Cognitive Decline and Late-Life Depression Trajectories Using Bayesian Consensus Clustering and Machine Learning *Mu Yang*	Caledon
OR15	Day-to-day variability in activity levels using wearable devices detects transitions to depressive episodes prior to changes in mood: analysis of densely-sampled data from a contactless longitudinal study *Abigail Ortiz*	Caledon
OR16	Detecting and Analyzing Potential Comorbid ADHD in People Reporting Anxiety Symptoms from Social Media Data Using Transformers *Michael Guerzhoy*	Caledon
OR17	Predictive Care: The False Promise of Fair AI Models in Acute Psychiatry *Laura Sikstrom*	Caledon
OR18	Serum Metabolomic Pathways in Predicting future onset of Crohn's Disease *Mingyue Xue*	Caledon
OR19	A computational approach to breath-by-breath ventilator waveform data extraction and analysis during ex vivo lung perfusion enables enhanced physiological lung assessment *Xuanzi Zhou*	Oakville

ORAL PRESENTATION INDEX

Abstract #	Abstract Title	Location
OR20	Automated Prognostication using Deep Learning Applied to Chest X-Rays of Patients with Suspected Pneumonia Presenting to the Emergency Department: A Prospective Shadow Deployment Study *Eduardo P. R. P. Almeida*	Oakville
OR21	ChatGPT and Retinal Disease: A Cross-Sectional Study on AI Comprehension of Clinical Guidelines *Michael Balas*	Oakville
OR22	Leveraging patient's longitudinal data to predict One-year Mortality Risk *Hakima Laribi*	Oakville
OR23	Measures of Overnight Oxygen Saturation can Characterize Sleep Apnea Severity and Predict Postoperative Respiratory Depression *Atousa Assadi*	Oakville
OR24	Measuring Respiratory Mechanics with Esophageal Catheter and Oscillometry *Shaghayegh Chavoshian*	Oakville

Friday, October 13 - 1:00 PM-2:00 PM

Abstract #	Abstract Title	Location
OR25	An artificial intelligence-driven magnetic resonance imaging synthesis framework *Timur Latypov*	Ballroom
OR26	Application of Machine Learning for Clinical Decision Support in the Treatment of Newly Diagnosed Pediatric Crohn Disease Patients *Ricardo Gabriel*	Ballroom
OR27	Can a general large language model augment clinical decision-making in pediatrics? *Esli Osmanlliu*	Ballroom
OR28	Clinical Features, Non-Contrast CT Radiomic and Radiological Signs in Models for the Prediction of Hematoma Expansion in Intracerebral Hemorrhage *Frank Chen*	Ballroom
OR29	Development of a Multi-modal Machine Learning-Based Prognostication Model for Traumatic Brain Injury Using Clinical Data and Computed Tomography Scans *Pascal Tyrrell*	Ballroom
OR30	Exploratory Analysis of Perfusion Index as a Screening Tool for Continuous Monitoring of Blood Pressure in Critically Ill Children *Mana Shahriari*	Ballroom
OR31	Machine learning enables detection of Li-Fraumeni Syndrome using tumor whole-genome sequencing *Brianne Laverty*	Ballroom
OR32	Pixels and Perspectives: Exploring Public Perceptions about AI-Generated Images of Children with Medical Conditions *Muhammed Mukadam*	Ballroom
OR33	Prediction of Emergency Department Readmission among Child and Youth Mental Health Outpatients Using Deep Learning Techniques. *Simran Saggu*	Ballroom
OR34	Relationship between Air Pollution and Crohn's disease (CD) Risk and their relation to Biomarkers of CD risk *Jingcheng Shao*	Ballroom

ORAL PRESENTATION INDEX

Abstract #	Abstract Title	Location
OR35	A Surgical Robot Simulation Framework for Reinforcement Learning to Automate Manipulation and Cutting Subtasks *Radian Gondokaryono*	Kingsway
OR36	Classifier to predict drug cardiac activity using human stem cell-derived cardiac tissues *Julia Plakhotnik*	Kingsway
OR37	Detecting Arrhythmogenic Right Ventricular Cardiomyopathy (ARVC) using Residual Neural Networks *Chris J. George, Sophie Sigfstead*	Kingsway
OR38	Development and evaluation of a live birth prediction model for evaluating human blastocysts *Hang Liu*	Kingsway
OR39	Echocardiogram Quality Enhancement with Vector Quantized Generative Adversarial Networks (VQ-GANs) *Alif Munim*	Kingsway
OR40	Improving Mortality Prediction in People with Cardiovascular Disease: A Random Survival Forests Approach Integrating Frailty Assessment *Jack Quach*	Kingsway
OR41	Novel Approaches in 12-Lead Electrocardiogram Signal Reconstruction *Yan Zhu*	Kingsway
OR42	Obstacle detection for persons with visual impairments using AI-powered smart glasses *Haining Tan*	Kingsway
OR43	Quality of Interaction Between Clinicians and Artificial Intelligence. A Systematic Review *Argyrios Perivolaris*	Kingsway
OR45	A User-Centered Design Approach to an Artificial Intelligence-Enabled Electronic Medical Record in Canadian Primary Care *Krizia Francisco, Puneet Seth*	Caledon
OR46	Assessing Prognosticators of Intracranial Metastatic Disease in Patients With HER2+ Breast Cancer Leveraging Supervised Machine Learning Algorithms *Marco V. Istasy*	Caledon

ORAL PRESENTATION INDEX

Abstract #	Abstract Title	Location
OR47	Bio-inspired modulation to enhance the robustness of deep learning models against healthcare data quality problems *Mohamed Abdelhack*	Caledon
OR48	Deep Learning-Enabled Fluorescence Quantification for Surgical Guidance: Benefits Over Analytical Methods *Anjolaoluwa Adewale*	Caledon
OR49	Development of an objective framework to optimize machine learning-based single-cell segmentation accuracy for multiplexed tissue cytometry *Trevor D. McKee, Mark Zaidi*	Caledon
OR50	Improving Patch-based Segmentation for Pediatric Cancer Detection *Abhishek Moturu*	Caledon
OR51	Innovative Detection of Diabetes Stigma in Digital Spaces with Large Language Models *Somayeh Amini, Mark Dayomi*	Caledon
OR52	Multi-task Machine Learning of the Electronic Medical Record Predicts Future Symptoms among Cancer Patients *Baijiang Yuan*	Caledon
OR53	Rapid Evolution of Artificial Intelligence in Medicine: Comparative Analysis of ChatGPT-3.5 and ChatGPT-4 in Generating Clinician-level Vascular Surgery Recommendations *Arshia Javidan*	Caledon
OR54	Temporal validation of SEPERA to inform nerve-sparing strategy during radical prostatectomy and comparison against expert surgeons *Lauren Pickel*	Caledon
OR55	Addressing class imbalance with data augmentation when training deep learning models to identify high quality clinical studies *Cynthia Lokker*	Oakville
OR56	Applications of Artificial Intelligence to Improve Patient Experiences During Care Transitions: An Informal Review *Nicole Bodnariuc*	Oakville
OR57	Associations between Sex, Race, and Sedation in Invasively Ventilated Patients *Sarah Walker*	Oakville

ORAL PRESENTATION INDEX

Abstract #	Abstract Title	Location
OR58	Challenges in Expert Labeling of Data to Leverage Machine Learning to Support Physiotherapy in the ICU *Adriana Ieraci*	Oakville
OR59	Clinical Outcome Prediction: Evaluating Quantization of Large Language Models *Raj Krishnan Vijayaraj*	Oakville
OR60	Evaluating the Efficacy of Transformer Networks for Audio Signal Classification in Dysphagia Detection *Hamza Mahdi, Eptehal Nashnoush*	Oakville
OR61	Patient and Stakeholder Engagement (PSE) in the Integration of Large Language Models (LLMs) in Healthcare Chatbots *Nikhil Jaiswal*	Oakville
OR62	Prediction of Trauma Bay Disposition Using Explainable Machine Learning *Seong Park*	Oakville
OR63	The use of large language models for therapeutic recommendations *Jean Marie Tshimula*	Oakville
OR64	TxT -Toronto-Technion Treatment Curator *Melanie Courtot, Michael Fralick*	Oakville

Presenting on Thursday, October 12

P#	Abstract Title
P01	3D Pose Estimation Using RGB-D Data for Rehabilitation *Gloria-Edith Boudreault-Morales*
P02	A Novel AI Clinician: Training an Intelligent System to Predict COVID-19 Pneumonia Hospital Outcomes *George Chen*
P03	Accelerating Personalized Breast Cancer Treatment: Validating an AI-Driven Quantitative Ultrasound Model for Predicting Neoadjuvant Chemotherapy Response *Matthew Shammas-Toma*
P04	An AI-assisted Chatbot for Patient Education and Care in Radiotherapy *James C. L. Chow*
P05	Applying mechanistic in-silico EEG biomarkers for improved diagnosis in depression *Frank Mazza*
P06	Beyond Hand-Crafted Features For Pretherapeutic Molecular Status Identification Of Pediatric Low-Grade Neuroepithelial Tumors *Kareem Kudus*
P07	Deep Learning Architectures for 3D Reconstruction of Oral Cancer Models from Intraoperative Spatial-Frequency Fluorescence Images *Natalie J. Won*
P08	Deep Learning-Enabled 3D Fluorescence Imaging for Surgery: A Simulation Study in the Second Near-Infrared Window *Jerry Wan*
P09	Designing a healthbot for varenicline adherence using healthcare provider perspectives *Mackenzie V. Earle*
P10	Detection and recognition of hand impairment level in stroke survivors using egocentric video of activities of daily living *Anne Mei*
P11	Evaluating the Effectiveness of Early Survival Prediction for Coronary Artery Disease Patients: Pre-Treatment vs. Combined Pre and Post Treatment Features *Anita Khalafbeigi*

POSTER INDEX

P#	Abstract Title
P13	Evaluation of Synthetic Data Augmentation for Mitigating Covariate Bias in Real World Health Data *Lamin Juwara*
P14	Identifying Trusted and Ambiguous Regions in Neural Network Predictions: High-Fidelity AI For Image Pathology *Kenneth Wenger*
P15	Iterative XAI Frameworks for Oncology Decision Making: Integrating Expert Feedback to Enhance Cancer Diagnosis *Shermineh Ghasemi*
P16	Machine Learning for Assessment of Capsulorhexis Performance in Cataract Surgery *Jonathan ZL. Zhao*
P17	Multi-Document Summarization of Patient Neurovascular Radiology Reports *Heet Sheth*
P18	Predicting self and care staff's evaluation of resident's rehabilitation potential in post-acute care setting using machine learning algorithms *Bonaventure A. Egbujie*
P19	Prediction Algorithms Driving the Making of the National Early Warning System Two Plus Capacity of the Electronic Casualty Card System in Any Disaster Situations – A Component of Vimy Multi-System *Mariem Abid, Stephane Bourassa*
P20	Prediction of Depression Relapse in Youth and Adolescence from Ecological Momentary Assessment (EMA) data through Machine Learning *Cheuk Hei Chung*
P21	Robust Detection of Seizure Onset Zone in Patients with Epilepsy using a Novel Graph-based Neural Network *Milad Lankarany*
P22	Sex-Based Differences in Speech Coherence for classifying Schizophrenia using BERT similarity *Alban Voppel*

POSTER INDEX

P#	Abstract Title
P23	Smoking Cessation Interventions in South Asian Region: a systematic scoping review *Rameesha Rehmani*
P24	Summarizing Clinical Trials with Large Language Models *Bryant Lim*
P25	Supervised Machine Learning Pipeline to Classify Pain using sEMG and MMG during Neuromuscular Electrical Stimulation to Combat Intensive Care Unit Acquired Weakness *Meg Sharma*
P26	Using Artificial Intelligence To Label Free-Text Operative And Ultrasound Reports For Grading Pediatric Appendicitis *Waseem Abu-Ashour, Dan Poenaru*
P28	Utilization of unsupervised image feature-based clustering to scale classifier design in histopathology *Minli Chen*
P30	Whole-Person Biopsychosocial Subtyping of Mental Illnesses in Treatment-Seeking Youth *Denise Sabac*

Presenting on Friday, October 13

P#	Abstract Title
P31	A Comprehensive Study of Radiomics-based Machine Learning for Liver Fibrosis Detection in CT Images *Jay Yoo*
P32	An Automatic Retinal Image Analysis Method to Estimate Comprehensive Glaucomatous OCT Parameters Using Two-dimensional Images *Chuying Shi*
P33	An integrated toolkit for measuring fairness of risk predictive models for healthcare *Shaina Raza, Amrit Krishnan*
P34	Artificial intelligence-enabled cough detection for monitoring pulmonary tuberculosis treatment response in Madagascar: a preliminary report *Alexandra Zimmer*
P35	Assessing Hand Function in Spinal Cord Injury Patients: A Personalized Egocentric Video-Based Hand Analysis Approach *Mehdy Dousty*
P36	Bridge2AI-Voice as a Biomarker of Health: Building an ethically sourced, bio-acoustic database to understand diseases like never before (Bridge2AI-Voice Consortium) *Jordan Lerner-Ellis, Lochana Jayachandran*
P37	Capacity of Language Learning Models to Generate Medical Residency and Undergraduate Medicine Progress Test Questions *Ryan S. Huang*
P38	Clinician and health system leader perspectives on the use of AI for deriving social determinants of health data in primary care settings *Stephanie Garies*
P39	Delineate cell-cell communication (CCC) in anti-cancer drug resistance by deep learning based multi-modal single-cell methods *Fatema Zohora*
P40	Exploring Canadian Sentiments on COVID-19 Vaccination: A Twitter-based Analysis *Hassan Maleki Golandouz*

POSTER INDEX

P#	Abstract Title
P41	Exploring patient perspectives on how they can and should be engaged in the development of artificial intelligence (AI) applications in health care *Samira Adus*
P42	Identifying Individualized Neurophysiological Causal Features for Working Memory Performance: Implications for Non-Invasive Brain Stimulation *Mina Mirjalili*
P43	Identifying, Characterizing and Tracking Suspicious Skin Moles *Mahla Abdolahnejad, Rim Mhedbi*
P44	Investigating model failures by patient (profiles) for safer clinical deployment *Olivier Lefebvre*
P45	Investigating the relationships between social media discourse and ICU bed demand to inform healthcare supply-chain decisions: COVID-19, Twitter and Causal Analysis *Mahakprit Kaur*
P46	Latent Variable Energy Based Model with Self-Supervised Approaches for Cancer Grading Problem *Kayvan Tirdad*
P47	Off-Label Drug Use during the COVID-19 Pandemic in Africa: Topic modeling and sentiment analysis of Ivermectin in South Africa and Nigeria as a case study *Zahra Movahedi Nia*
P48	On the factuality of Large language model-generated summaries of clinical abstracts *Wael Abdelkader*
P49	Predicting Pre-eclampsia in Pregnant Women: An ML-based Approach using the Lavndr App *Shveta Bhasker*
P51	Solving Healthcare's Last Mile Problem *Karim Keshavjee*
P52	Stability-Based Biomarker Development to Identify Pathological Brain Areas Responsible for Freezing of Gait in Parkinson's Disease *Nooshin Bahador*
P53	SurvdigitizeR: R Package to Automate the Digitization of Published Kaplan-Meier Curves *Jasper Zhongyuan Zhang, Qiyue Zhang*

POSTER INDEX

P#	Abstract Title
P54	Time Motion-Study for Artificial Intelligence Automation to Improve Family Medicine Workflow: Protocol for a Mixed Methods Study *Karen Li*
P55	Understanding Patterns in Variants of Uncertain Significance to Facilitate Reclassification Using Machine-learning Based Variant Effect Predictors *Cindy Zhang*
P56	Unveiling Potential Hidden Bias in Automated Lateral Spine Image Interpretation: Predicting Demographic and Anthropometric Characteristics using Convolutional Neural Networks *Barret A. Monchka*
P57	Using Speech Features in a Random Forest Machine Learning Model to Predict COPD Symptoms *Sashini Kosgoda*
P58	Utilizing image denoising and machine learning segmentation to quantify fluid volume in eyes with vascular retinal diseases: the STATIC study *Niveditha Pattathil*
P59	Evaluating the efficacy of an automated, voice-based swallowing dysfunction screening tool utilizing convolutional neural networks *Rami Saab*
P60	Using the Wizard of Oz methodology to build a healthbot to improve medication adherence for smoking cessation *Kamna Mehra*

[OR01] Artificial intelligence-based decision support predicts requirement for neurosurgical intervention in acute traumatic brain injury

Christopher Smith, Division of Neurosurgery, St Michael's Hospital
Armaan Malhotra, University of Toronto
Christopher Hammill, St Michael's Hospital
Derek Beaton, University of Toronto
Alun Ackery, Division of Emergency Medicine, St Michael's Hospital
Muhammad Mamdani, University of Toronto
Jefferson Wilson, Division of Neurosurgery, St Michael's Hospital
Robert Moreland, Division of Medical Imaging, St Michael's Hospital
Errol Colak, Division of Medical Imaging, St Michael's Hospital
Christopher Witiw, Division of Neurosurgery, St Michael's Hospital

Introduction

Artificial intelligence model integration into clinical workflow offers potential to optimize decision-support for transfer of acute traumatic brain injury (TBI) patients to appropriate tertiary care facilities. We aimed to develop an automated triage tool to predict neurosurgical intervention for brain-injured patients without any image-level labels.

Methods

We utilized a provincial trauma registry to identify TBI patients from 2005 to 2022 treated at a specialized Canadian trauma center. Model training, validation, and testing was performed using head CT scans with binary ground truth patient-level labels corresponding to whether the patient received neurosurgical intervention obtained retrospectively. The finalized model, termed Automated Surgical Intervention Support Tool for TBI (ASIST-TBI), was deployed in a simulated prospective fashion on all TBI patients presenting to our center between March 2021 and September 2022. An image-based vision transformer architecture demonstrated optimal performance metrics compared to convolutional neural network, principal component analysis logistic regression, and tabular clinical variable models.

Results

A dataset of 2,806 trauma patient head CT scans acquired between 2005-2021 were utilized for training, validation, and testing. In simulated prospective deployment, an additional 612 consecutive scans were used to assess the performance of ASIST-TBI. There was accurate prediction of neurosurgical intervention with an area under the receiver operating curve (AUC) of 0·92, accuracy of 0·87, sensitivity of 0·87, and specificity of 0·88 on the test dataset. Simulated prospective deployment resulted in an AUC of 0·89, 0·85 sensitivity, 0·84 specificity, and 0·84 accuracy. No sex or age specific gender bias was encountered between incorrect and correct classifications.

Discussion/Conclusion

We developed a novel deep learning model that accurately predicts requirement for neurosurgical intervention using acute TBI CT scan input. ASIST-TBI has potential application to optimize inter-facility triage efficiency and care pathways for brain-injured patients.

[OR02] Deep Learning for prediction of Neurodevelopmental Impairment in Preterm Babies from Cranial Ultrasound exam

Alessandro Guida , Nova Scotia Health Authority
Noah Barrett , Dalhousie University
Xiang Jiang , Dalhousie University
Samuel Stewart , Dalhousie University
Jeff Kowalski , Nova Scotia Health Authority
Stan Matwin, Dalhousie University
Michael Vincer, IWK Health
Jehier Afifi , IWK Health
Tahani Ahmad , IWK Health

Introduction

Cranial ultrasound (CUS) is widely used as a screening tool to assess brain injury in preterm infants. Severe abnormalities on CUS are known to be associated with Neurodevelopmental impairment (NDI) while, for less severe changes or apparently “normal” scans the prognostic outcome is variable. This study has two aims:

• To train a deep learning prognostic model based on CUS to early assess NDI outcomes.

• To train a diagnostic model that can identify abnormal US images and aid radiologist assessment as a computer-aided diagnosis tool.

Methods

This is a retrospective study of a cohort of very preterm infants born between 2004 and 2016 and admitted to the Neonatal Intensive Care Unit at IWK Health. The sample size includes images collected from 514 patients (4626 images) at 3-time points during the neonatal period. 3 coronal images were selected at each time point. The NDI outcome (i.e. patient outcome) and CUS diagnostic outcome (i.e. timepoint outcome) were imported based on the clinical assessment. Further image annotation was performed by a radiologist who binary labelled each single US image based on the presence or absence of visual abnormalities in the image (Image outcome label). We trained the model using different deep learning approaches: from simple classic CNN to more advanced strategies (e.g. multi-view CNNs, to multiple instance learning).

Results

We tested different CNN architectures and strategies; results for the prognostic model indicate an average of 0.72 AUC. The diagnostic model performs generally better, with an average AUC of 0.83. Model performance is greatly improved (AUC > 0.9) with the addition of model uncertainty estimation.

Discussion/Conclusion

During the preparation of this study, many challenges came to our attention and we deployed different strategies to address each of them learning valuable lessons.

[OR03] Improving Pediatric Low-Grade Neuroepithelial Tumors Molecular Subtype Identification Using a Novel AUC Loss Function for Convolutional Neural Networks

Khashayar Namdar, University of Toronto
Matthias Wagner, The Hospital for Sick Children
Cynthia Hawkins, The Hospital for Sick Children
Uri Tabori, The Hospital for Sick Children
Birgit Ertl-Wagner, The Hospital for Sick Children
Farzad Khalvati, The Hospital for Sick Children

Introduction

Pediatric Low-Grade Neuroepithelial Tumor (PLGNT) is the most common type of brain tumor in children, and Convolutional Neural Networks (CNNs) have been shown to be effective for classifying BRAF fusion and BRAF V600E mutation PLGNT molecular subtypes on MR images. Area Under Receiver Operating Characteristic (ROC) Curve (AUC) is a popular metric for evaluating CNNs, which is not differentiable. Thus, the models cannot be directly optimized for AUC. Here we introduce a novel AUC loss function to improve the performance of PLGNT molecular subtype classifiers.

Methods

Inspired by Wilcoxon-Mann-Whitney (WMW) loss function, we developed an AUC loss function in PyTorch, which will be open-sourced in GenuineAI Python library. We validated our loss function to improve the performance of CNN classifiers for PLGNT molecular subtype identification. Our REB-approved retrospectively acquired study cohort consisted of 143 BRAF fusion and 71 BRAF V600E mutation children with PLGNT. As the input to the CNNs, we used MRI Fluid-Attenuated Inversion Recovery (FLAIR) sequence and the tumor segmentations provided by a pediatric neuroradiology fellow and verified by a senior pediatric neuroradiologist.

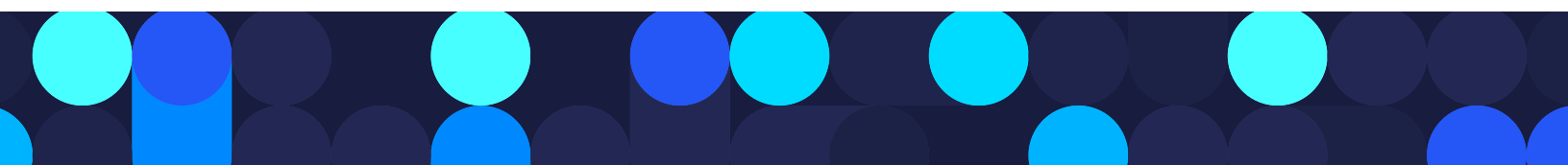
We used a shallow CNN architecture and repeated the train-validation-test (60/20/20) experiment with different data splits and model random initializations 100 times. To train the baseline model, we use binary cross entropy (BCE). Learning rate (0.1), optimizer (stochastic gradient descent), and the maximum number of epochs (40) were the same for baseline and the main models.

Results

We achieved a test AUC of 86.11 with 95% Confidence Interval (CI) [84.96, 87.25] for BRAF fusion vs BRAF V600E mutation binary classification with BCE loss function. AUC loss improved the mean AUC to 87.71 CI [86.64, 88.79] (p-value 0.0456).

Discussion/Conclusion

Directly optimizing CNNs for AUC results in statistically significantly improved performance of CNN classifiers for PLGNT molecular subtype identification.



[OR04] Machine Learning for the Prediction of Massive Transfusion in Trauma

Anton Nikouline , London Health Sciences Centre

Brodie Nolan, St. Michael's Hospital

Jinyue Feng, University of Toronto

Frank Rudzicz, Faculty of Computer Science, Dalhousie University

Avery Nathens, Sunnybrook Health Sciences Centre

Introduction

Several prediction tools have been developed for the initiation of massive transfusion (MT) protocols. Despite 11 validated prediction scores, they are limited in providing good predictability with early clinical data. We developed a MT prediction model using machine learning and early clinical data from the National Trauma Data Bank (NTDB).

Methods

From 2013 to 2018, all patients (age greater than or equal to 16) presenting to level I or II trauma centers with an Injury Severity Score (ISS) greater than 12 from the NTDB were included for analysis. Interfacility transfers, burns, isolated head and hip injuries were excluded. MT was defined as 5 units of blood product within 4 hours or 10 units within 24 hours of arrival. Patient variables were split into two groups- pre-hospital and emergency department data (ALL) or only pre-hospital (EMS). Six machine learning models were used to predict MT and their performances compared.

Results

A total of 326,758 patients were included with 18,871 (5.8%) meeting MT criteria. Patients requiring MT had a median ISS of 29 [IQR 22-41], median number of blood products within 4 hours of 12 [7-25] and within 24 hours of 14 [8-30]. Systolic blood pressure, time in emergency department, heart rate, Glasgow Coma Scale and temperature were the most important variables in predicting MT. Extreme Gradient Boost modelling demonstrated the best performance, while all machine learning models outperformed logistic regression. The model's performance using the EMS and ALL dataset, respectively, demonstrated an area under the receiver operating curve of 0.76 and 0.83, specificity of 78% and 84%, sensitivity of 74% and 83%, positive predictive value of 17% and 23% negative predictive value of 98% and 99%, and accuracy of 78% and 83%.

Discussion/Conclusion

Machine learning models are able to accurately predict massive transfusion in trauma using early clinical data.

[OR05] NLP and Machine Learning Pipeline to Automate Extraction of Clinical Injury Data

Evangeline Zhang, Hospital for Sick Children
Alper Celik, Hospital for Sick Children
Nicholas Singh, Hospital for Sick Children
Kevin Yao, Hospital for Sick Children
Bailey Ng, Hospital for Sick Children
Alyssia Naran, Hospital for Sick Children
Daniel Rosenfield, Hospital for Sick Children
Deborah Taylor, Hospital for Sick Children
Devin Singh, Hospital for Sick Children (SickKids)

Introduction

The Canadian Hospitals Injury Reporting and Prevention Program (CHIRPP) is an injury and poisoning surveillance system under the Public Health Agency of Canada (PHAC) that collects and analyzes data on injuries, poisonings, and mental health cases from Emergency Departments (ED) across Canada. CHIRPP data is utilized to recognize injury risk and trends, increase public awareness, and establish government policies to improve safety and quality of life. However, the current manual process of extracting injury information from ED cases is time-consuming which has resulted in over 18 months of unreported data backlog. We aim to improve this process using machine learning (ML) and natural language processing (NLP).

Methods

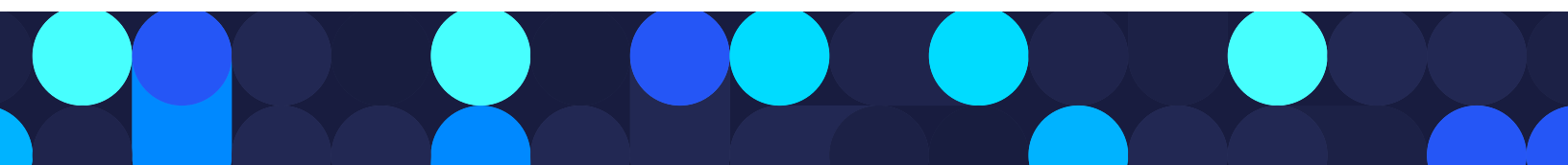
We developed a pipeline that processes electronic medical records (EMRs) from SickKids ED. This pipeline consists of pre-processing scripts to remove irrelevant sections, CHIRPP case detection, summarization and post-processing for human review. For pre-processing, we created a rule-based method for automating the identification of relevant sections within the EMR. These sections are then classified as either CHIRPP positive (i.e. needs reporting to PHAC) or CHIRPP negative, using a large language model (LLM) fine-tuned on historical data of CHIRPP cases (bert-large-uncased). 119,480 unique patient notes were used for training and 62,488 for testing (held out test-set). EMR notes for positive cases are then automatically summarized for reporting using another LLM (t5-small).

Results

Our classification model shows 94% precision and 87% recall with an f1 score of 95%. Our summarization methods show promising results with encouraging cosine distances compared to human summaries. In our first iteration, we achieved a 50% reduction in time-savings and aim to increase it to 70%.

Discussion/Conclusion

Our ultimate objective is to fully automate CHIRPP data entry, eliminating the need for manual inputting, reducing backlog, and enabling near real-time injury trend reporting.



[OR06] The State of Artificial Intelligence in Pediatric Surgery: A Systematic Review

Mo Elahmedi , Harvey E. Beardmore Division of Pediatric Surgery, The Montreal Children's Hospital, McGill University Health Centre

Riya Sawhney, Harvey E. Beardmore Division of Pediatric Surgery, The Montreal Children's Hospital, McGill University Health Centre, Montreal, Quebec, Canada

Fabio Botelho Mendes, Harvey E. Beardmore Division of Pediatric Surgery, The Montreal Children's Hospital, McGill University Health Centre

Elena Guadagno, Harvey E. Beardmore Division of Pediatric Surgery, The Montreal Children's Hospital, McGill University Health Centre

Dan Poenaru, Harvey E. Beardmore Division of Pediatric Surgery, The Montreal Children's Hospital, McGill University Health Centre

Introduction

Artificial intelligence (AI) has been recently shown to improve clinical workflows and outcomes - yet its potential in pediatric surgery remains largely unexplored. This systematic review details the use of AI in pediatric surgery.

Methods

Nine databases were searched from inception until January 2023, identifying articles focused on AI in pediatric surgery. Two authors reviewed full texts of eligible articles. Studies were included if they were original investigations on the development, validation, or clinical application of AI models for pediatric conditions primarily managed surgically. Studies were excluded if they were not peer-reviewed, were review articles, editorials, commentaries, or case reports, or did not employ at least one AI model. Data included study characteristics, clinical specialty, purpose, role, AI technique, performance metrics, key results, interpretability, validation, and bias. Missing performance metrics were imputed using a random forest algorithm.

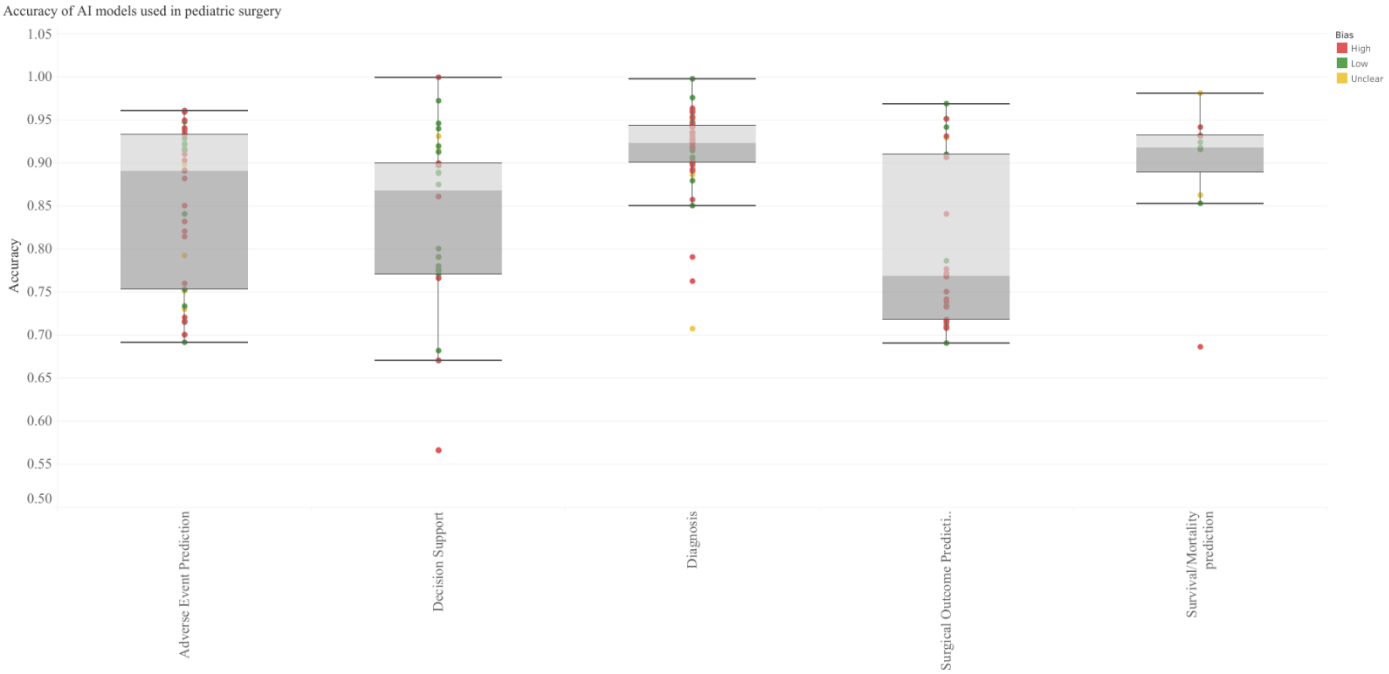
Results

Authors screened 8,178 articles and included 112 studies that reported on 155 models trained on data from 430,654 children and adolescents. Half of the studies (50%) reported predictive models (for adverse events [25%], surgical outcomes [16%] and survival [9%]), followed by diagnostic (29%) and decision support models (21%). Neural networks (44%) and ensemble learners (36%) were the most commonly used AI methods across application domains. The main pediatric surgical subspecialties represented across all models were general surgery (31%) and neurosurgery (25%). Overall mean accuracy was 0.86 ± 0.10. Forty-four percent of models were interpretable, and 6% were both interpretable and externally validated. Forty percent of models had a high risk of bias, and concerns over applicability were identified in 7%.

Discussion/Conclusion

AI has wide clinical applications in pediatric surgery. However, most AI models are not externally validated and lack interpretability. The current research prerogative includes prospective algorithm validation and integration in the clinical pipeline.

Supporting information


Accuracy of AI models used in pediatric surgery
Bias
High
Low
Unclear
Accuracy
1.05
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
Adverse Event Prediction
Decision Support
Diagnosis
Surgical Outcome Predicti..
Survival/Mortality prediction

[OR07] A Machine Learning Approach to Processing and Interpreting Ex Vivo Lung Radiographs Predicts Transplant Outcomes

Bonnie T. Chao, Latner Thoracic Research Laboratories, Toronto General Hospital Research Institute, University Health Network
Jun Ma, Department of Laboratory Medicine and Pathobiology, University of Toronto
Xuanzi Zhou, Latner Thoracic Research Laboratories, Toronto General Hospital Research Institute, University Health Network
Micheal C. McInnis, Division of Cardiothoracic Imaging, Joint Department of Medical Imaging, Toronto General Hospital, University Health Network
Jonathan C. Yeung, Latner Thoracic Surgery Research Laboratories,University Health Network
Marcelo Cypel, Latner Thoracic Research Laboratories, Toronto General Hospital Research Institute, University Health Network
Mingyao Liu, Institute of Medical Science, Temerty Faculty of Medicine, University of Toronto
Bo Wang, Department of Laboratory Medicine and Pathobiology, University of Toronto
Andrew T. Sage, Institute of Medical Science, Temerty Faculty of Medicine, University of Toronto
Shaf Keshavjee, Latner Thoracic Research Laboratories, Toronto General Hospital Research Institute, University Health Network

Introduction

Ex vivo lung perfusion (EVLP) is an advanced clinical platform for isolated donor lung assessment and treatment. Radiographs acquired during EVLP show a pristine view of donor lungs without confounding obstructions from the chest wall and heart. However, systematic and in-depth analysis of ex vivo lung radiographs has not been performed to realize the full potential of these images. In this study, we used neural networks to process EVLP lung radiographs and predict lung transplant outcomes.

Methods

N&#3f1,300 1h and 3h EVLP radiographs from 650 clinical cases were split 80:20 into train and validation sets. Convolutional neural networks (ResNet-50, ResNeXt-50, RexNet-100, EfficientNet-B2, EfficientNet-B3, DenseNet-121) were pretrained using radiographs publicly available from the National Institutes of Health Chest X-ray 14 database. The model was subsequently fine-tuned to first classify transplant decisions as well as donor lung outcomes (transplanted lungs with recipient time to extubation of < 72 hours, ≥72 hours, and lungs declined for transplant). For each EVLP case, 1h and 3h radiographs were concatenated together to yield one classification.

Results

EfficientNet-B2 best predicted transplant suitability with an area under the receiver operating characteristic (AUROC) curve of 89.8%, and ResNet-50 predicted donor lung outcome with an AUROC of 78.4%. The neural network classifications were equivalent to using manually labeled radiographic features to predict transplant suitability (AUROC=89.7%) and donor lung outcomes (AUROC=76.3%).

Discussion/Conclusion

Convolutional neural networks successfully automated EVLP radiograph interpretation when trained to predict lung transplant outcomes. This automatic image analysis allows clinicians to access diagnostic information in the ex vivo lung radiograph without the need for a Radiologist and can easily be adopted by EVLP transplant centers around the globe. Additional studies are underway to evaluate and integrate the predictive value of radiographic analysis with other donor lung assessments obtained during EVLP.

[OR08] Biological Pattern Discovery in Glioma Stem Cell Spatial Organization using Computer Vision and Transcriptomics

Shamini Ayyadhury, University of Toronto/University Health Network
Patty Sachamitr, Blue Rock Therapeutics, Toronto
Michelle Kushida, Peter Gilgan Centre for Research and Learning
Nicole Park, The Hospital of Sickkids
Fiona Coutinho, Brain Canada
Panagiotis Prinos, Structural Genomics Consortium, University of Toronto
Cheryl Arrowsmith, Structural Genomics Consortium, University of Toronto
Peter Dirks, Peter Gilgan Centre for Research and Learning
Trevor Pugh, Princess Margaret Cancer Centre
Gary Bader, The Donnelly Centre, University of Toronto

Introduction

New therapeutics have not improved glioblastoma's abysmal prognosis and there is an urgent need to improve therapeutic screening and discovery. Treatment outcomes in reality are a function of cellular ecosystems, where populations of cells exist in a collective graph-like embedding. Glioma stem cells(GSCs), thought to give rise to GBM, represent an aspect of that community patterning. The application of spatial pixel algorithms from computer vision and deep-learning pipelines can facilitate the understanding of organizational properties of cell communities and improve GSC therapeutic modeling.

Methods

Overall study was conceptualized by SA. Phase-contrast imaging was designed/executed by PS/MMK/NIP/FJC/PP, supervised by CHA/PBD. Computer vision modeling using 29 spatial pixel features was applied for each 17'601 phase-contrast images. Spatial pixel signatures were analyzed using PCA, CCA and multi-linear regression modeling by SA under the supervision of TJP/GDB.

Results

High PC2 scores correlated with mesenchymal/microglial/injury-response gene signatures, whereas the lower PC2 scores correlated with neurodevelopmental and brain cell-type signatures. Algorithms measuring "variance", "contrast" and "homogeneity" spatial pixel distributions were stronger in the neurodevelopmental images. The mid and higher end of the granularity spectrum was stronger in images from the mesenchymal/microglial/injury-response samples. The lower end of the granularity spectrum was enriched along low PC2, where the neurodevelopmental samples were found.

Discussion/Conclusion

Cells do not function in isolation but as complex communities of varying but interdependent cells. Treatments are delivered to these dense and complex structures. We propose evolution of current imaging models that will take into account the spatial architectural complexity of cell communities in high-throughput imaging studies. Therefore, we can include hierarchical patterns of community interaction during compound screening as well as post-screening metadata. These concepts are currently being used to understand tissue architecture and develop AI tools to aid in uncovering architectural principles that are biologically motivated, on the xenium/visium and other HTS platforms/datasets.

Supporting information

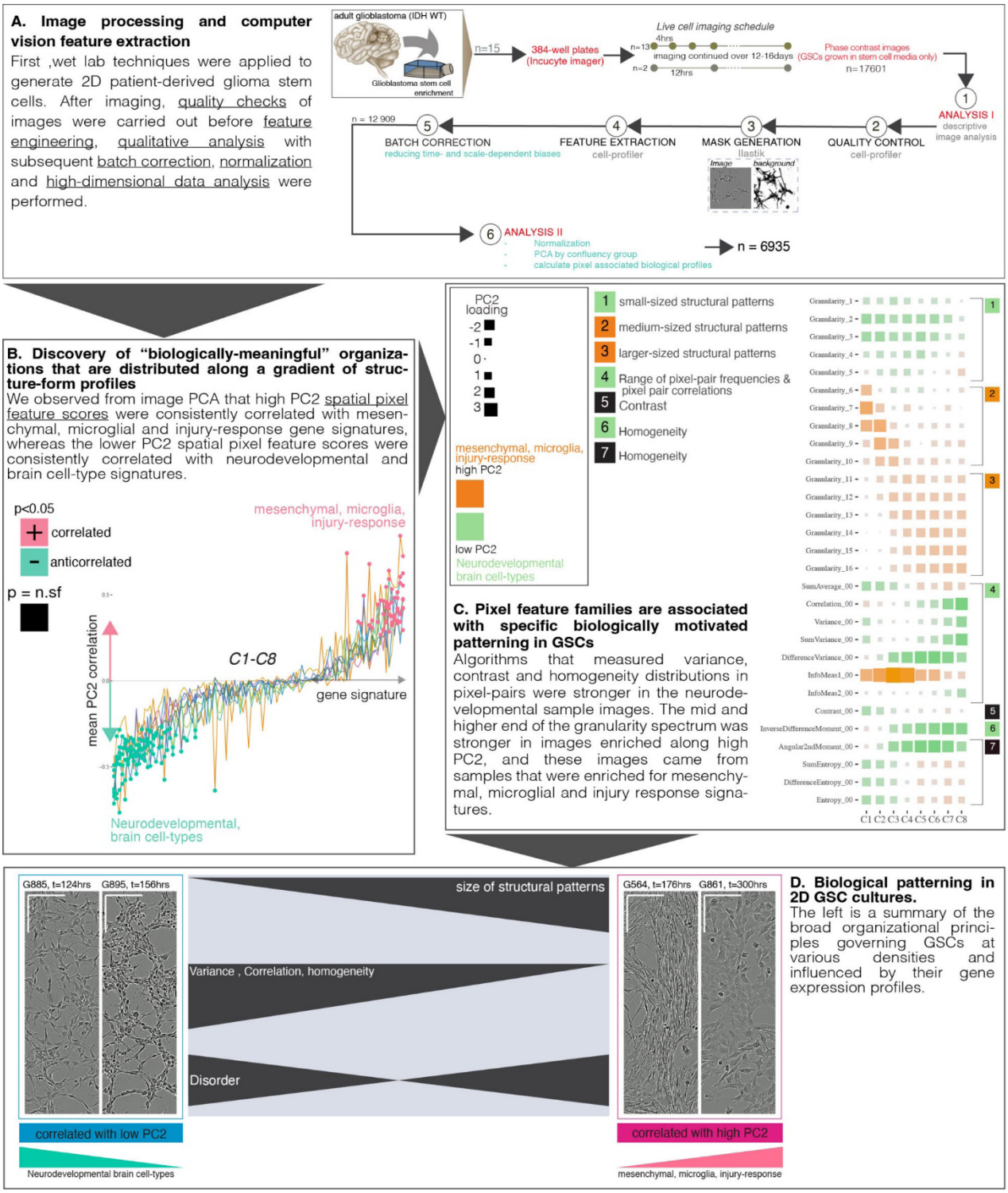


Figure summary of study. Using an unsupervised approach, we show that organizational patterns of glioma stem cells in culture are biologically motivated and hence can be used to understand organizational principles, be integrated with genomics data to complement molecular modeling. Together, the concept can be used to understand behavioral phenotypes in terms of architecture and intrinsic molecular expression, so as to enhance and improve current 2D culture modeling.

[OR09] Designing a Scalable Pipeline for ML Ops to Expedite AI Research and Deployment at Princess Margaret Cancer Centre

Benjamin Grant , Princess Margaret Cancer Centre - UHN
Muammar M. Kabir , Princess Margaret Cancer Centre - UHN
Tirth Patel, University Health Network
Sharon Narine, University Health Network
Tony Tadic, University Health Network
Geoffrey Liu, Princess Margaret Cancer Centre, UHN
Robert C Grant , Princess Margaret Cancer Centre - UHN
Tran Truong, Princess Margaret Cancer Centre - UHN

Introduction

The integration of AI in healthcare has the potential to improve the effectiveness and efficiency of medical practice. Many barriers impede clinical adoption of AI, yet few relate to models themselves. Rather, it is the administrative, logistical, and operational requirements across disciplines and departments that makes real-world clinical deployment of AI – frequently called the 'Last Mile Problem' – so challenging. These requirements include permissions (privacy, security, ethics), processes to reliably develop and validate models, and programs to responsibly deploy, maintain, and monitor them. To address these obstacles, Cancer Digital Intelligence (CDI) has designed a Machine Learning Operations (ML-Ops) pipeline to make AI more accessible, reliable, and impactful for researchers and clinicians at PM.

Methods

CDI is a technology innovation hub that is integrated within UHN Health Information Systems (HIS) and critical digital infrastructure. CDI works closely with oncologists and scientists across PM to facilitate technology-enabled clinical research and AI. Using the MIRA Clinical Learning Environment (MIRACLE), CDI has helped oncologists develop diverse clinical AI solutions, with several models deployed clinically at PM, and more in the pipeline ready for silent mode validation. Our proposed ML-Ops pipeline is a synthesis of our core methodologies and learnings from these efforts.

Results

Our pipeline covers data collection and processing, model training and testing, and real-world validation and deployment at PM. We have standardized strategies for QC, visualization, model development, silent mode, and clinician engagement and feedback. These steps can be automated within the MIRACLE platform, which we hope can serve as a one-stop-shop for clinical AI at PM.

Discussion/Conclusion

Clinical AI solutions need to clear an incredibly high bar, which is infeasible for clinicians or clinical researchers to achieve independently. By bringing like-minded individuals together from across PM and leveraging interdisciplinary expertise, we hope to pave the Last Mile of clinical AI deployment for everyone.

[OR10] Exploring deep-learning to accurately classify breast tissue morphology using Wide-field OCT.

Ali Yassine, Perimeter Medical Imaging AI -University of Toronto ECE Department
Yanir Levy, Perimeter Medical Imaging AI
Ersin Bayram, Perimeter Medical Imaging AI
Mark Nguyen, Perimeter Medical Imaging AI
Ervin Sejdic, University of Toronto ECE Department
Maggie Burns, Perimeter Medical Imaging AI

Introduction

Intraoperative tumor margin assessment is an unmet need as around 20% of breast-conserving surgery requires re-excision. Wide-field optical coherence tomography (WF-OCT) can aid in lumpectomy margin visualization and might help to reduce the need for reoperation. WF-OCT is a novel technology; hence the interpretation of OCT images could benefit from an AI-based clinical decision support system. The focus of this work is to build a multi-label classifier that can aid in training physicians and identify suspicious regions and mimickers in margins.

Methods

We utilized pathology correlated, manually labelled, WF-OCT data with a total of 34,404 training images, 6,072 validation, and 6,728 testing. The dataset was split into 11 categories which included both suspicious and non-suspicious breast features. After exhausting numerous AI network architectures such as ResNets, EfficientNetv2, VisionTransformers, CLIP, VGG16, Mobilenetv3, and customizedCNNs, the highest overall testing accuracy attained was 95%. Given the potential surgical oncology use, we decided to explore ensemble learning to aim for the highest accuracy possible. The deep learning framework used for development was Pytorch and the environment used was ColabPro+ and GCP.

Results

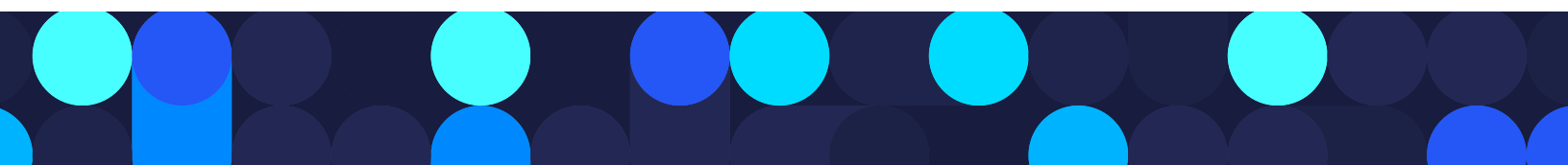
Our approach employs the ensemble of 10 EfficientNetv2Small, attaining a 98.4% accuracy with a 92% confidence threshold for the multi-classifier. An optimized binary classifier derived from the best-performing ensemble achieved recall and precision rates (99% and 100%) in detecting suspicious regions. Furthermore, the multi-classifier provided crucial insights into data distribution, recognizing under-representative tissues often misidentified as positive in smaller convolutional network binary classifier training.

Discussion/Conclusion

The ensemble binary classifier, despite high inference time, can enhance in-surgery margin assessments with precise feature recognition. Additionally, the multi-classifier acts as an educational tool for OCT breast tissue analysis and aids in data curation and auto-labelling. Future research should focus on adding tissue types, exploring larger models, and optimizing to reduce inferencing times.

Supporting information

	Precision	Recall	f1-score	Support
Adipose	1.00	0.98	0.99	660
Bag	0.91	1.00	0.95	39
BloodVessel	1.00	0.98	0.99	183
Clip	1.00	1.00	1.00	18
Cyst	1.00	0.99	1.00	257
Fibrocystic_Change	1.00	1.00	1.00	2
Fibrous	0.92	0.98	0.95	347
Interface	0.95	0.94	0.94	339
NormalDuct	0.85	0.88	0.86	130
Suspicious	1.00	0.99	0.99	2954
Suture	1.00	0.99	1.00	443
accuracy			0.98	5372
macro avg	0.97	0.98	0.97	5372
weighted avg	0.98	0.98	0.98	5372



[OR11] Identifying nucleosome positioning features based on Deep Residual Networks

Yosef Masoudi-Sobhanzadeh, Queen's University
Shuxiang Li, Queen's University
Yunhui Peng, Central China Normal University
Anna Panchenko, Queen's University

Introduction

Nucleosomes represent the elementary building units of eukaryotic chromosomes and consist of DNA wrapped around the histone octamer. Nucleosomes are located in genomes at certain positions which depend, in part, on DNA sequence. Previous studies attempted to investigate how sequence patterns may influence nucleosome positioning. However, they had to use only limited experimental data sets because of the high memory and time complexity of computational methods, required to tackle this problem on the whole genome scale.

Methods

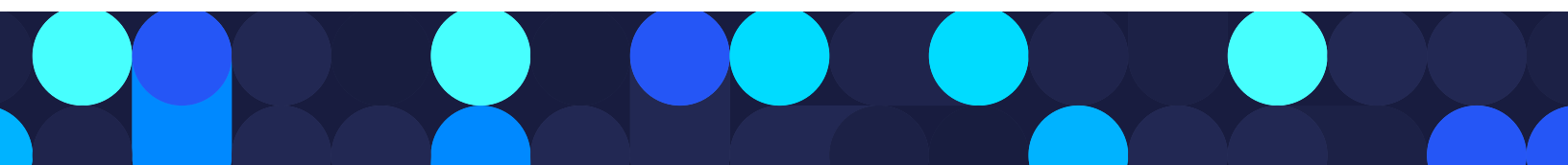
To fill in the above-mentioned gaps, we considered the whole human genome and introduced an efficient deep learning-based framework. Experimental MNase-seq fragments were mapped to the human reference genome, and nucleosome and non-nucleosome regions were determined. Instead of convolutional and pooling layers, a novel two-step feature selection method was introduced and applied to create several candidate subsets of features. For selecting the final subset and for generating an effective nucleosome positioning prediction model, deep residual neural networks were utilized, and the features were then ranked based on AUC analysis.

Results

44 out of 34,272 features were identified as being crucial in determining nucleosome positioning. These features consisted of 9 sequence patterns with 10-11 base pair periodicity; 13 composition-based and 22 position-based features. It is noteworthy that most of the detected features were associated with the minor grooves facing toward/away from the histone octamer.

Discussion/Conclusion

Discussion and conclusion: The periodicity-related features had the highest contribution to determining the nucleosome positioning. From the nucleosome positioning prediction perspective, our prediction model outperformed previous studies in terms of diverse classification criteria.



[OR12] Zero-Shot Medical Image Captioning with Frozen Vision Transformers and Large Language Models

David Li, Western University
Kartik Gupta, Western University
Jaron Chong, Western University

Introduction

The development of multi-modal large language models (LLMs) to enable radiological image interpretation will rely heavily upon billions of curated image-caption pairs for training. Constructing such large-scale medical image-caption datasets poses a significant challenge since images collected from the Internet may lack detailed captions. Recent innovations in foundation models could allow zero-shot captioning of unlabeled radiology images.

Methods

A kNN vector embeddings database search and image dataset augmentation were performed using the ROCO dataset to identify the most similar images in the LAION-5B dataset, which comprises 5 billion images. From the k=100 subset (N&#3f133,960), 100 images were randomly selected. BLIP-2 was utilized for zero-shot image captioning with a frozen CLIP ViT-L/14 vision transformer and FLAN-T5-XL LLM bridged with a querying transformer (Fig. 1). All experiments were conducted on a single server (Intel Xeon 2.2GHz CPU, Nvidia A100 40GB GPU). Captions were generated for each unlabeled image without any text prompts. Image-caption pairs were evaluated for accuracy of modality, anatomic region, diagnosis, image annotations, and overall subjective quality on a scale of 1-5.

Results

Generated captions had a modality and anatomic region accuracy of 86% and 90%, respectively. 61/100 images had a grossly apparent diagnosis of which only 9 were correctly diagnosed, and 2 partially correct. Mean subjective evaluation of overall caption quality was 3.04 with the mode (72%) classification being 3 (model identifies exam correctly but no findings or diagnosis). Annotations were successfully described in 13/16 images. Average inference time was 0.93 ± 0.58s/image.

Discussion/Conclusion

This study demonstrated that frozen vision transformers and LLMs can generate high-quality and accurate captions for radiology images. Zero-shot captioning could expedite the creation of large-scale image-caption datasets and facilitate the development of multi-modal LLMs. The integration of multi-modal LLMs into clinical practice holds great promise for enhancing patient care and efficiency throughout the entire healthcare system.

[OR13] A Motivational-Interviewing Chatbot with Generative Reflections for Increasing Readiness to Quit Among Smokers

Andrew Brown, Candidate
Ash Kumar, University of Toronto
Osnat Melamed, CAMH
Angus Wang, University of Toronto
Marta Maslej, CAMH
Nadia Minian, CAMH
Jodi Wolff, CAMH
Matt Ratto, University of Toronto
Peter Selby, CAMH
Jonathan Rose , University of Toronto

Introduction

The Motivational Interviewing (MI) therapeutic counseling approach to behaviour change, specifically applied to motivation to quit smoking, could be applied at the population level if it could be automated and delivered through the internet. Recent advances in generative AI hold the promise of more natural and contextual interactions for such chatbots.

Methods

An interdisciplinary collaboration among MI-expert clinicians, computer engineers and social scientists designed and evolved an MI-inspired chatbot system to engage with smokers. The conversation consists of 5 questions, a response from the participant, and a generated MI-style reflection based on the question and response. A fine-tuned version of the GPT-2 large language model was used to generate the reflection. A total of n=100 smokers were recruited online to test the chatbot, called MIBot version 5.2. The participants' readiness to quit smoking was measured before the conversation and 1 week later, and the CARE scale (CARE) that measures perceived empathy was applied.

Results

The average confidence level of the smokers that they would succeed in quitting increased (pre-conversation to 1 week later) by an average of 1.3 (SD 2.0, $p < .001$) on an 11-point scale. The average importance to quit increased 0.7 (SD 2.0, $p < .001$) and readiness to quit increased 0.4 (SD 1.7, $p < .05$). These increases were larger than those observed with a simpler version of the chatbot, which only asked the questions and did not provide reflections, although not significantly so. The version with generative reflections was perceived as significantly more empathetic ($p < .004$) than the version that only asked the questions without reflections.

Discussion/Conclusion

The MIBot chatbot has a significant effect on smokers' confidence to quit, motivating further development. More recent large language models, such as GPT-4, might lead to longer even more effective therapeutic conversations.

[OR14] Biopsychosocial Characterization of Cognitive Decline and Late-Life Depression Trajectories Using Bayesian Consensus Clustering and Machine Learning

Mu Yang, CAMH/Dalla Lana School of Public Health

Earvin Tio, CAMH

Rajith Wickramatunga, CAMH

Julie Schneider, Department of Neurological Sciences Neuropathologist, Rush Alzheimer's Disease Center

David Bennett, Rush Alzheimer's Disease Center

Zihang Lu, Queen's University

Daniel Felsky, CAMH/Dalla Lana School of Public Health

Introduction

Interest has been growing to study the comorbidity between cognitive decline and late-life depression (LLD). Existing studies have mostly grouped participants based on clinical diagnoses, without considering severity of symptoms simultaneously. To bridge the gap, we derived longitudinal trajectories of cognition and depression by integrating both self-report symptoms and diagnostic records. We then identified unique and shared biopsychosocial predictors and postmortem consequences of these trajectory subgroups using machine learning.

Methods

We analyzed 2,992 elderly participants enrolled without dementia from the Religious Orders Study and Memory and Aging Project. Integrating symptom and diagnosis records, latent trajectory subgroups were identified independently for cognitive decline and LLD using Bayesian consensus clustering. Logistic regression, elastic-net regression and XGBoost were used to model with predictors including baseline social-demographic measures, personality traits, physical abilities, genetic risk factors and medication burdens. Associations of cognitive and depressive trajectory subgroups on postmortem brainwide neuropathologies were assessed for 1,721 participants.

Results

Three subgroups were identified for cognitive decline and two for LLD were identified, which overlapped significantly (chi-square p=8.1x10-26). Individuals in the fast cognitive decline subgroup were the most likely to experience consistent late life depressive symptoms. Elastic-net regression performed best overall, while XGBoost excelled at predicting moderate subgroups. High neuroticism and low physical health at baseline predicted unhealthy subgroups for both cognition and LLD. Additionally, significant associations with postmortem neuropathologies were identified for cognitive decline trajectory subgroups but not for depression.

Discussion/Conclusion

We performed the most comprehensive analysis of dimensional measures of longitudinal cognitive decline and depression in older adults to date using cutting-edge Bayesian trajectory models. Using machine learning, we found that both physical and mental health, as well as medication burdens were predictive of healthy cognitive and depressive trajectories, providing opportunities to screen at-risk and resilient individuals before the onset of cognitive impairment.

[OR15] Day-to-day variability in activity levels using wearable devices detects transitions to depressive episodes prior to changes in mood: analysis of densely-sampled data from a contactless longitudinal study

Abigail Ortiz , University of Toronto
Ramzi Halabi, CAMH
Martin Alda, Dalhousie University
Ishrat Husain, University of Toronto
Benoit Mulsant, University of Toronto
Arend Hintze, Dalarna University

Introduction

Anticipating clinical transitions in bipolar disorder (BD) is essential for the development of clinically actionable predictions and prevention. We sought to detect the onset of depressive episodes in participants with a primary diagnosis of BD I or II. We hypothesized that changes in activity would be the earliest indicator of a future depressive episode.

Methods

To extract non-stationary hidden patterns from raw wearable data that could be indicative of an episode, we recursively performed self-adaptive decomposition of non-stationary univariate time-series. Then, we computed the Hilbert transform-based instantaneous frequency (IF) of each extracted univariate intrinsic mode function; and the spectral derivative of the individual IFs to highlight the onset of significant spectral variability. Next, to capture the instances for which the IF transitions changed in either direction, we implemented a spectral derivative spike detector (SD2) using a tuned peak detection algorithm. We considered an instantaneous spike to be a true positive if it fell within the 2 weeks preceding the onset of a depressive episode (defined as the start of at least 2 consecutive weeks with a PHQ-9 score ≥ 10).

Results

137 participants were followed for a mean (± SD) 12.6 ± 5.7 months, using both densely-sampled objective data (from wearable) and weekly self-ratings. Analyzing 184,179 wearable-derived datapoints for activity and sleep variables, and 6,985 weekly PHQ-9 scores, we found that the highest accuracy was based on day-to-day variability in number of steps, anticipating the onset of depressive episodes 7-9 days before they occurred with a mean ± SD sensitivity of 0.79 ± 0.27.

Discussion/Conclusion

Changes in activity were the earliest indicator of depressive episodes in our participants with BD. Transitions to dynamic representations of behavioral phenomena in psychiatry may facilitate episode forecasting and individualized preventive interventions.

[OR16] Detecting and Analyzing Potential Comorbid ADHD in People Reporting Anxiety Symptoms from Social Media Data Using Transformers

Noelle Lim, LinkedIn, University of Toronto

Claire Lee, Tesla

Michael Guerzhoy, University of Toronto

Introduction

Up to approximately 50% of adults with ADHD may also have an anxiety disorder. Patients presenting with anxiety may be treated for anxiety without ADHD ever being considered, possibly affecting treatment. We show how data on ADHD comorbid with anxiety can be obtained from social media data, and show that state-of-the-art NLP (Transformers) can be used to detect possible comorbid ADHD in people with anxiety symptoms. We automatically generate explanations for our method highlighting indications of possible comorbid ADHD.

Methods

We collected data from anxiety and ADHD online forums (subreddits). We identified posters who first started posting in the Anxiety subreddit and later started posting in the ADHD subreddit as well. We use this subset of the posters as a proxy for people who presented with anxiety symptoms and then became aware that they might have ADHD.

We fine-tune a Transformer architecture-based classifier to classify people who started in the Anxiety subreddit and then started posting in the ADHD subreddit vs. people who posted in the Anxiety subreddit without later posting in the ADHD subreddit, and report the correct classification rate.

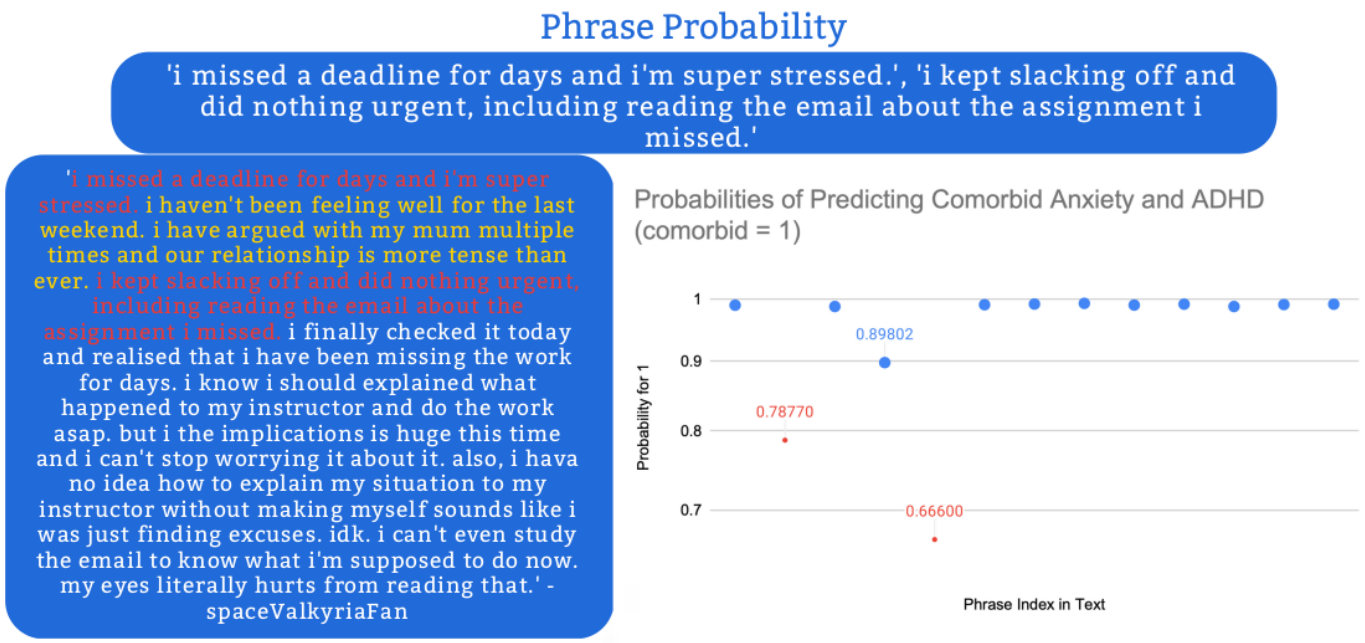
Results

We show that a Transformer architecture is capable of achieving good classification results (76% correct for RoBERTa vs. under 60% correct for the best classical model, both with 50% base rate). We also show that the explanations we obtained for why the classifier predicts that a poster would start posting in ADHD accord with intuition.

Discussion/Conclusion

We present a novel task that can elucidate the connection between anxiety and ADHD; use Transformers to make progress toward solving a task that is not solvable by traditional NLP techniques; and show a method for visualization of our classifier illuminating the connection between anxiety and ADHD presentations.

Supporting information



[OR17] Predictive Care: The False Promise of Fair AI Models in Acute Psychiatry

Christoffer Dharma, University of Toronto, Centre for Addiction and Mental Health
Peter Muirhead, University of Toronto, Centre for Addiction and Mental Health
Susan Bondy, University of Toronto
Juveria Zaheer, Centre for Addiction and Mental Health, University of Toronto
Laura Sikstrom, Centre for Addiction and Mental Health, University of Toronto

Introduction

AI applications have potential to improve mental health care, such as in managing inpatient violence in acute psychiatry. In this setting, machine learning models are being trained on electronic health records (EHRs) to predict which patients are at risk; however, biased training data can result in discriminatory harms. We examine how admission by police to a psychiatric Emergency Department (ED) can bias EHRs for racialized patients, with implications for training machine learning models to make predictions on these data.

Methods

We conducted a computational-ethnography, involving consultations with knowledge users (clinicians, patient advisors), ethnographic observations in the ED, and 22 qualitative interviews. We also compiled EHRs for 8,045 ED visits, to examine the impact of patient race and admission on false positive assessments (i.e., patients identified as high-risk on rating scales who did not become violent).

Results

According to ethnographic findings, police communicate factors resulting in apprehension of patients to ED staff, which increases perceptions of risk and contributes to subtle shifts in care practices and interpersonal interactions. These shifts can impact machine learning models in unmeasurable ways. EHR analyses suggest that Black, Indigenous, and Middle Eastern patients are at higher risk of false positive assessments relative to white patients; however, associations are attenuated or not significant when accounting for police admission, Prevalence Ratio=2.57 (95%CI: 2.30, 2.88).

Discussion/Conclusion

Systemic factors may contribute to unfair assessment of violence risk for racialized patients, which can lead to unfair machine learning predictions when models are trained on EHRs. Current approaches to mitigating bias in training data or modelling do not address the underlying social and political realities leading to unfair assessment (e.g., more frequent apprehension by police). Using machine learning to identify patients at risk of biased assessment may be a more promising approach to supporting assessment and promoting clinically relevant and equitable acute psychiatric care.

[OR18] Serum Metabolomic Pathways in Predicting future onset of Crohn's Disease

Mingyue Xue , Lunenfeld-Tanenbaum Research Institute
Jingcheng Shao , University of Toronto
Sun-Ho Lee , University of Toronto
Anna Neustaeter , Lunenfeld-tanenbaum research institute
Williams Turpin , Lunenfeld-tanenbaum research institute
Kenneth Croitoru , University of Toronto

Introduction

The serum metabolome contains numerous endogenously produced and environmentally absorbed biomarkers, potentially useful for Crohn's disease (CD) detection. Many metabolomic studies aiming to understand CD pathogenesis are comparing samples from patients to healthy participants. However, the presence of established disease is a major confounding factor limiting the capacity to identify biomarkers of CD pathogenesis. Our study assessed the predictive potential of pre-clinical phase metabolites for CD onset and compared their predictive capability with other pre-clinical biomarkers.

Methods

We prospectively followed a cohort of healthy first-degree relatives in the CCC-GEM Project. The cohort included 78 pre-CD participants matched with 312 participants who remained healthy. At enrollment, we measured serum metabolites, microbiome composition, fecal calprotectin (FCP) assay, and the urinary lactulose to mannitol ratio (LMR) assay. We used conditional logistic regression to select biomarkers associated with CD onset, and the selected variables were included in random forest models to assess their predictive potential. The area under the curve (AUCs) of these biomarkers based on 5-fold cross-validation was compared using the Delong test.

Results

Out of the 1001 pre-diagnostic serum metabolites, 63 were associated with CD onset ($1\times10\text{-}4 < q < 0.05$), showing enrichment of sphingomyelin and aspartate, and depletion of isoleucine, leucine, and valine. The metabolome-based CD-prediction model achieved an AUC of 0.89 (95%CI 0.85-0.92) and was significantly superior to LMR-based (AUC=0.48, 95%CI 0.40-0.56, $p < 0.001$), 16s RNA-based (AUC=0.65, 95%CI 0.85-0.92 $p < 0.001$), and FCP-based (AUC=0.70, 95%CI 0.59-0.81, $p < 0.001$) models.

Discussion/Conclusion

The serum metabolite-based model outperformed 16s RNA, LMR, and FCP in predicting CD, suggesting that metabolomic pathways contribute to CD development beyond intestinal permeability, gut microbiome, and inflammation. Identified metabolites offer potential disease pathogenesis insights into the disease pathogenesis, and the prediction model presents an opportunity for CD prevention and early diagnosis.

[OR19] A computational approach to breath-by-breath ventilator waveform data extraction and analysis during ex vivo lung perfusion enables enhanced physiological lung assessment

Xuanzi Zhou , University of Toronto
Lorenzo Del Sorbo, Latner Thoracic Surgery Research Laboratories, University Health Network
Olivia Hough, Latner Thoracic Surgery Research Laboratories, University Health Network
Bonnie T. Chao, Latner Thoracic Surgery Research Laboratories, University Health Network
Jonathan C. Yeung, Latner Thoracic Surgery Research Laboratories,University Health Network
Mingyao Liu, Latner Thoracic Surgery Research Laboratories, University Health Network
Marcelo Cypel, Latner Thoracic Research Laboratories, Toronto General Hospital Research Institute, University Health Network
Bo Wang, Peter Munk Cardiac Centre and the Techna Institute at the University Health Network
Shaf Keshavjee, Latner Thoracic Research Laboratories, Toronto General Hospital Research Institute, University Health Network
Andrew T. Sage, Institute of Medical Science, Temerty Faculty of Medicine, University of Toronto

Introduction

During ex vivo lung perfusion (EVLP), donor lungs are supported by mechanical ventilation and assessment of lung physiology is performed hourly. While physiological data is generated during every breath, it is impractical to capture and analyze this data for clinical decision-making. Herein, we describe a computational approach to ventilator waveform analysis to identify physiological features associated with patient outcomes after transplantation.

Methods

Flow and pressure data were recorded at a resolution of 100Hz using the Servo-i ventilator (Maquet, CA, USA) during n=60 clinical EVLP cases from 2019-2022. An automated algorithm was developed in R language to extract waveform data and analyze individual breaths for lung compliance (static, dynamic) and airway pressure (peak, mean, plateau) for all breaths during EVLP. Breath feature profiles were used to predict EVLP outcomes using a convolutional neural network model.

Results

Automated breath-by-breath data processing was successfully achieved for establishing a complete breath profile for EVLP lungs. We observed that a subtle decreasing trend in dynamic compliance during assessment (-0.01mL/cmH2O/breath, $p=0.05$) was associated with prolonged time to extubation. In addition, the magnitude of the difference in compliance from assessment to baseline tidal volume (10 to 7cc/kg) was significantly associated with the duration of mechanical ventilation post-transplant (+14.9mL/cmH2O, $p< 0.01$). Donor lungs that demonstrated an improvement in baseline, non-assessment, dynamic compliance during EVLP were strongly associated with good recipient outcomes (+7.85mL/cmH2O, $p< 0.001$). The high-resolution breath profiles also enabled the use of a convolutional neural network model to analyze ventilator waveform features and had an area under the receiver operating characteristic curve of 87.5+/-9.5% with 5-fold cross-validation in predicting EVLP outcomes.

Discussion/Conclusion

This study presents a novel computation approach to analyze breath-by-breath features of EVLP ventilator waveform data and its potential application in predicting post-lung transplant outcomes using a deep learning approach.

[OR20] Automated Prognostication using Deep Learning Applied to Chest X-Rays of Patients with Suspected Pneumonia Presenting to the Emergency Department: A Prospective Shadow Deployment Study

Eduardo P. R. P. Almeida, University Health Network - Toronto General Hospital
Shazia Akbar, Altis Labs
Thomas J.. Henessy, Altis Labs
Felix Baudalf-Lenschen, Altis Labs
Sameer Masood, Divison of Emergency Medicine, Department of Medicine, University of Toronto
Felipe S.. Torres, Department of Medical Imaging, University of Toronto

Introduction

This study aimed to validate an automated deep-learning model applied to chest X-rays (CXRs) to predict admission/discharge of patients presenting with signs and symptoms of pneumonia (PNE) to the emergency department (ED).

Methods

Consecutive patients presenting to the Emergency Department (ED) of a tertiary care hospital system from December 2022 to February 2023 who had a frontal CXR performed during the ED visit and had signs and symptoms potentially related to pneumonia were included (n=3,714; 52% males; median age 57 years [Q1:35; Q3:72; IQR 37]). The first CXR acquired from each patient was automatically analyzed by IPRO (Image-based PROgnostication), a deep learning 3D neural network, that generated a probability score ranging from 0 to 1 (0 = discharge, 1 = admission). IPRO performance was assessed using the area under the receiver operating characteristic curve (AUC), stratified by the type of chief complaint (PNE-related, including shortness of breath, general weakness, fever, chest pain with non-cardiac features, and cough/congestion).

Results

A total of 1,062 (29%) patients were admitted to the ward and 94 (3%) were admitted to the ICU, with an in-hospital mortality rate of 2% (67). IPRO predicted the probability of IP admission in all patients with an AUC of 0.795, and for patients with a PNE-related chief complaint, the AUC increased to 0.828 (Figure). Applying an IPRO threshold of < 0.1, IPRO would predict the discharge of 21% of patients with an accuracy of 93.7% and 18% of patients with PNE-related chief complaints with an accuracy of 95%.

Discussion/Conclusion

Deep learning applied to CXRs of patients presenting to the ED with symptoms related to pneumonia can predict IP admission and stratify discharge probability. The application of this model in the ED may help accelerate management and disposition decisions.

Supporting information

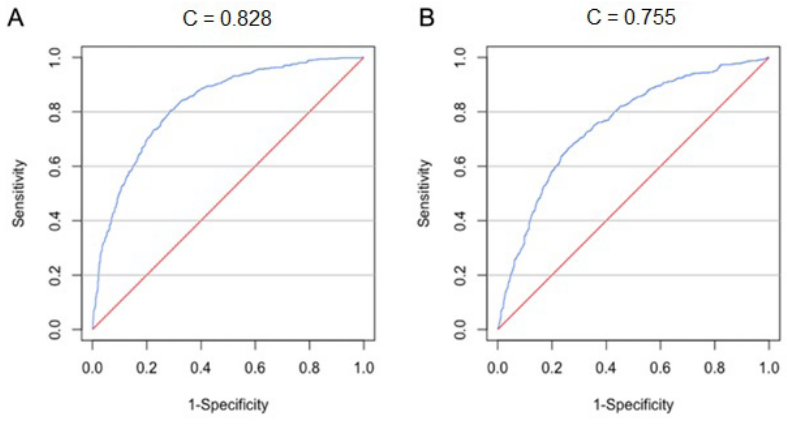


Figure ROC curves for predicting IP admission. A, PNE-related (n=1812), and B, other symptoms (n=1902)

[OR21] ChatGPT and Retinal Disease: A Cross-Sectional Study on AI Comprehension of Clinical Guidelines

Michael Balas, Temerty Faculty of Medicine, University of Toronto
Efrem Mandelcorn, University of Toronto
Peng Yan, University of Toronto
Edsel Ing, University of Alberta
Sean Crawford, University of Toronto
Parnian Arjmand, Mississauga Retina Institute

Introduction

The objective of this study was to evaluate the performance of an artificial intelligence (AI) large language model, ChatGPT (version 4.0), for common retinal diseases in accordance with the American Academy of Ophthalmology (AAO) Preferred Practice Pattern (PPP) guidelines. Although such models have been extensively trained on a broad swath of Internet text, their ability to offer medically accurate and contextually relevant advice remains largely unprobed.

Methods

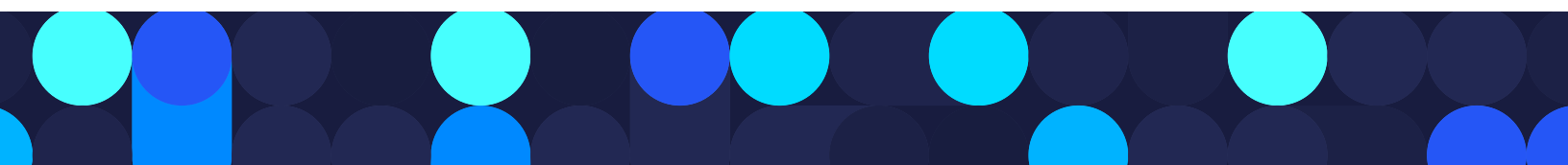
One hundred and thirty questions across 12 AAO PPP domains of retinal disease were input to ChatGPT. Responses were independently evaluated on a Likert scale from 1-5 by three vitreoretinal specialists based on their relevance, accuracy, and adherence to the AAO PPP guidelines. The readability of responses was evaluated using Flesch Readability Ease and Flesch-Kincaid grade level scores. The model's performance across domains was compared using descriptive statistics and the Kruskal-Wallis test.

Results

ChatGPT achieved an overall average score of 4.9/5, suggesting high alignment with the AAO PPP guidelines. Scores varied across domains, with the lowest in the surgical management of disease, including retinal detachment, full-thickness macular hole, and retinal tear domains. Inter-rater reliability for the three raters, as assessed by percent agreement, was substantial at 83.9% overall, implying a high degree of agreement. The responses had a low reading ease score and required a college-to-graduate level of comprehension. Identified errors were related to diagnostic criteria, treatment options, and methodological procedures.

Discussion/Conclusion

This study provides a robust evaluation of the ChatGPT 4.0 AI language model for common retina diseases. ChatGPT 4.0 demonstrated significant potential in generating guideline-concordant responses, particularly for common medical retinal diseases. However, its performance slightly decreased in surgical retina, highlighting the ongoing need for clinician input, further model refinement, and improved comprehensibility. Our findings contribute to the broader understanding of AI language model utilization in ophthalmology and healthcare.



[OR22] Leveraging patient's longitudinal data to predict One-year Mortality Risk

Hakima Laribi, Université de Sherbrooke
Nicolas Raymond, Université de Sherbrooke
Martin Vallières, Université de Sherbrooke

Introduction

Predicting long-term patient survival after admission is crucial for improving palliative care and initiating goals of care discussions. Previous research successfully predicted the Hospital-patient One-year Mortality Risk for all causes using routine admission data, but overlooked valuable temporal information from a patient's multiple hospital visits. Our study proposes to leverage patients' longitudinal data to predict one-year mortality risk at admission. Our code is publicly available on: https://github.com/MEDomics-UdeS/POYM

Methods

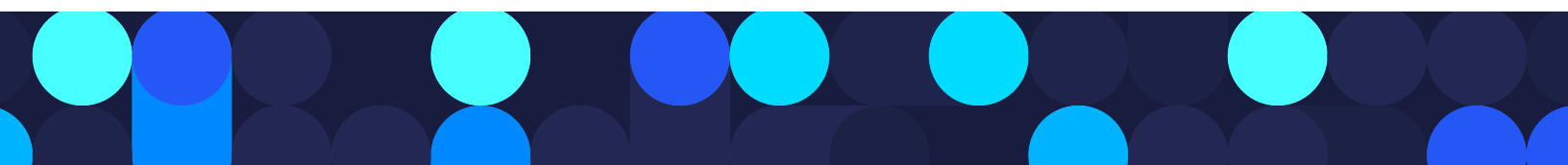
We analyzed 197,905 hospital visits of 116,002 adult patients admitted to a nonpsychiatric service at the University Hospital of Sherbrooke from July 2011 to June 2021. Two models were compared: a previously developed Random Forest (RF) treating patient visits independently, and a new recurrent neural network (LSTM) model considering longitudinal patient data. 5-folds cross-validation on 82,104 patients admitted between 2011 and 2017 assessed the benefits of temporal analysis. Modeled data types were "AdmDemo" (patient demographics and admission characteristics) and "AdmDemoDx" (also incorporating admission and old comorbidities diagnoses). A final LSTM model was trained on admissions between 2011 and 2017 and tested on a separate holdout set of admissions between 2017 and 2021 (33,898 patients), excluding non-eligible end-of-life care visits.

Results

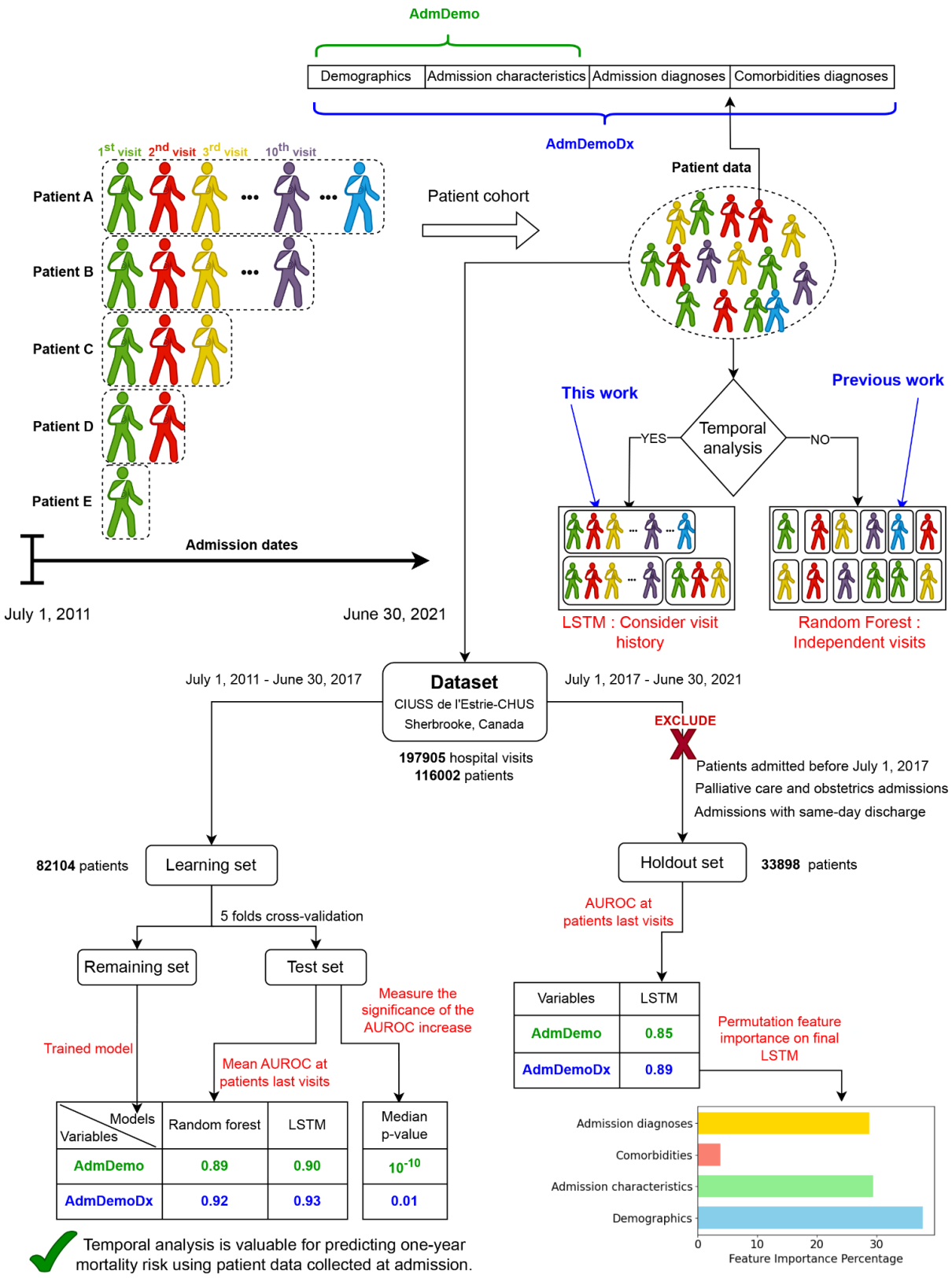
Our LSTM model demonstrated a raise in performance with an AUROC of 0.93 for "AdmDemoDx" and 0.90 for "AdmDemo". The DeLong statistical test confirmed significant differences (p-values < 0.05) between the RF model and our approach in both settings. On the holdout set, our model demonstrated AUROC values of 0.89 for "AdmDemoDx" and 0.85 for "AdmDemo".

Discussion/Conclusion

Our findings demonstrate the effectiveness of temporal analysis in predicting one-year mortality from hospitalization data of different types. Feature importance analysis revealed that simple data available at admission (AdmDemo) are the most discriminant in identifying high-risk patients. Overall, this study emphasizes the significance of leveraging patients' longitudinal data to advance predictive models and enhance patient care.



Supporting information



[OR23] Measures of Overnight Oxygen Saturation can Characterize Sleep Apnea Severity and Predict Postoperative Respiratory Depression

Atousa Assadi, University of Toronto, University Health Network

Frances Chung, University Health Network, University of Toront

Azadeh Yadollahi, University of Toronto-University Health Network

Introduction

Sleep apnea, a highly undiagnosed comorbidity, is a major risk factor for postoperative respiratory depression (PRD). Sleep apnea can be characterized with measures from overnight oxygen saturation (SpO2). Our goal was to extract features from SpO2 signals which are correlated with measures of sleep apnea severity (apnea-hypopnea-index: AHI) and compare their performances in predicting PRD.

Methods

In this retrospective study, we analyzed SpO2 signals of 158 surgical patients, which were recorded continuously (Embletta X100, Embla, Broomfield, CO) preoperatively and on the 3rd postoperative night. PRD was defined as ≥1 hypoxemia episode ≥3 minutes where SpO2≤85%. Preoperative SpO2 signals were analyzed to extract overnight entropy and nocturnal desaturation burden (NDB). NDB was calculated as the area under the curve of desaturation episodes with ≥3% drops. Correlation was used to examine the association between the extracted features and AHI. To compare the performance of sleep apnea related features in predicting PRD, 3 separate machine learning models based on logistic regression were developed using 100 random runs over 80% and 20% of the data for training and validation. Area under the receiver operative cure (AUC), sensitivity, and specificity were used to examine the performance of the models. All 3 models included sex, body mass index, and pre-existing cardiorespiratory disorders in their models to adjust for patients' demographics.

Results

Based on postoperative SpO2, 27 patients (17%) had PRD. Overnight entropy of preoperative SpO2 and NDB were significantly correlated with AHI ($r = 0.73, 0.85$, respectively, $p < 0.0001$ for both). The performance of the models for predicting PRD with preoperative AHI, overnight entropy, and NDB were AUC: 0.81±0.09, 0.80±0.09, 0.81±0.08, sensitivity: 0.70±0.18, 0.73±0.18, 0.70±0.18, and specificity: 0.70±0.08, 0.68±0.08, 0.69±0.09, on validation sets, respectively.

Discussion/Conclusion

Preoperative overnight SpO2 entropy and NDB can be used as alternative measures to characterize the severity of sleep apnea and predict postoperative respiratory depression.

[OR24] Measuring Respiratory Mechanics with Esophageal Catheter and Oscillometry

Shaghayegh Chavoshian , University of Toronto-University Health Network
Nasim Montazeri Ghahjaverestan, University Health Network-University of Toronto
Xiaoshu Cao, University of Toronto-University Health Network
Azadeh Yadollahi, University of Toronto-University Health Network

Introduction

Lung mechanical properties changes are essential for diagnosing or monitoring purposes of respiratory conditions. There are several methods to evaluate lung mechanics such as measuring pleural pressure which is technically challenging and significantly invasive. On the other hand, Oscillometry is a non-invasive device to measure respiratory impedance and lung mechanical functions. Oscillometry device delivers sinusoidal sound waves over a range of frequencies to the mouth during tidal volume. Our objective was to first develop an algorithm for processing pressure signals and then investigate the relationship between lung elastance using Oscillometry and pleural pressure measurements.

Methods

In this study, 19 adults with/without asthma simulated obstructive apnea by breathing against a closed mouth and nose. Oscillometry was performed before and after the intervention. Pleural pressure (via an esophageal catheter) and airflow were measured continuously. Volume signal was estimated as integral of airflow signal. Noise removal, validation algorithms, and curve fitting techniques were employed to reduce pressure signal irregularities. Subsequently, a multiple linear regression model was used to estimate lung elastance using pleural pressure, airflow, and volume data. For Oscillometry measurements, we considered reactance at a high frequency of 37 Hz representing the elastic properties of lung. The elastance difference from pre to post-upper airway obstruction was calculated for both methods. The association between the two approaches was examined using the Bland-Altman plot and Pearson correlation.

Results

Figure 1. presents the association and correlation between the lung elastance changes obtained from pleural pressure and Oscillometry measurements. Our results indicated that two methods to measure lung elastance were highly correlated ($r = 0.84$, $p < 0.0001$).

Discussion/Conclusion

Loss of elastic recoil in asthma during obstructive episodes may predispose to reductions in airflow. This study offers Oscillometry as a non-invasive and reliable method for assessing lung elastic properties changes due to upper airway obstruction in asthma.

Supporting information

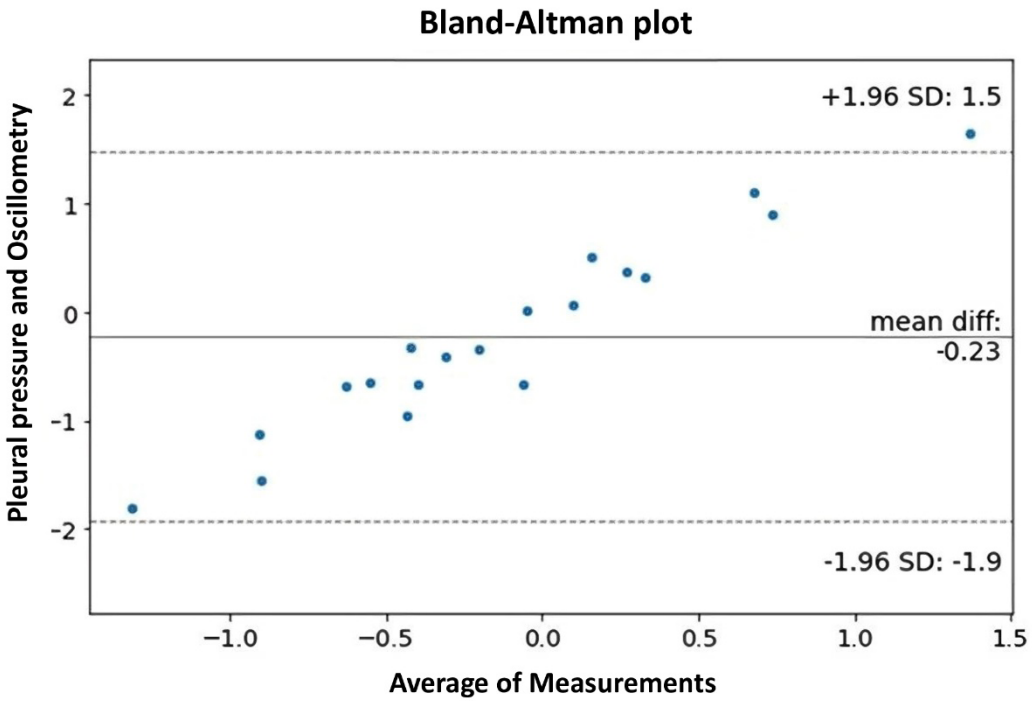


Figure 1. The association between the lung elastance changes obtained from pleural pressure and Oscillometry measurements.

[OR25] An artificial intelligence-driven magnetic resonance imaging synthesis framework

Timur Latypov , Institute of Medical Science, University of Toronto
Marina Tawfik, Department of Computer Science, University of Toronto
Daniel Joergens, Krembil Research Institute, University Health Network
Patcharaporn Srisaikaew, Krembil Research Institute, University Health Network
Paula Alcaide Leon, Joint Department of Medical Imaging, University Health Network
Benjamin Fine, Trillium Health Partners
David Mikulis, Joint Department of Medical Imaging, University Health Network
Frank Rudzicz, Faculty of Computer Science, Dalhousie University
Mojgan Hodaie, Department of Surgery, University of Toronto

Introduction

Magnetic Resonance Imaging (MRI) is essential for deriving information about brain structures and diagnosing neurological disorders. However, variations in MRI acquisition parameters across scanners impede image comparisons, especially between different facilities, necessitating repeat sessions. We explore artificial intelligence (AI) for MRI data synthesis, employing pre-existing T1-weighted images, which are the most prevalent, to generate the often-missing Diffusion Tensor Imaging (DTI) and contrast-enhanced T1 sequences (T1c).

Methods

We used 3D T1 and DTI images from the Amsterdam Open MRI Collection dataset (n=928) as well as T1 and T1c from Brain Tumour Segmentation Challenge 2021 (n=1435) dataset. DTI data was processed, and fractional anisotropy (FA) was extracted using the “Tractoflow” pipeline. We used a 3D U-Net trained in a supervised fashion using the combination of structural similarity index (SSIM) loss and L1 loss to generate synthetic FA and T1c images using contrast-free T1 as an input. Reconstruction accuracy was confirmed for the whole image using voxel-wise coefficient of determination and SSIM. In addition, Johns Hopkins University (JHU) White matter atlas and a two one-sided T-test (TOST) were used for regional FA similarity assessment.

Results

We trained models synthesizing FA and T1c volumes respectively from contrast-free T1 data. Testing subsets of FA and T1c data showed high similarity between real and synthesized images (SSIM>0.91, Rsq>0.9). Regional intensity values of FA within all major 48 white matter structures from the JHU White Matter Atlas were significantly similar across all testing subjects (p-corrected< 0.0001). These measures confirmed the capability of frameworks to produce realistic FA and T1c from T1 sequence.

Discussion/Conclusion

The approaches proposed in this study show high performance and will have immediate impacts on the way patients undergo brain imaging, making procedures more efficient, thereby improving the patient experience, and avoiding negative outcomes.

[OR26] Application of Machine Learning for Clinical Decision Support in the Treatment of Newly Diagnosed Pediatric Crohn Disease Patients

Ricardo Gabriel. Transcripts, University of Alberta

Daniel McClement, University of Alberta

Thomas Walters, SickKids

Russell Greiner, University of Alberta

Eytan Wine, University of Alberta

Introduction

The inflammatory bowel diseases (IBD) - Crohn disease (CD) and ulcerative colitis (UC) - are chronic, debilitating gastrointestinal diseases with no cure.

CD is a highly heterogeneous disease with highly variable response to therapies. This heterogeneity poses significant challenges for clinicians in selecting the most appropriate therapy for individuals.

Dietary therapy is a recommended therapy, for pediatric CD (pCD) patients, with exclusive enteral nutrition (EEN) recognized as the first line therapy for mild-to-moderate pCD. The efficacy of EEN varies greatly from patient to patient.

This study uses baseline and longitudinal data in creating machine learning models that learn to predict clinical respond status after EEN treatment.

Methods

CIDsCaNN prospectively enrolls and follows new onset pediatric IBD cases. Prospective data are available for 336 with pCD who received EEN for eight weeks. Therefore, we collected comprehensive baseline features for each of the 336 pCD patients, at time of diagnoses.

We applied machine learning algorithms to data, to produce a model predicting EEN treatment respond, defined as eight weeks Pediatric Crohn Disease Activity Index (PCDAI) reduction of at least 12.5 points of a patient's PCDAI baseline score. We trained the model on a randomly selected 80% of the data, which involved internal cross-validation. Then, we used 5-fold (external) CV to estimate the quality of that learned model.

Results

This project is in very early stages as we are still performing feature engineering and expanding our dataset to improve the quality of our model. We also still need to demonstrate that the learning process is stable.

Discussion/Conclusion

This research aims to deploy a classifier capable of predicting EEN clinical respond with an accuracy over 60% (i.e., above chance prediction).

This research has the potential to provide better quality of life for children who live with IBD and reduce healthcare burden in the healthcare system of Canada.

[OR27] Can a general large language model augment clinical decision-making in pediatrics?

Esli Osmanlliu , McGill University Health Centre

Introduction

General large language models (LLMs) have demonstrated impressive capabilities for medical challenges in adults. Their performance in pediatrics may differ, given a much smaller availability of online textual data in this population. Moreover, it is unclear if LLMs adequately integrate concepts of sex, gender and race/ethnicity when providing diagnostic suggestions for pediatric conditions. Awareness of the abilities and limitations of LLMs in pediatrics will help clinicians integrate these tools responsibly.

Methods

This study evaluated the diagnostic performance of GPT-4 (OpenAI, San Francisco) using ten clinical vignettes published in the “Clinician’s Corner” section of Paediatrics & Child Health. After submitting clinical, radiological and laboratory data, GPT-4 was asked to provide 1) the most likely diagnosis and 2) the differential diagnosis for each vignette. It was then asked if the true diagnosis for a given case was associated with a) sex, b) gender and c) race/ethnicity. Data were analyzed descriptively, as the proportion of accurate GPT-4 responses for each question.

Results

GPT-4 identified the correct diagnosis in 2/10 cases (table 1). It included the correct diagnosis in 6/8 (75%) of the remaining cases. Overall, GPT-4 correctly identified the presence or absence of any association with sex, gender, and race/ethnicity in 9/10, 5/10 and 8/10 cases respectively. In one answer, GPT-4 stated that “in everyday language and in much medical literature, ‘sex’ and ‘gender’ are often used interchangeably”, while specifying the actual differences between these concepts. Yet, it confused the concepts of sex and gender in 50% of the cases.

Discussion/Conclusion

GPT-4 can augment clinical decision-making for pediatric patients with challenging presentations through the formulation of an adequate differential diagnosis. It was however unable to select the right diagnosis in most cases. A more consistent use of sex and gender as distinct concepts constitutes an area for improvement in future iterations.

[OR28] Clinical Features, Non-Contrast CT Radiomic and Radiological Signs in Models for the Prediction of Hematoma Expansion in Intracerebral Hemorrhage

Frank Chen , Queen's University

Introduction

Rapid identification of hematoma expansion (HE) risk at baseline is a priority in intracerebral hemorrhage (ICH) patients and may impact clinical decision making. Predictive scores using clinical features and Non-Contract Computed Tomography (NCCT)-based features exist, however, the extent to which each feature set contributes to identification is limited. This paper aims to investigate the relative value of clinical, radiological, and radiomics features in HE prediction.

Methods

Original data was retrospectively obtained from three major prospective clinical trials ["Spot Sign" Selection of Intracerebral Hemorrhage to Guide Hemostatic Therapy (SPOTLIGHT)NCT01359202; The Spot Sign for Predicting and Treating ICH Growth Study (STOP-IT)NCT00810888] Patients baseline and follow-up scans following ICH were included. Clinical, NCCT radiological, and radiomics features were extracted, and multivariate modeling was conducted on each feature set.

Results

317 patients from 38 sites met inclusion criteria. Warfarin use (p=0.001) and GCS score (p=0.046) were significant clinical predictors of HE. The best performing model for HE prediction included clinical, radiological, and radiomic features with an area under the curve (AUC) of 87.7%. NCCT radiological features improved upon clinical benchmark model AUC by 6.5% and a clinical & radiomic combination model by 6.4%. Addition of radiomics features improved goodness of fit of both clinical (p=0.012) and clinical & NCCT radiological (p=0.007) models, with marginal improvements on AUC. Inclusion of NCCT radiological signs was best for ruling out HE whereas the radiomic features were best for ruling in HE.

Discussion/Conclusion

NCCT-based radiological and radiomics features can improve HE prediction when added to clinical features.

[OR29] Development of a Multi-modal Machine Learning-Based Prognostication Model for Traumatic Brain Injury Using Clinical Data and Computed Tomography Scans

Atsuhiro Hibi, Institute of Medical Science, University of Toronto
Pascal Tyrrell , Department of Medical Imaging, University of Toronto
Michael Cusimano, Division of Neurosurgery, St. Michael's Hospital
Alexander Bilbily, Sunnybrook Health Sciences Centre
Rahul G Krishnan, Department of Computer Science - UofT

Introduction

Computed tomography (CT) is a preferred imaging modality for prognostication in traumatic brain injury (TBI) patients. However, due to the specialized expertise required, timely and reliable TBI prognostication based on CT imaging and other clinical data remains challenging. This study aimed to enhance the efficiency and reliability of TBI prognostication by employing machine learning (ML) techniques on CT images and non-imaging clinical data.

Methods

A retrospective analysis was conducted on the CENTER-TBI, a multi-center TBI dataset collected across Europe (n = 1,016). Our prognostic model was based on the three-dimensional Vision Transformer, which is primarily used for video recognition. We adapted this model to treat a video as analogous to a CT scan and addressed dimension imbalance between CT and non-imaging variables to incorporate multi-modal input. Our developed model predicted favorable or unfavorable outcomes at six months post-injury. The prognostic performance was assessed using the area under the curve (AUC) over five-fold cross-validation and compared to conventional models relying on clinical variables and CT scoring systems. External validation was performed using a CINTER-TBI dataset collected in India (n = 348).

Results

The developed model achieved superior performance without manual CT assessments (AUC = 0.846) compared to the model based only on the clinical and laboratory variables (AUC = 0.817) and major CT scoring systems necessitating manual interpretations (AUC = 0.829 for Marshall and 0.838 for IMPACT). External validation demonstrated the prognostic capacity of the developed model to be significantly better (AUC = 0.859) than the model using clinical variables (AUC = 0.809).

Discussion/Conclusion

Our study demonstrated the potential of a multi-modal ML model in providing more efficient and reliable TBI prognosis based on CT scans. This approach may have implications for earlier intervention and improved TBI patient outcomes.

[OR30] Exploratory Analysis of Perfusion Index as a Screening Tool for Continuous Monitoring of Blood Pressure in Critically Ill Children

Mana Shahriari, CHU Sainte-Justine Research Centre
Carla Said, CHU Sainte-Justine Research Centre
Clara Macabiau, CHU Sainte-Justine Research Centre
Guillaume Emeriaud, CHU Sainte-Justine Research Centre
Rita Noumeir, CHU Sainte-Justine Research Centre
Philippe Jouvet, CHU Sainte-Justine Research Centre

Introduction

Perfusion index (PI) is derived from PPG signal and represents the ratio of pulsatile to non-pulsatile blood flow in the peripheral tissue. PI is an indicator of peripheral circulation and central perfusion; it represents local blood volume variations which is the resultant of systemic and local hemodynamic status. As such, it has the potential to be used for continuous and non-invasive hemodynamic monitoring. We plan to develop an algorithm to predict blood pressure (BP) from PPG and EKG signals. In this work, we studied if PI needs to be integrated into the algorithm, i.e., if there is a correlation between BP and PI measurement.

Methods

We conducted a study including children aged 0 –18 years, admitted to the pediatric intensive care unit of Sainte-Justine University Hospital whose PPG and invasive arterial blood pressure recordings were available simultaneously. We used our high-resolution database (HRDB)1 and patient characteristics such as age to investigate the correlation between systolic and mean arterial blood pressure (SaBP and MaBP) and PI values.

Results

1206 children were included in this study. We used generalized estimation equations to examine the correlation between PI and SaBP & MaBP. The analysis showed that PI and SaBP & MaBP were inversely related with a coefficient that varied among age groups and ranged from -0.82 to -2.39. Additionally, a stronger correlation was observed for SaBP and PI than for MaBP and PI.

Discussion/Conclusion

As there is a correlation between PI and BP in children, we consider including PI into the algorithm that predicts BP from PPG and EKG signals.

Supporting information

	Coefficient for PI and SaBP (p< 0.001)	Coefficient for PI and MaBP (p< 0.001)
Neonate	- 1.27	- 1.26
Infant	- 2.11	- 1.74
Toddler and Preschool	- 0.95	- 0.82
School Age Child	- 1.71	- 1.99
Adolescent	- 2.39	- 1.07

[OR31] Machine learning enables detection of Li-Fraumeni Syndrome using tumor whole-genome sequencing

Brianne Laverty, The Hospital for Sick Children
Vallijah Subasri, The Hospital for Sick Children
Nicholas Light, The Hospital for Sick Children
Scott Davidson, The Hospital for Sick Children
Mehdi Layeghifard, The Hospital for Sick Children
Rose Venier, The Hospital for Sick Children
Adam Shlien, The Hospital for Sick Children
Marianne Koritzinsky, University Health Network
Bo Wang, Department of Laboratory Medicine and Pathobiology, University of Toronto
David Malkin, The Hospital for Sick Children

Introduction

Li-Fraumeni syndrome (LFS) is a hereditary cancer predisposition syndrome caused by germline mutations in the tumour suppressor gene TP53. LFS is associated with an 80% lifetime cancer risk, with 46% of patients developing a second primary tumour. Therefore, diagnosis is critical to implement surveillance for secondary malignancies. However, diagnostic protocols fail to detect 25% of individuals. We hypothesized that LFS tumours evolve uniquely from sporadic cancers, providing them with characteristic features. To test this, we

Methods

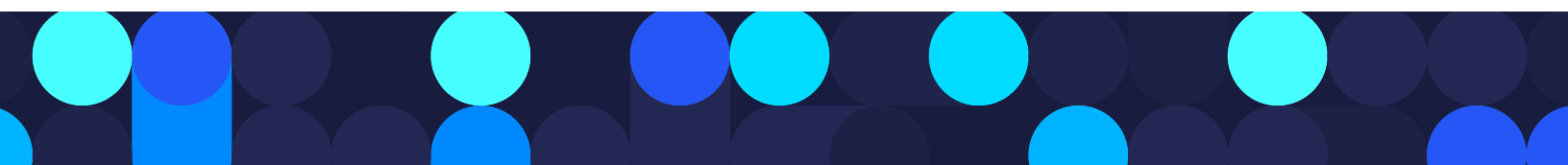
We developed and compared the performance of four machine learning algorithms to discriminate LFS (n=52) from non-LFS (n=193) patients using tumour variant patterns. The models were trained and tested using 5-fold nested cross-validation which allowed the use of the entire dataset to tune the hyperparameters and evaluate the performance without data leakage. We performed an ablation study to determine the necessary model features. Shapley Additive exPlanations (SHAP) was used to evaluate the feature importance to ensure the features made biological sense and determine areas of the tumour genome that may mutate uniquely in LFS patients.

Results

Our model achieved an F1 score of 0.69 (95% CI: 0.66-0.72) and area under the precision recall curve of 0.82 (95% CI: 0.80-0.85). Importantly, the model had an excellent precision of 0.85 (95% CI: 0.82-0.88). An ablation study determined 5 features that made biological sense were necessary for prediction.

Discussion/Conclusion

We have developed a classifier that uses tumour features to detect LFS. This represents the first machine learning tool to identify a cancer predisposition syndrome (CPS) from the tumour genome. As precision oncology expands and tumour sequencing programs become more widespread, a tool to detect CPS from the tumour genome will facilitate early diagnosis, leading to entrance into surveillance programs and improved outcomes.



[OR32] Pixels and Perspectives: Exploring Public Perceptions about AI-Generated Images of Children with Medical Conditions

Dave Lysecki, McMaster Univeristy

Gregorio Zuniga-Villanueva, McMaster University

Muhammed Mukadam, Quality of Life and Advanced Care Program

Introduction

Images of children with medical conditions are routinely used by organizations to educate, fundraise, or increase awareness. Using generative Artificial Intelligence (AI) to create images of children with medical is possible; however, this raises ethical and societal questions regarding AI biases, copyright, and privacy. Currently, it is unknown how accurately generative AI can represent vulnerable populations like children with medical conditions through images and whether using these images can impact public perception or acceptance.

Methods

Images of children with various medical conditions were through the generative AI software Midjourney® and were compiled into a 1-minute video. The video was shown to a purposive sample of healthcare professionals and parents of children with medical conditions, including one with expertise in AI. Feedback on the video was assessed at two different moments, after the video was shown and after disclosing that the images were AI-generated.

Results

The software could not accurately recreate certain medical conditions (e.g. trisomy 21) or medical technologies (e.g. tracheostomy). Sample responses revealed contrasting viewpoints. Concerns included feeling deceived when disclosing they were AI-generated images, misrepresentation of vulnerable groups, the inherent bias of the AI, and the use of unlicensed content in the creation of the AI tool and generation of these images. However, benefits included the potential for more diversity among images, protection of patient privacy, the absence of coercion or favouritism when choosing real-life patients as models, and cost-effectiveness.

Discussion/Conclusion

AI-generated images of children with medical conditions have limitations in the accuracy of representation based on biased datasets. Public perception showed contrasting perspectives that may inform policymakers and stakeholders on using AI-generated images of children with medical conditions. While there are benefits to using this emerging technology, risks and biases need to be addressed before adopting an open and wide use of this resource.

[OR33] Prediction of Emergency Department Readmission among Child and Youth Mental Health Outpatients Using Deep Learning Techniques.

Simran Saggu, Offord Centre for Child Studies

Introduction

The proportion of Canadian youth seeking mental health supports from emergency departments (ED) has risen in recent years. As EDs typically address urgent mental health crises, readmission to EDs could represent unmet mental health needs. Accurate ED readmission prediction could aid early intervention and ensure efficient healthcare resource allocation. The potential increased accuracy and better fit of graph neural network (GNN) machine learning models for predicting ED readmission in electronic health record (EHR) data is unexplored.

Methods

This study used EHR data for child/youth outpatients seeking services at McMaster Children's Hospital's Child and Youth Mental Health Program to develop and evaluate GNN models to predict: a) 30-day ED readmission (classification) and b) ED readmissions within 90 days (regression) of a child/youth's initial mental health outpatient visit. Models were evaluated and compared using F1 accuracy and root mean square error (RMSE) scores and compared to the most comparable recurrent neural network (RNN) models.

Results

For classification, the GNN model (PTO) outperformed the RNN model (Bi-GRU) by an increase in F1-score of 3.12%. For the regression task, the GNN model (BEGO) outperformed the RNN model (Bi-GRU) by a decrease in error score of RMSE of 5.3645.

Discussion/Conclusion

This study demonstrates the improved accuracy and potential utility of GNN models in predicting ED readmissions among child/youth mental health outpatients. These models have potential for use to inform clinical decision-making in a way that can facilitate targeted interventions, optimize resource allocation, and improve mental health outcomes for children/youth.

[OR34] Relationship between Air Pollution and Crohn's disease (CD) Risk and their relation to Biomarkers of CD risk

Jingcheng Shao , University of Toronto
Mingyue Xue , Lunenfeld-Tanenbaum Research Institute
Anna Neustaeter , Lunenfeld-tanenbaum research institute
Sun-Ho Lee , University of Toronto
Williams Turpin , Lunenfeld-tanenbaum research institute
Kenneth Croitoru , University of Toronto

Introduction

The etiology of Crohn's disease (CD) remains unknown, but its higher prevalence in polluted areas suggest a possible contribution of air pollution to CD pathogenesis. Therefore, we investigated the relationship between exposure to air pollution levels, CD risk and its biomarkers.

Methods

Healthy participants were enrolled in the CCC-GEM Project. Baseline samples included the urinary fractional excretion ratio of lactulose-to-mannitol (LMR, marker of gut barrier function), fecal-calprotectin (FCP), and stool 16SrRNA sequencing data in 2,256 participants. Eight pollutants were obtained from the National Air Pollution Surveillance Database. Inverse distance weighting via participants' postal code was used to calculate the average pollutant exposure three months before enrolment.

The impact of air pollution on CD risk was evaluated using an Additive Cox Proportional Hazard Model. Xgboost, DeepSurv and Deephit were used to compare the pollutant' predictive power. Generalized estimating equation models evaluated the association of pollutants with LMR, FCP, and 16SrRNA.

Results

Exposure to Particulate matter< 10um (PM10) was associated with increased CD risk ($p< 0.031$). PM10 was positively associated with LMR ($p=0.047$ and 0.044) but inversely associated with FCP ($p=0.005$). Presence of Proteobacteria was negatively associated with exposure to nitrogen oxides ($fdr=0.005$), and positively with ozone ($fdr=0.034$).

Three-month pollution data predicted CD risk best with a c-index of 0.60 by Xgboost. DeepSurv and DeepHit did not improve the c-index much (0.58 and 0.60) but had good Integrated Brier scores (0.04 each).

Discussion/Conclusion

Exposure to air pollutants were associated with CD risk, related biomarkers, and microbiome composition. Our prediction models illustrate the predictive potency of pollutants in anticipating CD risk, suggesting that air pollution has a significant role in modulating gastrointestinal health and potential implications in CD pathogenesis.

[OR35] A Surgical Robot Simulation Framework for Reinforcement Learning to Automate Manipulation and Cutting Subtasks

Mustafa Haiderbhai, University of Toronto
Radian Gondokaryono, University of Toronto
Andrew Wu , University of Toronto
Ivy Tan, University of Toronto
Lueder A.. Kahrs, University of Toronto

Introduction

Surgical robotics is a growing area of research with constant advancements towards shared and full autonomy. Reinforcement Learning (RL) methods can enable vision-based control for manipulation and cutting of surgical robotics subtasks. Simulations facilitate efficient large-scale data collection in a simulated environment. Recent simulators for surgical robotics allow for RL training of surgical robotics tasks, but are not well suited for training vision-based agents that need to be executed on real robots. Furthermore, these simulators are often designed for a specific robot setup, or lack a sim2real pipeline for working with different robot/camera setups.

Methods

We build a flexible surgical robotics simulation in Unity3D. Our simulation can support any robot setup, including our implemented da Vinci Research Kit (dVRK) robot, and the Franka Emika Panda robot equipped with EndoWrist instruments of the da Vinci Surgical System (dVSS). Our simulation includes a sim2real pipeline that can transfer trained agents from simulation directly onto the real robot system, using modular kinematics and shared interfaces between simulation and real. We train agents for subtasks such as block pushing, block rolling, rope cutting, and block picking. Our agents are trained purely in simulation using Domain Randomization.

Results

We achieve a high success rate of greater than 90% across all subtasks after sim2real transfer of our trained agents. Our pipeline includes testing with multiple cameras, including USB cameras and the dVSS endoscope. Our domain randomization creates robust agents that can adapt to changes in camera, lighting, shapes, and sizes.

Discussion/Conclusion

Our flexible and robust simulation allows for training of RL agents for surgical robotics tasks purely in simulation across different robots. Future work will examine more complex tasks with sequences of subtasks towards automating surgery.

Supporting information

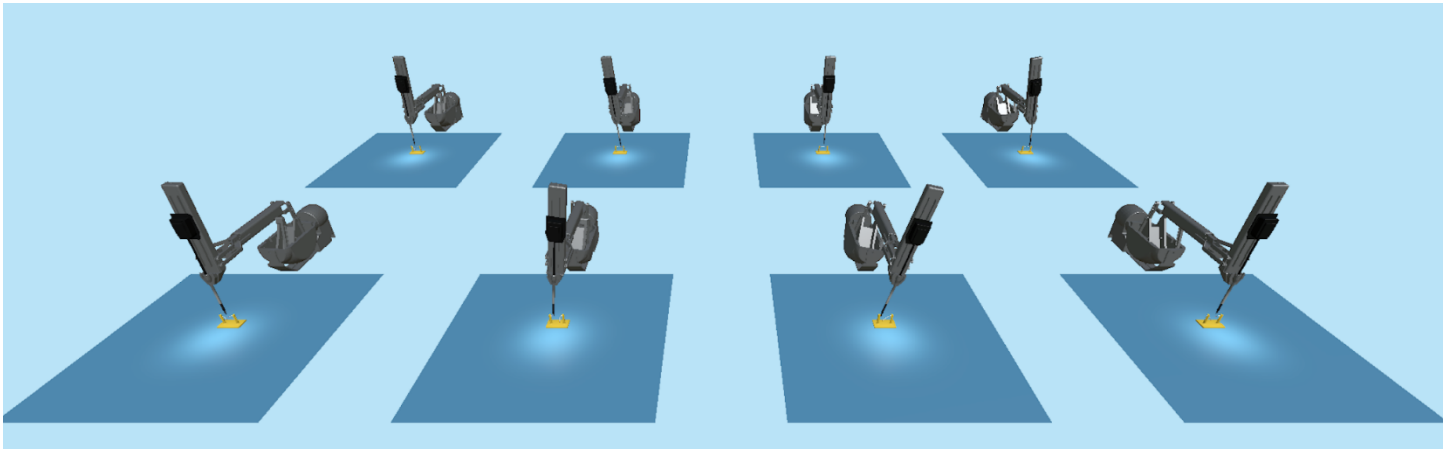


Figure 1. Our simulation running parallel training environments with the da Vinci Research Kit for a rope cutting task.

[OR36] Classifier to predict drug cardiac activity using human stem cell-derived cardiac tissues

Julia Plakhotnik, Hospital for Sick Children
Veronika Feherova, Deloitte
Kaley Hogarth, Hospital for Sick Children
Jason Maynes, Hospital for Sick Children

Introduction

Human-derived cardiac tissues can offer better clinical translational value than animal models in drug development pipelines. However, quickly identifying potential cardiac activity remains challenging. To improve cardiac testing efficiency, we built a proof-of-concept predictive cardiac activity ML classifier for drug candidates.

Methods

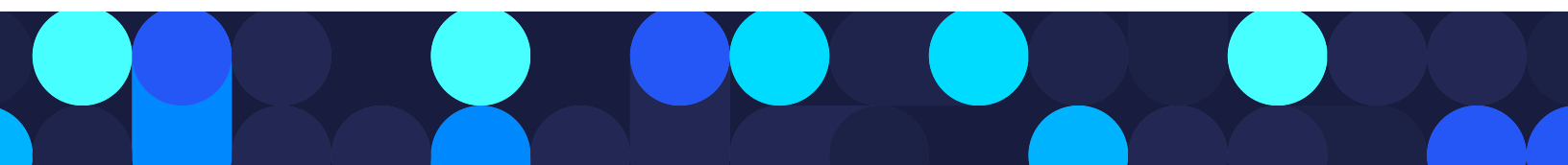
Spontaneously beating human stem cell-derived cardiac tissues were cultured on biomechanical substrates mimicking the native cardiac microenvironment. Tissues were treated with cardioactive drugs spanning 17 distinct molecular target-based categories. Classifier training data (N&#3f2312) consisted of drug-mediated tissue contractile function changes, extracted with optical flow-based tissue motion tracking. Test data (N&#3f1117), reserved for final model evaluation, contained 3 “new-to-classifier” drugs, including those with cardiotoxic side effects, and remaining categories from repeat experiments. Along with traditional cross-validation, a leave-drug-out validation scheme was employed to simulate encountering “new-to-model” drugs. Hyperparameter tuning, feature selection and engineering strategies were informed by preliminary modeling using XGBoost, Shapley explanations, feature ANOVA F-scores and multi-collinearity.

Results

We tuned 5 different classifiers (logistic regression, KNN, naïve Bayes, random forest, boosted trees) on 4 different feature sets, training with 100-200 observations per drug category and 20 cardiac contractility features. After tuning, random forests performed best across all feature sets (0.83 mean F1); KNN the worst (0.76 F1). Random forest scored similarly on the final test set, with high scores for “new-to-model” drug categories (0.82-0.93 F1), indicating classifier generalizability. Misclassifications occurred within medically-related drug classes, and per-class feature importance profiles matched the drugs’ known mechanisms of action, highlighting the biological relevance of classifier predictions.

Discussion/Conclusion

Using data from in-vitro cardiac testing mimicking human cardiac mechanobiology, random forest successfully classified 17 different cardiotropic drug categories with well-established molecular effects. Retroactively, our classifier identified drugs with known cardiotoxic side effects, illustrating how human-derived cardiac tissues combined with ML may mitigate potential downstream toxicity issues in drug pipelines.



[OR37] Detecting Arrhythmogenic Right Ventricular Cardiomyopathy (ARVC) using Residual Neural Networks

Chris J. George, University of Toronto
Sophie Sigfstead, University of Alberta
River Jiang, University of British Columbia
Brianna Davies, St. Paul's Hospital
Andrew Krahn, University of British Columbia
Christopher Cheung, University of Toronto

Introduction

Arrhythmogenic Right Ventricular Cardiomyopathy (ARVC) is a rare and life-threatening cardiac disorder, and can be associated with sudden cardiac arrest. Genetic in origin, ARVC affects approximately 1-2 in 5000 individuals worldwide, leading to fatty infiltration of the right ventricular wall and eventually resulting in ventricular tachycardia. Early diagnosis is crucial to implement preventive measures, yet it remains challenging. Given the significant morbidity associated with this condition, there is a pressing need to explore novel diagnostic approaches, such as utilizing neural networks to detect ARVC in electrocardiograms (ECGs). This project aims to validate these models and develop tools in an area where the existing literature is currently limited.

Methods

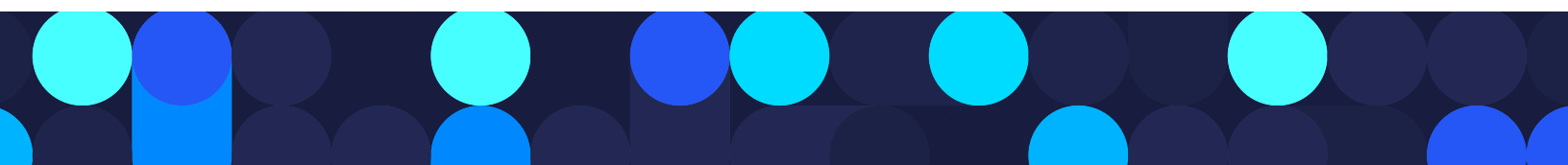
ECG records were gathered from the Hearts in Rhythm Organization (HiRO) ARVC registry, encompassing data from 12 Canadian sites. The dataset comprised 811 ECGs, including 340 ECGs from gene-positive ARVC patients (disease affected individuals with an identified ARVC-related gene) and 471 ECGs from unaffected individuals. A cardiologist manually evaluated the ECGs, excluding 45 samples that presented significant quality issues, resulting in a training set of 766 ECGs.

Results

The model produced strong results when evaluated on a 115-record test set, exhibiting an average specificity of 84.7% and sensitivity of 93.0%. The average accuracy and F1 score were 81.0% and 85.1% respectively. Improved performance is likely possible through further development of model architecture and hyperparameter tuning.

Discussion/Conclusion

Cost-effective machine learning models show great promise in identifying ARVC, warranting further investigation using larger and more diverse sample populations. Expanding analyses with Convolutional Neural Networks and incorporating samples from other patient groups, such as gene negative ARVC patients (affected patients without an identified ARVC-related gene), can provide deeper insights into this evolving area, enabling earlier diagnoses and ultimately improving patient outcomes.



[OR38] Development and evaluation of a live birth prediction model for evaluating human blastocysts

Hang Liu , University of Toronto
Guanqiao shan , University of Toronto
Yu Sun, University of Toronto

Introduction

Since 1978, more than 8 million children have been conceived through IVF. Yet, only about 30% of IVF attempts result in a successful birth. As a result, fertility patients often undergo multiple rounds of IVF, which can be expensive and emotionally draining. Several factors determine IVF success, one of which is the health of the blastocysts selected for transfer to the uterus. Specialists select the blastocysts using several criteria. But these human assessments are subjective and inconsistent in predicting which ones are most likely to result in a successful birth. Recent studies suggest artificial intelligence technology may help select blastocysts.

Methods

This study was conducted on retrospectively collected data of blastocyst images, patient couple's clinical features, and live birth outcomes, involving 17,580 blastocysts in frozen embryo transfer (FET) from 2016 to 2020 in a single IVF center. The individual effect of blastocyst images and the combined effect of patient couple's clinical features on live birth prediction were quantified.

Results

The live birth prediction model using only blastocyst images achieved an AUC of 0.67 (95% confidence interval (CI): 0.65-0.70). The live birth prediction model using both blastocyst images and patient couple's clinical features achieved an AUC of 0.77 (95% CI: 0.75-0.79). Sixteen clinical features were identified to be predictors of live birth outcomes and helped improve live birth prediction. Among these features, maternal age, the day of blastocyst transfer, antral follicle count (AFC), retrieved oocyte number, and endometrium thickness measured before transfer are the top five features contributing to live birth prediction.

Discussion/Conclusion

Artificial intelligence-aided blastocyte selection using patient and blastocyst characteristics may improve IVF success rates and reduce the number of treatment cycles patient couples undergo. Before specialists can use artificial intelligence in their clinics, they must conduct confirmatory clinical studies that enroll patient couples to compare conventional methods and artificial intelligence.

Supporting information

https://elifesciences.org/articles/83662#digest

[OR39] Echocardiogram Quality Enhancement with Vector Quantized Generative Adversarial Networks (VQ-GANs)

Alif Munim, Peter Munk Cardiac Centre (UHN)

Zeinab Navidi, Vector Institute

Wendy Tsang, Peter Munk Cardiac Centre (UHN)

Bo Wang, Department of Laboratory Medicine and Pathobiology, University of Toronto

Introduction

Echocardiograms are a vital tool in diagnosing cardiovascular diseases, but their interpretation can be hindered by inherent noise and artifacts. While autoencoder-based architectures have been previously employed for denoising echocardiogram videos, there's a pressing need for more robust techniques. In this study, we explore the utilization of Vector Quantized Generative Adversarial Networks (VQ-GANs), a more powerful architecture, to address this challenge effectively.

Methods

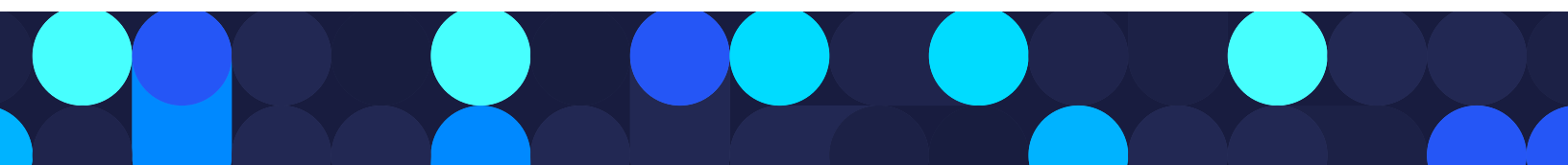
We employed VQ-GANs to denoise echocardiogram videos, leveraging their ability to generate high-quality and realistic data. We used the Stanford EchoNet-Dynamic dataset to train the model, comprising over 10,000 echocardiogram videos. Introducing Gaussian, speckle, and salt & pepper noise, commonly found in ultrasound imaging, we created a dataset with over 100,000 noisy videos. We conducted a comparative evaluation between the proposed VQ-GAN-based approach and existing autoencoder-based methods.

Results

Results showed that the VQ-GAN-based model outperformed autoencoder-based models in terms of denoising effectiveness. Qualitative and quantitative evaluation using metrics such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) showed significant video quality improvement, demonstrating the superiority of the VQ-GAN model.

Discussion/Conclusion

The application of VQ-GANs for denoising echocardiogram videos represents a promising approach in medical imaging, providing clinicians with clearer, more interpretable imaging for accurate diagnoses. Future research should focus on further optimizing the VQ-GAN architecture and validating the approach with a broader range of medical imaging modalities.



Supporting information

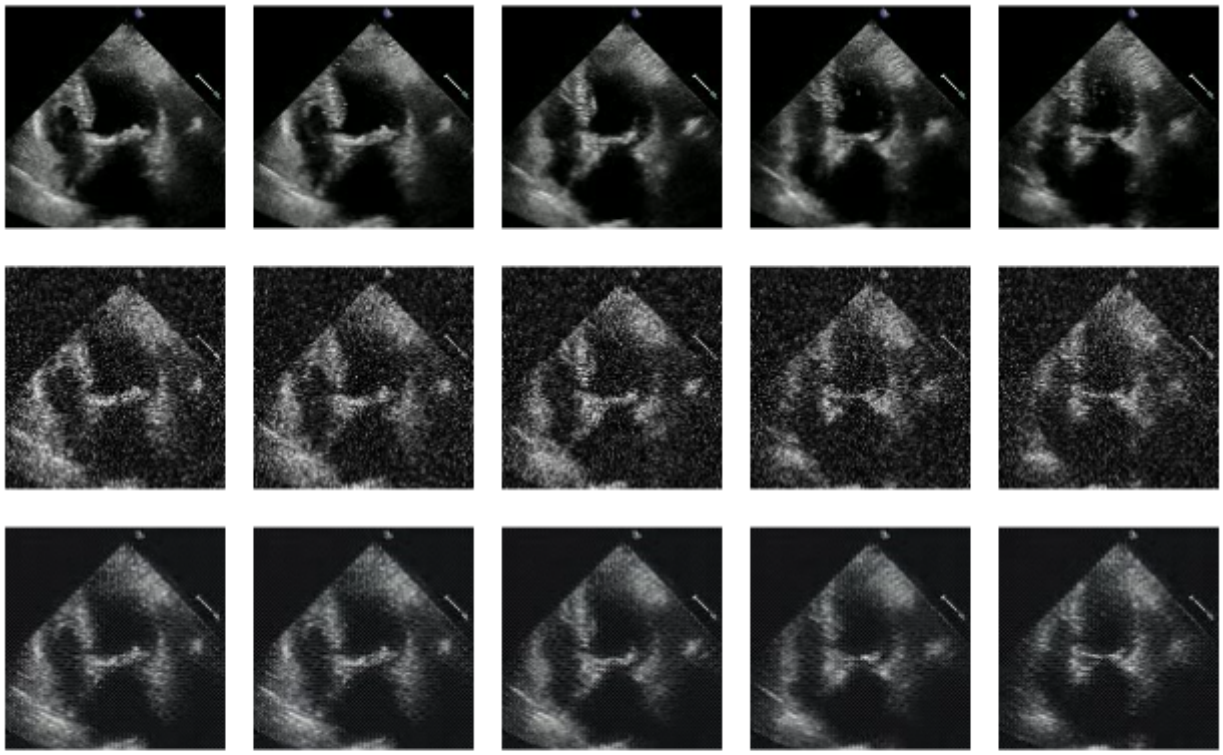


Figure 1. Original ground-truth video frames from a sample echocardiogram in the Stanford EchoNet-Dynamic dataset (top). The video frames are corrupted with random amounts of salt & pepper, gaussian, and speckle noise, which are commonly found in ultrasound images (middle). The VQ-GAN learns a latent representation for the frames which can be reconstructed back into noise-free frames (bottom).

[OR40] Improving Mortality Prediction in People with Cardiovascular Disease: A Random Survival Forests Approach Integrating Frailty Assessment

Jack Quach, Dalhousie University

Dustin Scott Kehler, Dalhousie University

Olga Theou, Dalhousie University

Joanna M Blodgett, University College London

Introduction

Frailty is a state of increased vulnerability characterized by accumulation of health deficits across physiological systems. The deficit accumulation-based frailty index (FI) is a widely used frailty measure that strongly predicts mortality and adverse outcomes in individuals with cardiovascular disease (CVD). However, FIs weight each deficit equally and do not consider potential interactions between deficits. Random survival forests (RSF) models offer a way to model nonlinear relationships and interactions between deficits to optimize mortality predictions. We evaluated whether RSF models with individual FI items better predicted mortality than RSF models with only the composite FI score.

Methods

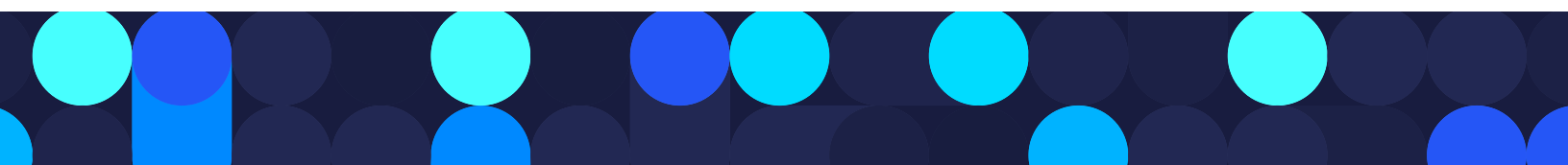
Individuals with a CVD history (n=2918) from the 1999-2015 National Health and Nutrition Examination Survey were included. A 46-item FI was used to measure frailty. We trained RSF models to predict all-cause mortality and CVD mortality with 1) individual FI items and 2) composite FI score. Hyperparameters were tuned using a random grid search and 5-fold cross-validation. Permutation variable importances were used to determine which deficits were most predictive of mortality. C-index was used to evaluate model performance (greater C-index indicates better mortality discrimination).

Results

The mean (standard deviation) age of the sample was 66.4 (13.2), 41.9% (n=1224) of whom were female. The RSF model with 46 individual FI items was better at predicting mortality (C-index: 0.708 all-cause; 0.707 CVD) than the model with the FI score (C-index: 0.651 all-cause; 0.638 CVD). Variations in pulse pressure had a large impact on ability to accurately predict both all-cause and CVD mortality.

Discussion/Conclusion

Using individual FI items provided superior mortality prediction compared to using the composite FI in this CVD cohort. As such, the RSF is a potential method to optimize mortality prediction of the widely used FI by modeling complex relationships between deficits. Identifying key deficit drivers of mortality may enable targeted frailty interventions.



Supporting information

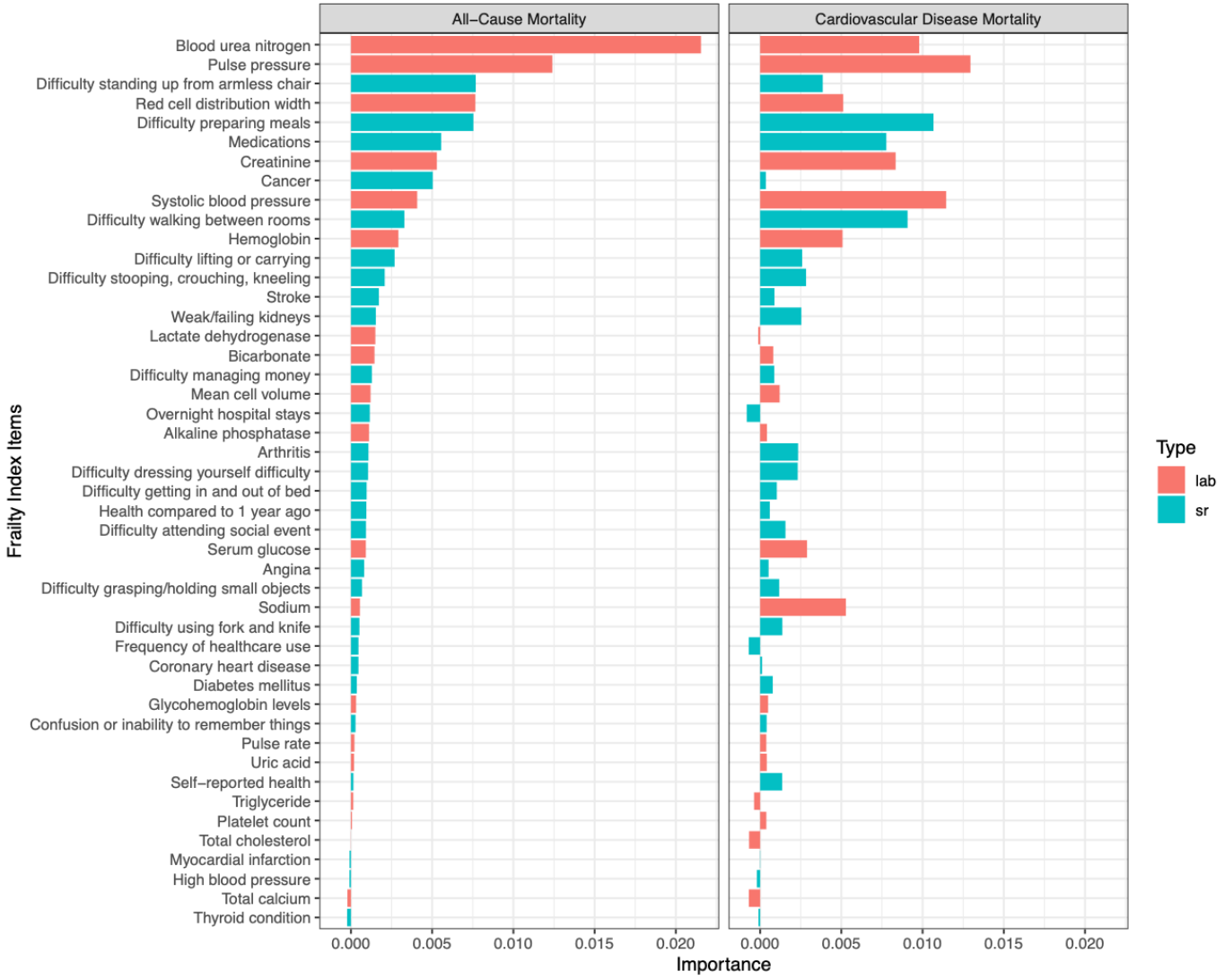


Figure 1. Permutation variable importances of the 46 frailty index items modelled by random survival forests. Permutation variable importances measure the impact of each variable in the mortality prediction model by observing the degree that prediction error increases when values of that variable are randomly shuffled. Lab test items coloured in red and self-report items are in blue.

[OR41] Novel Approaches in 12-Lead Electrocardiogram Signal Reconstruction

Yan Zhu, University of Toronto
Chunsheng Zuo, University of Toronto
Guanghan Wang, University of Toronto
Yunhao Qian, University of Toronto
Christopher Cheung, University of Toronto

Introduction

Reconstructing the complete 12-lead electrocardiogram (ECG) signals from fewer leads can enable ECG diagnostics when limited lead data is available. While many prior works have proposed machine learning (ML) models for signal synthesis, none have discussed what ML models should be used in various clinical settings. Furthermore, although many models, such as linear regression and convolutional neural networks (CNN), have been evaluated, little attention has been given to novel models such as U-Net and transformers.

Methods

We investigate seven ML models and their performances of ECG signal reconstruction on two large public datasets: PTB-XL and CODE-15%. Among these models, three of them are widely used in prior works, and the rest are more novel and have not been well-studied for ECG reconstruction yet. We evaluate the models in terms of reconstruction accuracy and computational costs and provide a complete comparison in ECG reconstruction performance.

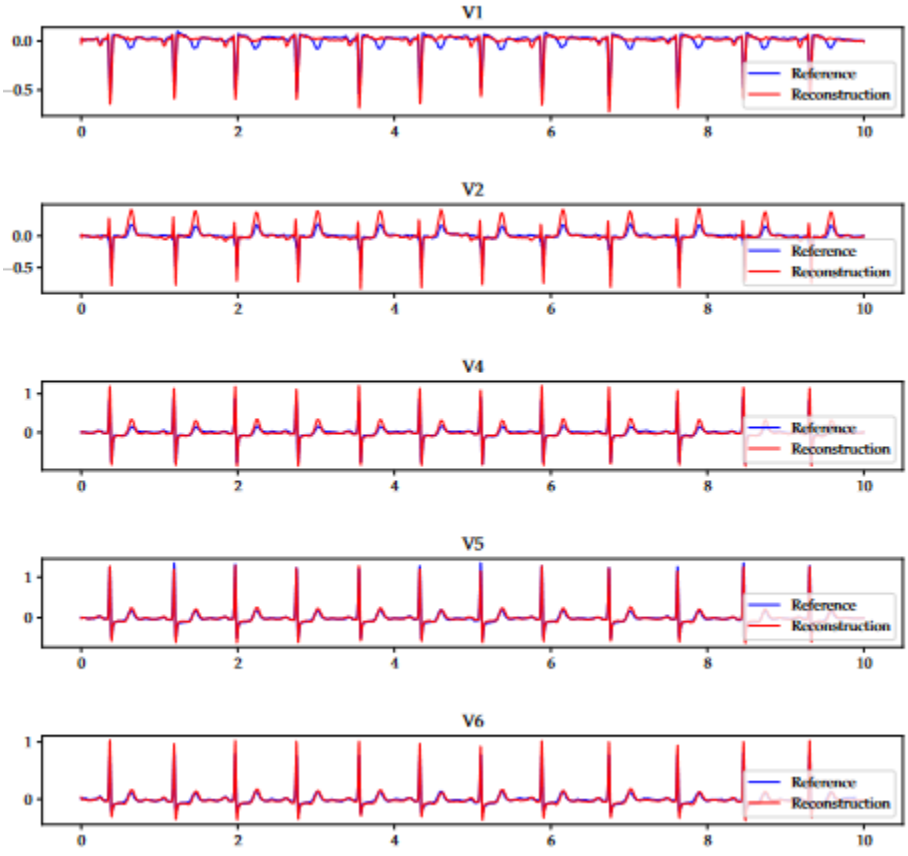
Results

All models generally performed well using the PTB-XL and CODE-15% datasets. In the PTB-XL dataset, the best performance was achieved using the CNN-long short-term memory (CNN-LSTM) model, with an RMSE of 0.1051 and Pearson R coefficient of 0.9297. In the CODE-15% dataset, the best performance was achieved using the LSTM model, with an RMSE of 0.1832 and Pearson R coefficient of 0.9112. An example of ECG reconstruction is provided in Figure 1.

Discussion/Conclusion

We conclude that CNN-based models performed better than other deep learning algorithms when the dataset is clean and noise-free, while LSTM-based models achieve the best reconstruction results when the data is noisy and/or data denoising is not accessible. Transformer models, which are usually considered the better alternatives of LSTMs, do not show clear advantages for ECG signal synthesis, but may have various opportunities for future research.

Supporting information



[OR42] Obstacle detection for persons with visual impairments using AI-powered smart glasses

Haining Tan, University of Toronto

Andrew Garrett Kurbis, University of Toronto

Alex Mihailidis, Department of Occupational Science and Occupational Therapy

Brokoslaw Laschowski, University of Toronto

Introduction

There are over 2.2 billion people worldwide with visual impairments. Smart glasses with integrated cameras could help provide sensory feedback and restore vision during activities of daily living. These technologies could be especially useful for detecting objects or obstacles that present potential safety risks during navigation.

Methods

In this study, we developed a stair recognition system powered by deep learning and deployed our model on a pair of smart glasses with onboard perception and computation. Our proof-of-concept focused on stair recognition because of the high risk of injury if users do not recognize the stairs themselves. We trained and optimized a lightweight and efficient convolutional neural network using ~515,000 images from StairNet, the largest open-source image dataset of real-world stair environments. We fine-tuned our deep learning model using ~7,250 manually labelled images from the Meta Ego4D dataset, which was collected using head-mounted cameras. Our model was compiled using TensorFlow Lite Micro and deployed on the smart glasses for onboard real-time inference.

Results

Our model was able to detect complex stair environments with 98.2% accuracy, with an average inference speed of 1.47 seconds on the embedded device.

Discussion/Conclusion

Our AI-powered smart glasses serve as a first step towards the development of a sensory feedback system for persons with visual impairments. The output of our glasses (i.e., information about the walking environment) could communicate with users via audio or haptic feedback, or eventually through neuromodulation to directly interface with the visual cortex.

[OR43] Quality of Interaction Between Clinicians and Artificial Intelligence. A Systematic Review

Argyrios Perivolaris , University of Toronto
Robert Chris Adams-McGavin, University of Toronto
Yasmine Madan, McMaster University
Tony Antoniou , University of Toronto
Muhammad Mamdani, University of Toronto
James J. Jung , University of Toronto

Introduction

Artificial intelligence (AI) has the potential to improve quality of healthcare when thoughtfully integrated into clinical practice. However, many clinicians remain hesitant towards its adoption. A crucial factor affecting successful adoption is the quality of interactions between AI solutions and clinicians. Current evaluations of AI solutions tend to focus solely on model performance. Thus, there is a critical knowledge gap in the assessment of AI-clinician interactions. We aimed to synthesize existing literature to identify interaction traits that can be used to assess the AI-clinician quality of interactions.

Methods

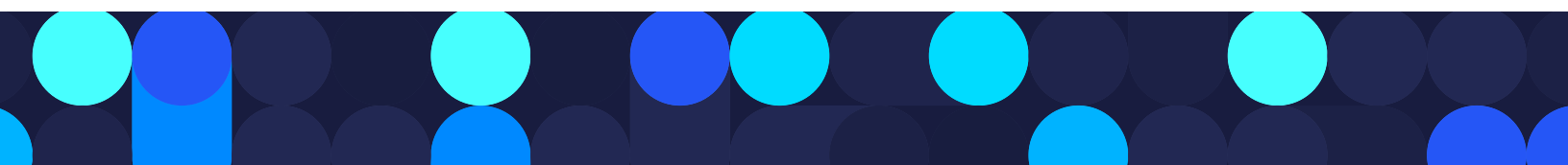
We performed a systematic review of published studies to June 2022 using OVID Medline, PsychINFO, Embase, and Scopus databases that reported interaction traits between clinicians and AI-enabled clinical decision support systems. Due to study heterogeneity, we conducted a narrative synthesis of the different interaction traits. Two authors categorized AI-clinician interaction traits based on their shared constructs.

Results

From 34 included studies, we identified 210 interaction traits. The most commonly reported interaction traits included usefulness, ease of use, trust, satisfaction, willingness to use, and usability. After removing duplicate or redundant traits, 90 unique interaction traits were identified. Unique interaction traits were then classified into seven categories: usability, system performance, clinician trust and acceptance, impact on patient care, communication, ethical and professional concerns, and clinician engagement and workflow.

Discussion/Conclusion

We identified seven categories of interaction traits between clinicians and AI systems. The proposed categories may serve as a foundation for a framework assessing the AI-clinician quality of interactions.



[OR45] A User-Centered Design Approach to an Artificial Intelligence-Enabled Electronic Medical Record in Canadian Primary Care

Krizia Francisco , University of Waterloo

Puneet Seth , TELUS Health

Sukhman Tamber , McMaster University

Introduction

Primary care physicians are at the forefront of the clinical process that can lead to diagnosis, referral, and treatment. Their ability to navigate the clinical interaction whilst using technology, and level of engagement can profoundly impact the patient experience. As Electronic Medical Records (EMRs) have become a more integrated part of primary care delivery, limitations in their usage and functions have been recognized to be a contributor to physician burnout. This demonstrates that EMRs have the potential to be optimized as a tool that can support care delivery. With these learnings, it's important to highlight how we can optimize and design AI in the primary care setting to ensure its ultimately supporting primary care delivery. The primary objective of this research is to understand if the provider-centered design approach, rooted in contextual design, can enhance the use of an AI-enabled tool embedded in the primary care EMR.

Methods

An industry partnership has been established with TELUS Health, to use their EMR, the Collaborative Health Record (CHR).

A total of 5 phases have been designed for this study.

• Phase 1 will interview primary care physicians to understand their current workflow in the traditional CHR using pre-appointment questionnaires and encounter templates,

• Phase 2 will interview the same primary care physician on the use of an AI-enabled CHR using enhanced pre-appointment questionnaires and enhanced encounter templates,

• Phase 3 focuses on a qualitative analysis of the data collected in the interviews to develop user-centered requirements,

• Phase 4 is the re-design of an AI-Enabled CHR based on the data analyzed in phase 3 and;

• Phase 5 is the validation of the new designs in a secondary interview with the primary care physicians.

Results

Preliminary results will be available by the conference.

Discussion/Conclusion

Preliminary results will be available by the conference.

[OR46] Assessing Prognosticators of Intracranial Metastatic Disease in Patients With HER2+ Breast Cancer Leveraging Supervised Machine Learning Algorithms

Marco V. Istasy, Faculty of Medicine, University of Toronto

Sunit Das, Division of Neurosurgery, University of Toronto

Introduction

Intracranial metastatic disease (IMD) is an increasingly common and devastating complication of breast cancer primaries which decimates both patient quality of life and overall survival. Despite the perniciousness of this disease, there is a lack of a formalised and cogent prognostic algorithm whereby the likelihood of IMD development may be interrogated as a function of patient- and tumour-specific features.

Methods

Six unsupervised machine learning algorithms (Decision Tree (DT), Random Forest (RF), CatBoost (CB), XGBoost (XGB), LightGBM (LGBM), and Feed Forward Neural Network (FNN)) were trained, tested, and validated on a multi-institutional dataset of over 8,000 patients diagnosed with HER2+ breast cancer between January 1, 2014 and March 1, 2021 at Sinai Health System and Sunnybrook Health Sciences Centre. Individual model capacity was robustly evaluated using calibration analyses, Brier scores, and area under the receiver operator characteristic curve (AUC) and the Delong test was applied to assess statistically significant differences across model AUCs. To elucidate model epistemology, all algorithms were interrogated using Shapley Additive Explanations (SHAP) and tree-based models were visually inspected.

Results

All models except the DT exhibited acceptable calibration ($r > 0.95$), accuracy (Brier scores 0.75). The Delong test demonstrated statistically significant differences in discrimination between four model groups: 1) DT, 2) RF, 3) CB, XGB, LGBM, and 4) FNN ($p < 0.05$; organised in order of increasing AUC). SHAP interrogation revealed concordance in feature importance among all six models.

Discussion/Conclusion

Five of the six algorithms employed in our analysis demonstrated acceptable discriminative ability in predicting the development of IMD from breast cancer primaries. Furthermore, epistemic insight into each algorithm uniquely allows for employment in a clinician-in-the-loop paradigm, whereby follow-up care and preventive strategies for this patient population may be modulated by realisations from such algorithms.

[OR47] Bio-inspired modulation to enhance the robustness of deep learning models against healthcare data quality problems

Mohamed Abdelhack, Krembil Centre for Neuroinformatics
Jiaming Zhang, University of California, San Diego
Sandhya Tripathi, Washington University in St. Louis
Bradley Fritz, Washington University in St. Louis
Daniel Felsky, CAMH/Dalla Lana School of Public Health
Michael S Avidan, Washington University in St. Louis
Yixin Chen, Washington University in St. Louis
Christopher R King, Washington University in St. Louis

Introduction

Data quality is an issue of concern in machine learning model deployment which limits its application in high-stake fields such as healthcare. Training datasets are usually well-curated and of high quality; however, in production environments, missing data is a common occurrence.

Methods

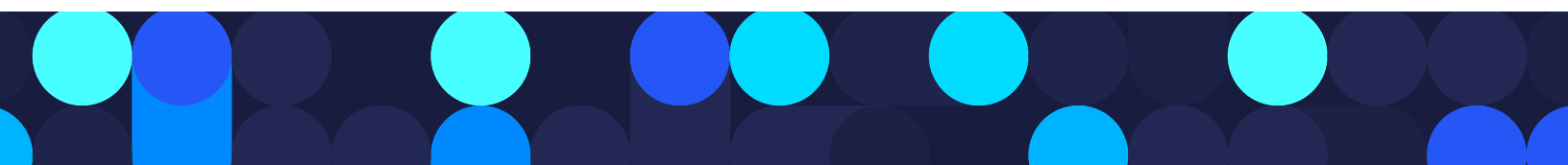
This study introduces a new neural network architecture that allows the weights of a fully-connected layer to be modulated in real time via an additional input. This takes its inspiration from modulatory neurons that can alter the behavior of a neuron by up- or down-regulating inputs. We tested these layers on medical datasets with missing data flags using three different missingness paradigms (random, quantile, and feature missingness) and data quality measures as the modulation input.

Results

Results showed the resulting networks to be more robust against increasing degradation of data quality at the test phase. This addition has advantages over imputation as it skips the imputation process and trains the model end-to-end. It also allows the introduction of other data quality measures that imputation cannot handle. Additionally, it showed superior performance to the addition of data quality and missingness flags at the input as it allows for more complex interactions between the data and the quality measures such as accounting for data missing-not-at-random.

Discussion/Conclusion

These results reveal that the incorporation of modulation layers with data quality measures allows us to enhance the robustness of the neural network models thus allowing their use in applications with frequent data quality issues such as healthcare.



[OR48] Deep Learning-Enabled Fluorescence Quantification for Surgical Guidance: Benefits Over Analytical Methods

Anjolaoluwa Adewale , Princess Margaret Cancer Centre
Natalie J. Won , Princess Margaret Cancer Centre
Jerry Wan , Princess Margaret Cancer Centre
Mandolin Bartling , Department of Otolaryngology – Head & Neck Surgery, University of Toronto
Alon Pener-Tessler , Department of Otolaryngology – Head & Neck Surgery, University of Toronto
Esmat Najjar , Department of Otolaryngology – Head & Neck Surgery, University of Toronto
Jonathan C. Irish , Department of Otolaryngology – Head & Neck Surgery, University of Toronto
Brian C. Wilson , Princess Margaret Cancer Centre
Michael J. Daly , Princess Margaret Cancer Centre

Introduction

Intraoperative fluorescence imaging is an emerging technique to guide cancer surgery. Clinical imaging devices, however, are limited to qualitative assessments that depend not only on fluorophore uptake but other factors including optical properties and surface topography. As such, there is an unmet need to develop imaging devices that can quantify in vivo fluorescence concentration, as this may help surgeons answer the question: is this tumor or healthy tissue? This simulation study is a first step in fulfilling this need as it investigates the performance of a novel deep learning (DL) approach to fluorescence quantification, in comparison to traditional analytical methods.

Methods

A Siamese convolutional neural network (Fig 1.a) was trained with 10,000 fluorescence images (65x65 pixels) of composite spherical harmonics (CSH) varying in width and height, along with corresponding optical property maps. These images were generated using a numerical tissue simulator modeling near-infrared light propagation. The DL model, (~1.9 million parameters) underwent 2 hours of training. Subsequently, the model's performance was assessed using 11 fluorescence images of CSH with varying widths. These results were compared to those obtained from an analytical model based on diffusion theory, computed by dividing fluorescence by a function of optical properties.

Results

The DL model (MSE=0.44) outperformed the analytical model (MSE=4.12) in estimating concentration. Notably, the analytical model consistently underestimated fluorescence concentration (Fig 1.c).

Discussion/Conclusion

This study provides preliminary insights into a novel DL algorithm's accuracy in estimating fluorescence concentration from in-silico fluorescence images. Ongoing efforts are underway to validate these findings experimentally in more realistic settings (e.g., phantoms, animals).

Supporting information

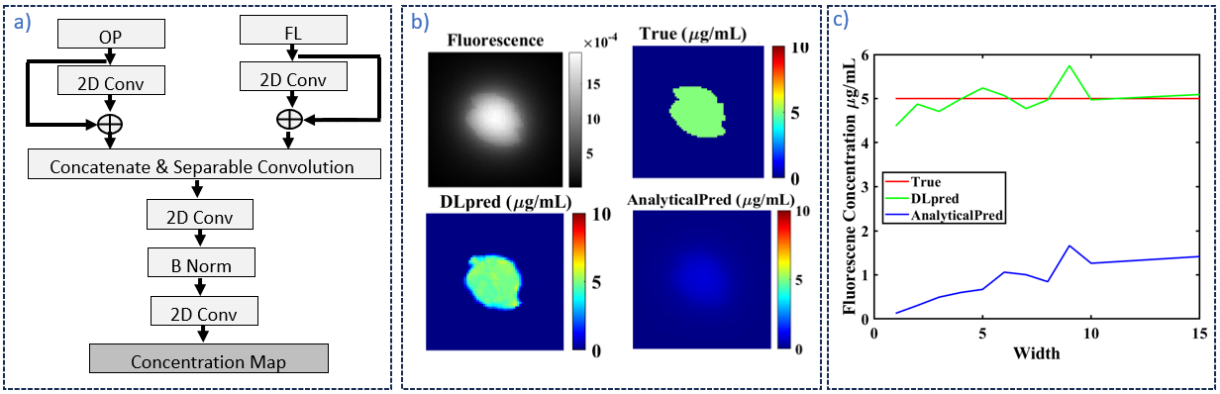


Figure 1: a) Siamese Deep Learning architecture b) Fluorescence image of 8 mm simulated tumor with estimated and true concentration maps c) Estimated concentrations across width.

[OR49] Development of an objective framework to optimize machine learning-based single-cell segmentation accuracy for multiplexed tissue cytometry

Trevor D. McKee , Pathomics.io

Mark Zaidi , University of Toronto

Phoebe Lombard , University of Toronto

Justin Grant , Pathomics.io

Introduction

Multiplexed immunostaining methods enable “tissue cytometry” – flow cytometry-like analysis on tissue sections, but single-cell cytometric analysis needs reliable cell segmentation. No common framework to quantify segmentation errors exists; this remains a major challenge, as cellular segmentation errors propagate into spurious findings in resultant single-cell data; with unknown error magnitude.

Methods

This framework permits objective assessment of single-cell segmentation within multiplexed immunohistochemistry datasets, alongside methodologies for visualization of the resultant accuracy against the gold standard of manual annotations by an expert observer. One set of manual annotations can compare multiple different segmentation strategies; highlighting segmentation errors (over-, under-, or missed-segmentation).

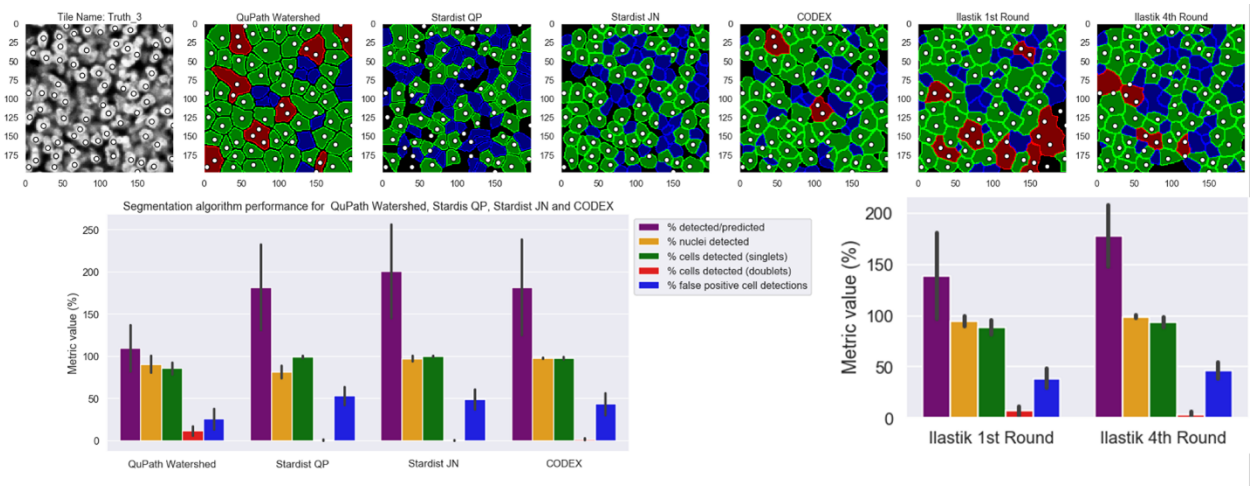
Results

Manual annotations, single points placed within nuclear intercalator signal in “ground truth” image, are compared to computer vision-based approaches (QuPath Watershed and CODEX MAV), machine learning approaches (Ilastik, with two different rounds of training); and deep-learning approaches (StarDist, implemented in QuPath or via Jupyter Notebook). Visualized in figure are: a 1:1 correspondence between annotation and segmentation (green); over-segmentation (breaking one cell into more than one segmentation object) produces cell fragments with no corresponding annotation (blue); while under-segmentation, more than one annotation per segmentation (red); and in some cases manual annotations with no corresponding segmentations appear as solo annotation dots.

Discussion/Conclusion

Quantification of the primary measures of segmentation accuracy calculated as: percent predicted (number of segmented cells / number of annotated points), percent nuclei found (nuclei found inside a cell mask / all annotated nuclei), percent of singlets (cell masks with only one nuclei annotation / total number of cell masks with at least one nuclei annotation); and the percent of mis-segmentations (doublets / under-segmentation events, and false positives / over-segmentation events). Optimal strategy may depend on the question being asked; but at least accuracy can be objectively assessed across multiple annotated image patches and segmentation methodologies.

Supporting information



[OR50] Improving Patch-based Segmentation for Pediatric Cancer Detection

Abhishek Moturu, University of Toronto, Vector Institute, SickKids

Sayali Joshi, University of Toronto, SickKids

Andrea Doria, University of Toronto, SickKids

Anna Goldenberg, University of Toronto, Vector Institute, SickKids

Introduction

Identifying early signs of cancer in whole-body pediatric MRI images is a challenging task for various reasons - whether it be the size of the tumour compared to the size of the whole-body MRI, the rarity of cancer cases, the heterogeneity of tumour type, shape, and location, the age and sex of the child, the differences in the MRI machines, the noisiness of the MRI scans, or the variations in radiologists' labels.

Methods

We utilize a model based on U-Net to perform segmentation with several metrics in mind: tumour area with respect to data imbalance and foreground and background pixel imbalance and tumour boundary. The model was trained and validated on patches from 675 whole-body MRI volumes, with ground truth tumour segmentation labels obtained from our collaborating radiologists. Augmentations and synthetic data were also used to increase the amount of data with tumours.

Results

We perform analysis to study the strengths and weaknesses of our model based on tumour size, brightness, and location within the body and also report the results of the radiologists' evaluation of our cancer segmentation model to see if it helps reduce the time, increase the accuracy, or improve radiologist experience.

Discussion/Conclusion

This work stresses the importance of properly utilizing patch-based methods and the common pitfalls to consider when using these methods on medical imaging data. The effect on radiologist performance with and without our methods highlights the need for better clinically-relevant metrics and design to incorporate AI-based tools into the diagnostic practice. We emphasize the importance of clinician input for the safe and effective deployment of diagnostic models in healthcare applications.

[OR51] Innovative Detection of Diabetes Stigma in Digital Spaces with Large Language Models

Somayeh Amini , University of Toronto
Mark Dayomi, IHPME/University of Toronto
Zahra Shakeri , IHPME/Dalla Lana School of Public Health/University of Toronto
Tanav Bajaj , HIVE Lab, IHPME, Dalla Lana School of Public Health, U of T
Aryan Sadeghi , HIVE Lab/IHPME/Dalla Lana School of Public Health/University of Toronto
Shengjie Tony Zou , HIVE Lab, IHPME, Dalla Lana School of Public Health, University of Toronto

Introduction

Diabetes-related stigma presents a significant burden to affected individuals, impacting their well-being and disease management. With the growing prominence of online platforms as go-to health information sources, addressing this stigma within these digital environments is crucial. This study explores the potential of Large Language Models (LLMs) in detecting and assessing the reach of diabetes-related stigma across diverse digital sources and conversations. The overarching aim is to create an open-source and publicly accessible platform to highlight the nuances of this ongoing issue, providing potential strategies to reduce the harmful effects of stigma associated with diabetes.

Methods

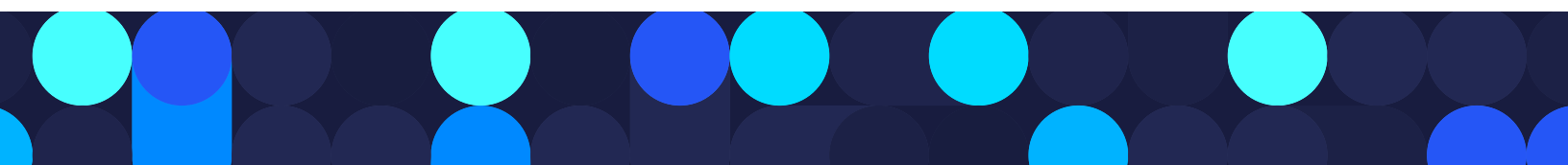
In our study, we collected substantial data from online sources, such as social media platforms and Q&A websites, in partnership with Diabetes Canada. This data was then analyzed using ChatGPT, a well-established LLM. Additionally, we created a web system to illustrate the extent of diabetes-related stigma in digital conversations. To validate our findings, we involved domain experts, confirming the results obtained from the model. This approach supports our objective to understand and address diabetes-related stigma in the digital space.

Results

Our results demonstrated an initial agreement of 77.91% between domain experts and ChatGPT in identifying non-stigmatized comments, while detection of stigmatized comments showed lower concordance at 40%. Through improved prompt engineering and fine-tuning of the operational definition, we managed to enhance the agreement rate to 84% for non-stigmatized and 70% for stigmatized comments. Moreover, we have made our web-based platform publicly available, serving as a useful resource for the general public and researchers interested in delving further into the study of diabetes-related stigma.

Discussion/Conclusion

LLMs can effectively detect diabetes-related stigma across diverse digital sources. With comprehensive web scraping and careful refinement of prompts and operational definitions, these models show improved accuracy in stigma detection. Their application can contribute to supportive, stigma-free online environments for individuals living with diabetes.



[OR52] Multi-task Machine Learning of the Electronic Medical Record Predicts Future Symptoms among Cancer Patients

Baijiang Yuan , Princess Margaret Cancer Centre - UHN
Muammar M. Kabir , Princess Margaret Cancer Centre - UHN
Kevin He , University Health Network
Benjamin Grant , Princess Margaret Cancer Centre - UHN
Sharon Narine, Princess Margaret Cancer Centre - UHN
Wei Xu, Princess Margaret Cancer Centre, UHN
Rahul G Krishnan, Department of Computer Science - UofT
Tran Truong, Princess Margaret Cancer Centre - UHN
Geoffrey Liu, Princess Margaret Cancer Centre, UHN
Robert C Grant , Princess Margaret Cancer Centre - UHN

Introduction

Cancer and its treatments often cause symptoms. Automated warning systems could mitigate symptoms by alerting healthcare teams and enabling personalized preventative interventions. We developed a general-purpose longitudinal system for predicting symptomatic deterioration among outpatients undergoing intravenous systemic anti-cancer therapy

Methods

Patients treated for aerodigestive cancers at the Princess Margaret Cancer Centre were randomly divided into development and testing cohorts. For each treatment, machine learning was applied to preceding electronic medical record (EMR) data to predict patient-reported symptom deterioration, defined as at least a four point worsening on the Edmonton Symptom Assessment Scale. Features included diagnostic and treatment characteristics, laboratory tests, and patient-reported symptoms. Single-task (e.g., LASSO and XGboost) and multi-task (e.g., temporal CNNs, LSTM and Transformer) models were trained, tuned, and evaluated based on discrimination, calibration, and net benefit

Results

The cohort consisted of 3,998 patients who underwent 45,904 treatment sessions, with data across 400 features. Among these patients, 1,547 (38.6%) were female; median age was 64.0 (interquartile range 13.0). The most common diagnoses were lung (1,505, 37.6%), head and neck (696, 17.4%), and pancreatic cancers (685, 17.1%). The best model, a multi-task transformer, predicted symptom deterioration with an AUROC range of 0.732-0.822, marking a 1.4-6.2% improvement over the best single-task model. At a 10% alert rate, treatments associated with alerts would be enriched 4-13 fold for symptom deterioration ($P < 0.001$). The system was calibrated and would provide a net benefit across a wide range of threshold probabilities in decision curve analysis

Discussion/Conclusion

Longitudinal general-purpose multi-task machine learning systems trained using EMR data can accurately predict a wide range of symptoms. Based on these results, automated warning systems for symptoms should be implemented and evaluated in real-time clinical practice to guide preventative interventions

[OR53] Rapid Evolution of Artificial Intelligence in Medicine: Comparative Analysis of ChatGPT-3.5 and ChatGPT-4 in Generating Clinician-level Vascular Surgery Recommendations

Arshia Javidan, University of Toronto
Tiam Feridooni, University of Toronto
Lauren Gordon, University of Toronto
Sean Crawford, University of Toronto

Introduction

ChatGPT-3.5, an AI language model has demonstrated remarkable clinical potential across several domains and applications. Its newer version, chatGPT-4, was released only several months later and outperforms its predecessor across almost all benchmarks. Comparison of the two language models and evaluations of these models in surgical fields remain sparse. We compared the capability of both language models in providing clinician-level vascular surgery recommendations.

Methods

Forty questions related to four distinct areas of vascular surgery were generated by experts and inputted into both AI models. Responses were independently evaluated using a 5-point scale, with scores 4 and 5 rated as “appropriate,” and 1-3 as “inappropriate” by two blinded reviewers. Independent t-tests and Fisher’s exact test were used for comparisons, while Flesch Reading Ease scores were used to assess readability.

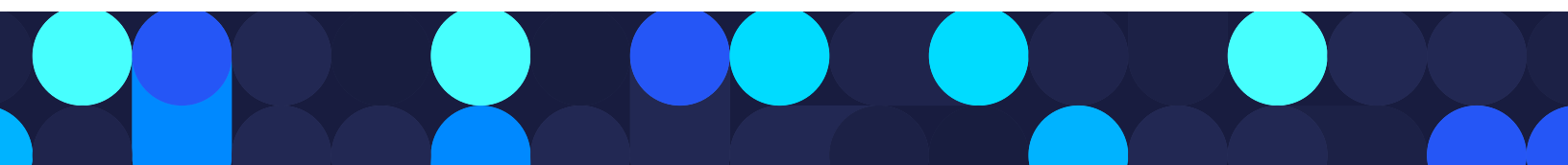
Results

ChatGPT-4 provided accurate recommendations almost three times more frequently than chatGPT-3.5 (95% vs. 32.5%, $p < 0.001$). ChatGPT-4’s responses were longer (317 vs. 265 words), with no difference in reading ease score between the models (17.8 vs. 19.4, $p = 0.547$), corresponding to college-level graduate texts.

Discussion/Conclusion

AI language models can offer accurate, clinician-level recommendations to complex vascular surgery prompts. Notably, the newer model’s accuracy improved nearly threefold only over the course of several months. These advancements underscore their potential as reliable clinical tools. Yet, their use requires expert validation due to occasional nonsensical or false content.

Notably, the achieved performance did not involve model fine-tuning on surgery-specific databases or prompt engineering, both of which can increase accuracy and readability. Rapid AI progression suggests it may soon outperform physicians in knowledge expertise, making discussions about its integration into medical education and practice essential.



[OR54] Temporal validation of SEPERA to inform nerve-sparing strategy during radical prostatectomy and comparison against expert surgeons

Lauren Pickel, University of Toronto
Jethro Kwong, University of Toronto
Kevin Zhang, University of Toronto
Ryan Booth, University of Toronto
Aiman Shahid, University of Toronto
Maximiliano Ringa, Trillium Health Partners
Amna Ali, Trillium Health Partners
Alistair E W Johnson, The Hospital for Sick Children
Andrew Feifer*, Trillium Health Partners
Alexandre R Zlotta*, Mount Sinai Hospital

Introduction

Extra-prostatic extension of prostate cancer is associated with worse oncological outcomes following RP. Functional outcomes after RP (urinary continence, potency) depend on the extent to which the neurovascular bundles lying adjacent to the prostate are preserved during RP. This nerve-sparing strategy involves a delicate balance between maximizing functional outcomes while minimizing risk of positive surgical margins. Therefore, accurate prediction of side-specific extra-prostatic extension (ssEPE) is essential to inform nerve-sparing strategy.

Methods

We previously developed and validated a machine learning model, SEPERA (Side-specific Extra-Prostatic Extension Risk Assessment). Temporal validation was performed on 695 patients undergoing RP at University Health Network and Trillium Health Partners from 2020-2022. Nerve-sparing strategy at time of surgery was compared to SEPERA's recommendations. Since SEPERA only outputs risk of ssEPE, thresholds that achieved 95% sensitivity (0.16) and 95% specificity (0.57) on the original dataset were used to generate recommendations for "Complete" (risk 0.57) nerve-sparing. The proportion of cases with ssEPE and positive margins (ssPSM) on surgical pathology was determined.

Results

ssEPE was found in 467 out of 1390 prostatic lobes (34%). SEPERA performed similarly compared to the original study with AUROC 0.75 (95% CI: 0.73-0.78) and AUPRC 0.63 (95% CI: 0.58-0.67). SEPERA's recommendations differed from clinical decisions in 49% of cases (Figure 1). Where SEPERA recommended "Minimal" nerve-sparing but a greater degree of nerve-sparing was performed, 52% of cases had ssEPE and 35% had ssPSM. Conversely, where SEPERA recommended "Complete" nerve-sparing but less was performed, only 13% of cases had ssEPE and none had ssPSM.

Discussion/Conclusion

In both academic and community settings, SEPERA improves current practice and pathological outcomes by accurately predicting ssEPE. SEPERA can inform nerve-sparing strategy during RP, especially when "Minimal" nerve-sparing is recommended.

Supporting information

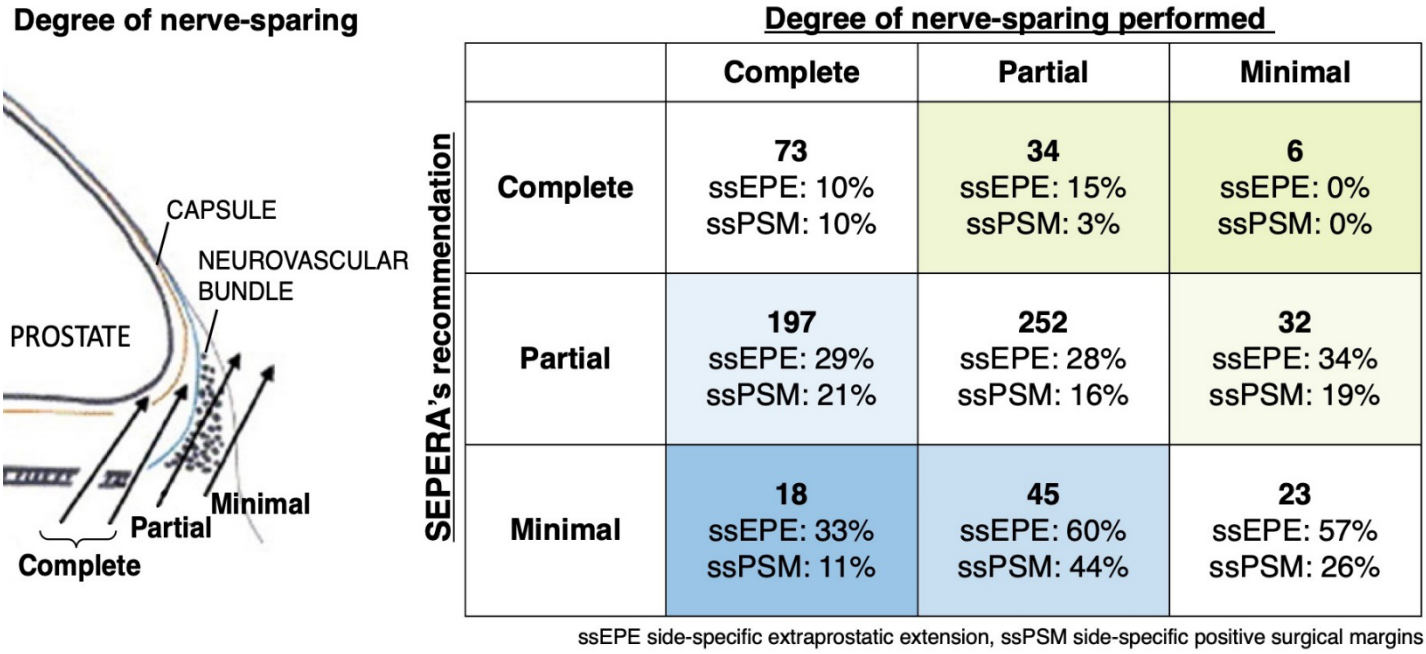


Figure 1. Left: Illustration of degree of nerve-sparing (Minimal ≤40% spared, Partial 40-80%, Complete ≥80%). Right: Comparison of SEPERA's recommendation for degree of nerve-sparing to actual clinical decision made during radical prostatectomy. Green represents cases where SEPERA recommends a greater degree of nerve-sparing than what was performed, and blue cases where SEPERA recommends less nerve-sparing (i.e., wider resection) than what was performed. The number of side-specific cases is shown in bold. The percentage of cases in each category found on surgical pathology to have side-specific extra-prostatic extension (ssEPE) and side-specific positive surgical margins (ssPSM) is given below.

[OR55] Addressing class imbalance with data augmentation when training deep learning models to identify high quality clinical studies

Cynthia Lokker, McMaster University
Elham Bagherie, McMaster University
Wael Abdelkader, McMaster University
Rick Parrish, McMaster University
R. Brian Haynes, McMaster University
Alfonso Iorio, McMaster University
Muhammad Afzal, Birmingham City University

Introduction

Only ~1% of PubMed articles are high quality and ready for practice. We recently trained BioBERT-based models to classify clinically relevant articles by methodologic quality using a dataset of 160,071 articles characterized by a 4:1 class imbalance. Our selected model, trained with a balanced undersampling of the data, maintained 99% recall and achieved 70% specificity. Here we explore data augmentation approaches on model performance.

Methods

We used class weights, synonym substitution, backtranslations, and text summarization using pretrained language models to balance the dataset. We applied a grid search strategy to fine-tune BioBERT models across hyperparameter combinations. Performance was compared with models trained using the undersampled and original unbalanced datasets. Models with >99% recall and optimal specificity were selected.

Results

The Table shows that AUC was >0.96 for all selected models. Synonym replacement improved performance slightly but performance of models trained using other augmentation approaches did not differ from the undersampled, balanced dataset.

Discussion/Conclusion

Imbalance in datasets can result in poorly performing models. In our experiments, synonym replacement was the most promising approach, but overall, augmentation did not greatly improve specificity. Future research will investigate the effects of adjusting class ratios and size of the dataset on model performance.

Supporting information

Table. Performance of BioBERT models trained using data augmentation in a hold-out validation dataset of 16,071 articles.

Dataset configuration	Performance (95% CI)			
	Specificity	Accuracy	Precision	AUC
Undersampled dataset (reference standard)	70.2% (69.4-71.0)	75.5% (74.8-76.2)	42.8% (41.6-44.0)	0.97
Unbalanced dataset	66.6% (65.8-67.4)	72.6% (71.9-73.3)	40.1% (38.9-41.2)	0.97
Class weights	71.5% (70.7-72.3)	76.6% (75.9-77.2)	43.9% (42.7-45.1)	0.97
Synonym replacement (PPDB)	72.3% (71.6-73.0)	77.2% (76.6-77.9)	44.6% (43.8-45.4)	0.98
Synonym replacement (Wordnet)	73.3% (72.7-74.0)	78.1% (77.3-79.7)	45.6% (44.8-46.3)	0.98
Backtranslated (French)	70.3% (69.6-71.0)	75.6% (74.9-76.3)	42.9% (42.1-43.7)	0.97
Backtranslated (German)	70.3% (69.6-71.0)	75.6% (74.9-76.3)	42.9% (42.1-43.7)	0.97
Text summarization	70.8% (70.1-71.5)	76.0% (77.4-78.7)	43.3% (42.5-44.1)	0.96

PPDB =Paraphrase Database

[OR56] Applications of Artificial Intelligence to Improve Patient Experiences During Care Transitions: An Informal Review

Nicole Bodnariuc, Trillium Health Partners

Terence Tang, Trillium health partners

Carolyn Steele Gray, Mt Sinai

Introduction

Inadequate hospital-to-home transitions can lead to medication errors, emergency visits, and rehospitalizations. These care transitions are complex, involving multiple parties who are often unfamiliar with each other. The Digital Bridge team is designing a digitally enabled care transition intervention from hospital to home. With the recent advances in artificial intelligence (AI), we are interested in exploring how AI can be used during care transitions to inform future iterations of our tool.

Methods

EMBASE and PUBMED, were searched from January 1, 2018, to July 1, 2023. We included peer-reviewed articles published in English that described the use of AI in hospital-to-home transitions to improve patient experience, provider experience, or care outcomes. One researcher reviewed the search results and applied the above inclusion criteria. The use of AI was then categorized according to intended technology function, type of AI technology, and the outcomes of the experience summarized.

Results

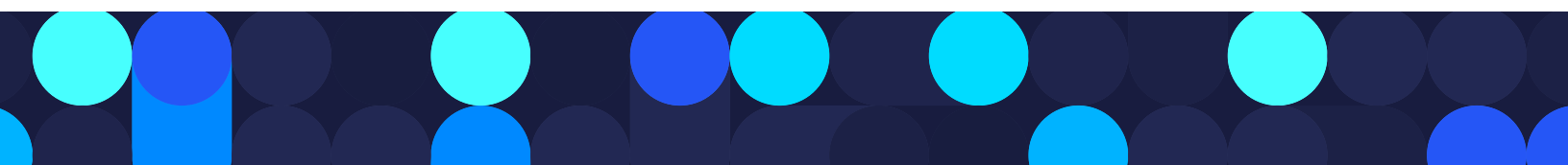
Thirteen studies examined AI-based techniques for predicting hospital readmission using clinical data. However, recent research underscores the importance of integrating social determinants of health data to better understand patients' lived experiences. Moreover, most of the AI models were disease-specific, limiting their broader applicability.

Bias in medical AI models is a significant concern that may impact clinical care. Despite this, most studies did not address bias detection and mitigation strategies. Additionally, many studies reported model performance using metrics like AUC, but practical patient-level solutions were lacking.

Three studies focused on using AI to generate prescriptions and discharge summaries, with one attempting to convert clinical concepts into patient-friendly language.

Discussion/Conclusion

AI has been used to predict those at risk of readmission using clinical data. There have been few reports on using generative AI for facilitating communication between clinicians or improving patient communication. Further investigations into these technologies may improve the care transition experience for patients and clinicians.



[OR57] Associations between Sex, Race, and Sedation in Invasively Ventilated Patients

Sarah Walker, University of Toronto
Federico Angriman, Sunnybrook Health Sciences Center
Lisa Burry, Mount Sinai Hospital
Leo Celi, Massachusetts Institute of Technology
Alistair Johnson, The Hospital for Sick Kids
Kuan Liu, Institute of Health Policy, Management and Evaluation
Sangeeta Mehta, Mount Sinai Hospital
George Tomlinson, Dalla Lana School of Public Health
Christopher J. Yarnell, University of Toronto

Introduction

Propofol and benzodiazepines are sedating medications used to facilitate care for patients receiving invasive ventilation. However, over-sedation is associated with worse outcomes. It is unknown whether sedation dose varies by patient sex or race and ethnicity.

Methods

We performed a retrospective cohort study of adults receiving invasive ventilation for over 24 hours using the Medical Information Mart for Intensive Care IV (MIMIC-IV, 2008-2019) database. We compared the association between patient sex and race and ethnicity (Asian, Black, Hispanic, white) and the sedation dose using a multilevel Bayesian proportional odds model that adjusted for baseline and time-varying covariates. Co-primary outcomes were the doses of propofol and benzodiazepine, secondary outcome was the observed level of sedation. We reported posterior odds ratios and 95% credible intervals [CrI].

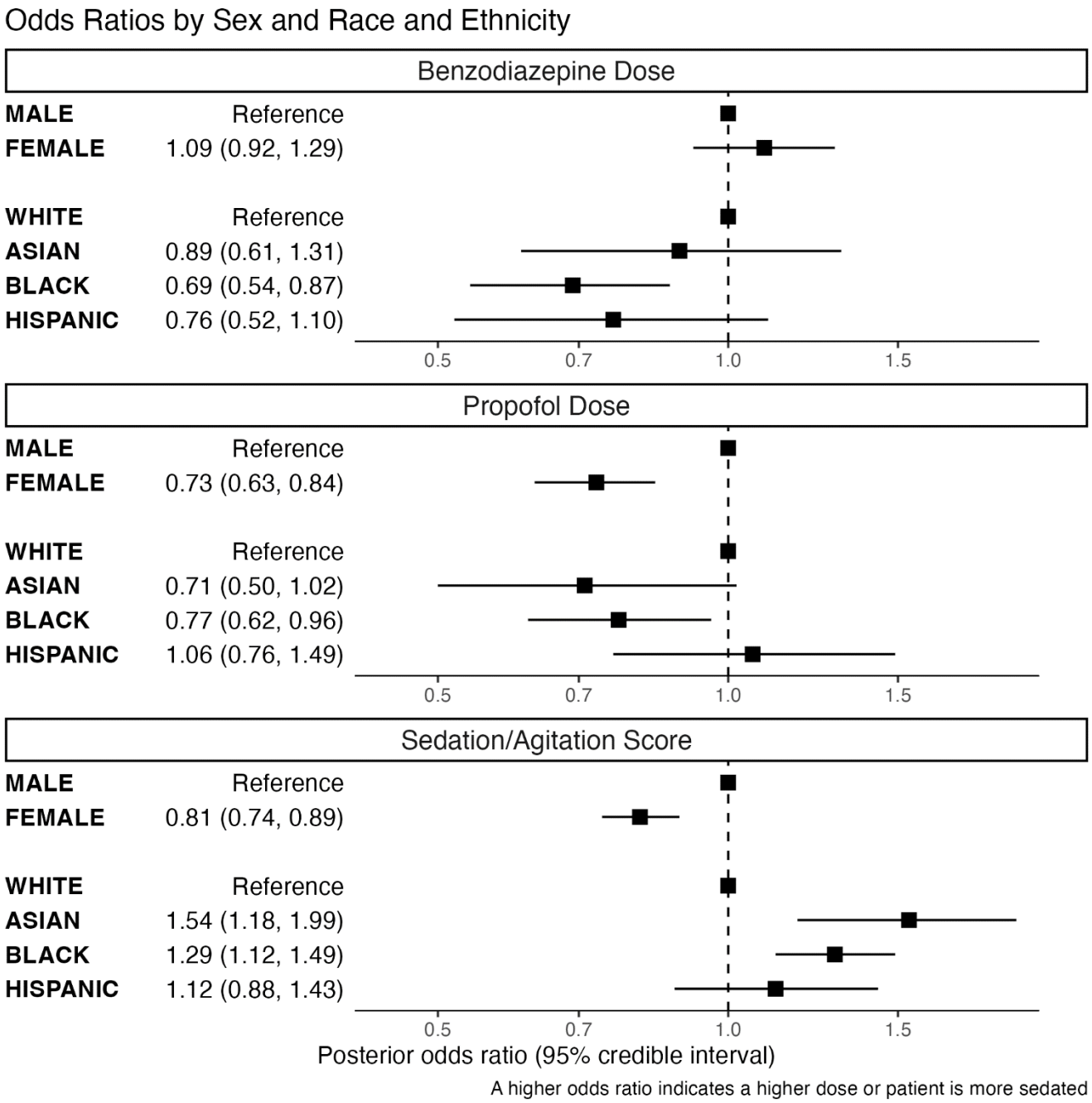
Results

We studied 10,617 patients from MIMIC-IV: 43% female (4,564); 3% Asian (356), 13% Black (1386), 5% Hispanic (494) and 79% white (8,381). The adjusted dose of propofol was lower in female (OR 0.73 [0.63 to 0.84]) compared to male patients, and Black (OR 0.77 [0.62 to 0.96]) compared to white patients. The adjusted dose of benzodiazepine was lower in Black (OR 0.69 [0.54 to 0.87]) compared to white patients. The observed level of sedation was higher in Asian (OR 1.54 [1.18 to 1.99]) and Black (OR 1.29 [1.12 to 1.49]) compared to white patients, and lower in female (OR 0.81 [0.74 to 0.89]) compared to male patients.

Discussion/Conclusion

Women received less propofol and were less sedated than men, suggesting an opportunity to reduce propofol doses for men. Compared to white patients: Asian patients were more sedated, implying propofol and benzodiazepine doses may be too high; Black patients received less propofol and benzodiazepines but were more sedated, which may reflect both protocol-driven sedation and excess propofol and benzodiazepines; and Hispanic patients had similar doses and levels of sedation.

Supporting information



[OR58] Challenges in Expert Labeling of Data to Leverage Machine Learning to Support Physiotherapy in the ICU

Adriana Ieraci, Laboratory Medicine and Pathobiology, Faculty of Medicine, University of Toronto

Vlad Porcilla, Faculty of Applied Sciences and Technology, Humber College

Maryam Davoudpour, Faculty of Applied Sciences and Technology, Humber College

Sunita Mathur, School of Rahabilitation Therapy, Queen's University

Kenneth Wu , Department of Medicine, University of Toronto; Department of Respirology, St. Michael's Hospital, Unity Health Toronto

Jane Batt, Department of Medicine, University of Toronto

Sharon Gabison, Department of Physical Therapy, University of Toronto

Alireza Sadeghian, Toronto Metropolitan University

Introduction

ICU-Acquired Weakness (ICUAW) is neuromuscular dysfunction characterized by muscle weakness and atrophy.(1) Patients who are older than 66 and mechanically ventilated for at least 2 weeks, are more likely to become dependent in their daily living post-ICU(2) due to ICUAW. Neuromuscular electrical stimulation (NMES) can maintain muscle mass and strength(3) but is resource-intensive to deliver in ICU and patients are not (always) able to provide feedback to therapists to adjust stimulation settings.

Surface electromyography (sEMG) and mechanomyography (MMG) are used to quantify muscle contraction and muscle fatigue.(4) If sEMG and MMG could be used in ML models to monitor muscle response to NMES, new tools would be possible to facilitate longer and more frequent sessions with current limited therapist resources and in the absence of patient feedback.

Methods

We built custom hardware and software to capture physiological sEMG/MMG data. Healthy subjects self-administered NMES over video chat while physiotherapy students remotely controlled data acquisition and visually assessed muscle contraction. ICU patients received NMES by a physiotherapist while sEMG/MMG was collected. A modified Medical Research Council Scale (mMRC) for Quality of Contraction (QoC) was used to label the sensor data for each contraction.

Results

We have collected sEMG/MMG data from 24 healthy subjects from the tibialis anterior (TA) muscle , and 10 ICU patients from the TA and rectus femoris. We have assessed the correlation of sEMG/MMG metrics (e.g., amplitude) with QoC labels. We are developing a deep learning pipeline with classification models to detect muscle contraction during NMES.

Discussion/Conclusion

Challenges in obtaining expert labeled data include; cost of resources, ensuring reliable labels with validated scales such as the mMRC, resolving time dependence of labels(6), obtaining data for all classes of contraction due to time constraints and insufficient muscle response. Future work will explore patient-specific and population-specific models and evaluating utility of the detection window.

[OR59] Clinical Outcome Prediction: Evaluating Quantization of Large Language Models

Raj Krishnan Vijayaraj, University of Toronto

MarK Chignell, University of Toronto

Lu Wang, Texas State University

Shurui Zhou, University of Toronto

Introduction

AI integration in medicine can revolutionize clinical outcome prediction using valuable patient histories. Transformer-based large language models (LLMs) show promise in predicting outcomes like mortality risk. However, LLMs require substantial computing resources. Quantization, representing model parameters with low-precision data, reduces the computational burden. Our research evaluates quantized LLMs for clinical outcome prediction, assessing the trade-off between accuracy and efficiency in clinical prediction to determine the viability of deployment.

Methods

The MIMIC-III dataset was chosen for mortality prediction.

The CORe model, a transformer-based language model, was used. Evaluation metrics included accuracy, precision, recall, F1 score, and confusion matrix.

The model was quantized to reduce computational and memory requirements. Comparisons were made between the original non-quantized model and 8-bit and 4-bit quantized versions.

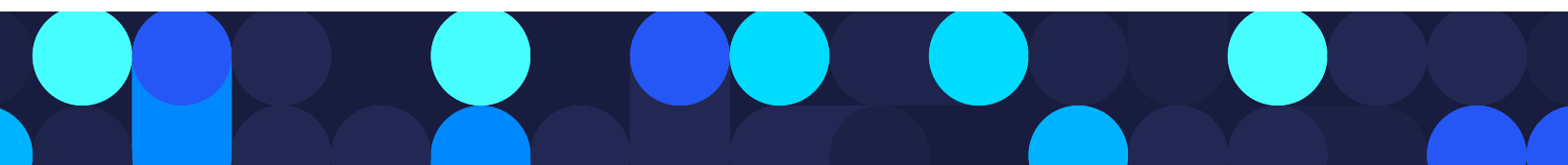
Evaluated llm_int8_threshold for 8-bit quantization and nf4/fp4 methods for 4-bit quantization, assessing efficiency and accuracy trade-offs.

Results

The baseline (non-quantized model), achieved an accuracy of 74.36%, precision of 0.87, recall of 0.74, F1 score of 0.79 and processing time of 192.78 s. For 8-bit quantization using LLM.int8() parameter, recommended value of 6 resulted in the lowest accuracy. While a value of 4 resulted in best accuracy with a processing time of 296.79s. This highlights the importance of tuning this hyperparameter for optimal performance. Surprisingly, both nf4 and fp4 4-bit quantization strategies yielded identical performance, suggesting 4-bit quantization's independence from the strategy.

Discussion/Conclusion

Our study explored the impact of quantization on a transformer-based language model for clinical outcome prediction. The performance degradation from an 8-bit quantized model can be reduced by tuning the LLM.int8() parameter. Careful consideration of trade-offs between efficiency and predictive accuracy is essential when implementing quantization. While the present results don't endorse adopting quantization, evaluating larger and more sophisticated models, fine-tuning, and architecture modifications may offer the potential to improve quantization performance in AI-driven clinical predictions.



[OR60] Evaluating the Efficacy of Transformer Networks for Audio Signal Classification in Dysphagia Detection

Hamza Mahdi, Sunnybrook Research Institute

Eptehal Nashnoush, Sunnybrook Research Institute

Rami Saab, Sunnybrook Research Institute

Arjun Balachandar, Department of Medicine, University of Toronto, Toronto, Canada

Houman Khosravani , Division of Neurology, Department of Medicine, Sunnybrook Health Sciences Centre, University of Toronto

Introduction

Dysphagia, a common post-stroke complication, can lead to significant morbidity and mortality. Machine learning has shown promise in dysphagia detection using audio analysis, but the potential of Transformer Networks remains largely unexplored. This study assessed Vision Transformer (ViT) and Swin Transformer models in dysphagia screening.

Methods

We recruited 68 post-stroke patients, segmented audio samples into 1,579 clips, and converted them into 6,655 Mel-spectrograms. These images underwent processing to create a 3-channel color map or a composite image from stacked single-channel spectrograms. The pass/fail status of the TOR-BSST swallowing-screening test was used as ground truth.

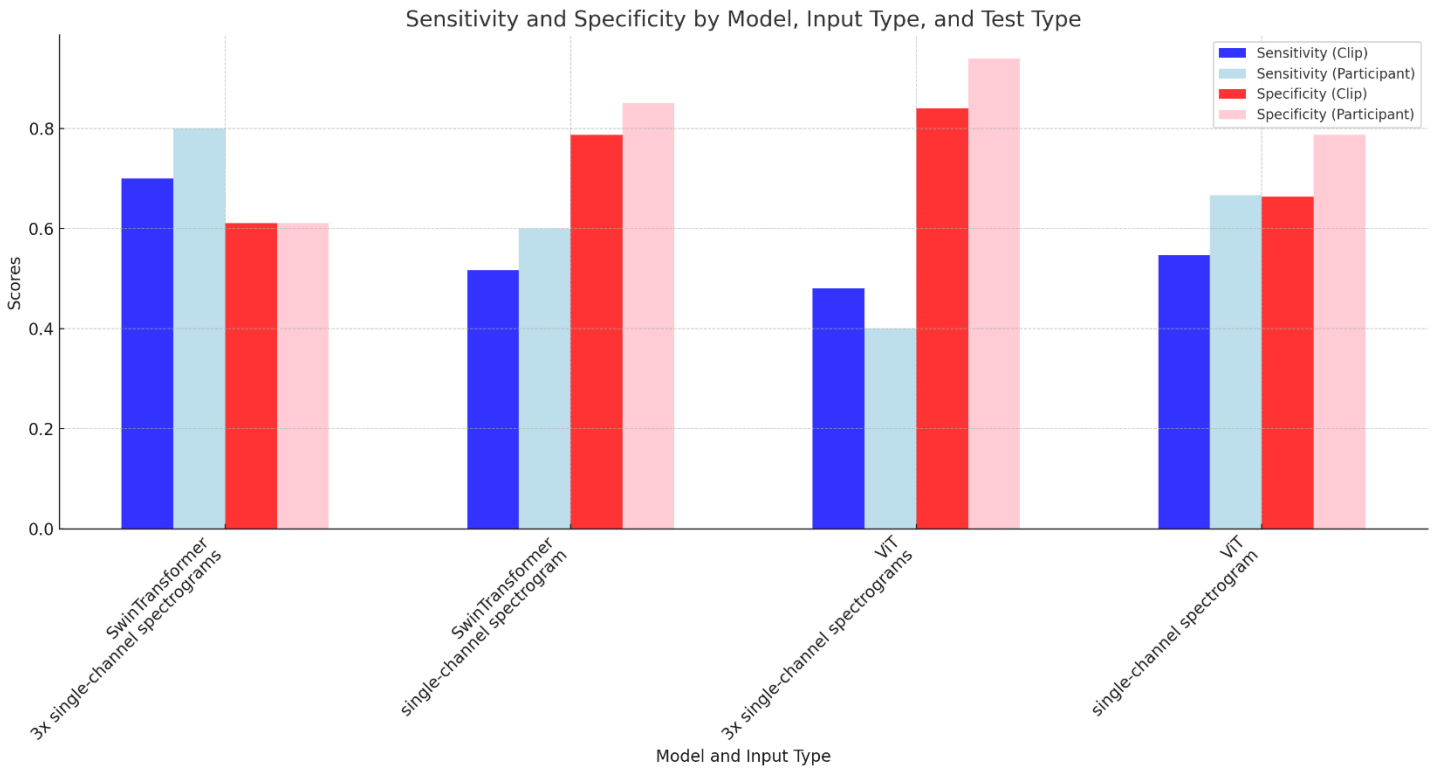
Results

In our study, both ViT and Swin Transformer models demonstrated notable results across different input types. The Swin Transformer model achieved an AUC ranging from 0.69 to 0.72 at the clip level and 0.78 to 0.83 at the participant level, sensitivity ranging from 0.45 to 0.80, and specificity ranging from 0.61 to 0.94. The ViT model achieved an AUC ranging from 0.59 to 0.73 at the clip level and 0.68 to 0.84 at the participant level, sensitivity ranging from 0.4 to 0.78, and specificity ranging from 0.61 to 0.94.

Discussion/Conclusion

Our findings highlight the potential of Transformer Networks, specifically ViT and Swin Transformer models, in audio signal classification for dysphagia screening. While the Swin Transformer model demonstrated higher sensitivity, the ViT model showed better specificity, indicating a trade-off between the models. Future work should focus on optimizing these models with diverse and larger datasets to validate their performance and explore their applicability in real-world healthcare settings.

Supporting information



[OR61] Patient and Stakeholder Engagement (PSE) in the Integration of Large Language Models (LLMs) in Healthcare Chatbots

Nikhil Jaiswal , Department of Family Medicine, Faculty of Medicine, McGill University
Yuanchao Ma , Research Institute of the McGill University Health Centre (RI-MUHC)
Kim Engler, Research Institute of the McGill University Health Centre (RI-MUHC)
David Lessard, Research Institute of the McGill University Health Centre (RI-MUHC)
Bertrand Lebouché, Department of Family Medicine, Faculty of Medicine, McGill University
Esli Osmanlliu , McGill University Health Centre

Introduction

In 2020, our multidisciplinary team successfully co-constructed an intelligent chatbot (MARVIN) with and for people with HIV. We are now adapting MARVIN to other specialties (e.g., oncology, pediatrics), while modifying its algorithmic infrastructure to further integrate LLMs. The power of PSE is central to this project, but is often an afterthought in the field.

Methods

This narrative piece summarizes key challenges of LLM integration into healthcare chatbots, identified through critical literature appraisal. Challenges were organized following the Software Development Life Cycle. Using our experience with MARVIN, we identify corresponding strategies to optimize PSE value in chatbot co-construction.

Results

PSE benefits successful LLM integration throughout the lifecycle of healthcare chatbots (Figure 1).

In the conception stage, early PSE through brainstorming workshops and focus groups identifies: a) needs that are most relevant to the population, b) where LLMs can be most valuable, and c) acceptable trade-offs between chatbot performance and use of powerful, but closed-source, products.

During development, the static and broad training data used by general-purpose LLMs limits their usefulness in specialized healthcare domains. A dedicated committee of patients and providers can improve this process by validating and fine-tuning knowledge databases. Efficiency gains are possible through methods including reinforced learning from human feedback.

At the testing and evaluation phase, LLM-based chatbots lack standardized assessment frameworks and guidelines for transparent reporting. PSE is also essential to develop patient-reported outcome and experience measures for use in upcoming healthcare chatbot trials.

During implementation, LLM-based healthcare chatbots face challenges related to low adoption, trustworthiness, and access. Through a multidisciplinary and diverse governance board, development teams will identify “blind spots” early and make design inclusive. Once implemented, the benefits of healthcare chatbots can be equitably distributed in society.

Discussion/Conclusion

PSE stands as the linchpin for the successful integration of LLMs in healthcare chatbot development, ensuring responsible AI-powered healthcare innovation.

[OR62] Prediction of Trauma Bay Disposition Using Explainable Machine Learning

Seong Park, Schulich School of Medicine, Western University
Anton Nikouline, London Health Sciences Centre, Western University
Ian Ball, London Health Sciences Centre, Western University
Pingzhao Hu, Western University
Kelly Vogt, London Health Sciences Centre, Western University

Introduction

Trauma is a global health concern, accounting for 9.2% of deaths and 10.9% of disability-adjusted life-years. Although established algorithms for trauma care, such as the Advanced Trauma Life Support, aim to alleviate this burden by supporting timely and more appropriate clinical decisions, pressured environments can elicit errors and deviations. Traumatically injured patients often require high-cost resources, including operating rooms, ventilators, intensive care unit stays and long admissions.

Machine learning offers a promising avenue for predicting necessary hospital resources. We sought to use machine learning to predict the disposition of traumatically injured patients from the emergency department using early clinical data from a retrospective dataset.

Methods

We conducted a retrospective cohort study of trauma patients from the National Trauma Data Bank (NTDB). Four years of data (January 2017 - December 2020) were analyzed to include patients with discharge dispositions of ‘Admitted,’ ‘Discharged,’ ‘Operating Room,’ ICU, ‘Deceased,’ or ‘Transferred to Another Hospital,’ with associated demographic and early clinical (pre-hospital, emergency department) information. Data processing was performed on the training dataset, including resampling, scaling, and imputation, prior to training six machine learning models (linear regression, random forest, adaptive boosting, gradient boosting, extreme gradient boosting and multilayer perceptron).

Results

The six models were tested on the validation dataset, and the area under the receiver operating characteristic curve (AUC), sensitivity, and specificity were derived to evaluate the models’ prediction performances.

Discussion/Conclusion

Our models were able to predict discharge dispositions with strong prediction characteristics. The utility of demographic and early clinical markers can be used for accurate prediction of patients in trauma care. With continued development and validation, these prediction models can help support hospital preparedness for these patients requiring high-cost resources.

[OR63] The use of large language models for therapeutic recommendations

Jean Marie Tshimula, McGill University

Dan Poenaru, McGill University & McGill University Health Center – Research Institute (MUHC-RI)

Introduction

With the growing demand for effective, evidence-based therapies, new approaches to therapeutic decision-making are worth exploring. Large language models (LLMs) have shown promise in supporting clinicians through natural language processing (NLP) techniques. In this systematic review, we investigate the existing literature on LLMs as an aid to therapeutic recommendations.

Methods

We systematically identified studies employing LLMs to predict treatment outcomes or develop personalized treatment plans in 15 scientific databases between 2019 and 2023. Studies were included if they used NLP techniques to analyze electronic health records (EHRs), other medical texts, or multimodal data, and were published in English and based on original datasets that were collected prospectively and retrospectively. Data extraction focused on the methodology, results, and conclusions drawn from each study.

Results

Our search yielded 37 potential abstracts, out of which we identified 29 eligible studies. Of these, 6 studies demonstrated improved diagnostic accuracy when LLMs were applied to EHRs. Another 8 studies evaluated the effectiveness of LLMs in generating personalized treatment plans, showing statistically significant improvements in terms of efficacy, disease severity/symptoms and higher patient satisfaction rates. Furthermore, 15 studies explored the integration of LLMs into clinical workflows, revealing potential benefits such as increased efficiency and reduced cognitive burden on clinicians.

Discussion/Conclusion

These findings underscore the potential of LLMs as a valuable tool for therapeutic recommendations. By analyzing vast amounts of medical text data, LLMs can uncover complex relationships and patterns that might otherwise go undetected. While further research is needed to validate these initial results, the present review suggests that LLMs hold great promise for fostering personalized medicine. As the field continues to evolve, we can expect to see even more innovative applications of LLMs in clinical decision-making, ultimately leading to improved patient outcomes and more informed treatment decisions.

[OR64] TxT -Toronto-Technion Treatment Curator

Melanie Courtot , Ontario Institute for Cancer Research
Michael Fralick , Department of Medicine, Sinai Health System
Bryant Lim , Department of Medicine, University of Toronto
Justin Richardsson , Ontario Institute for Cancer Research
Oren Caspi , Technion-Israel Institute of Technology
Yonatan Belinkov , Technion Taub Faculty of Computer Science

Introduction

Existing knowledge dissemination platforms for randomized controlled trials fall severely short of their intended goals: at present, they fail to enable informed patient decision-making and limit clinician access to state-of-the-art treatment options. ClinicalTrials.gov, housing over 300,000 trials, lacks user-friendly navigation and curation, and published clinical trial papers are scattered across multiple sources.

Methods

We have built TxT, a tool that scrapes clinical trial metadata (e.g., authors, year of publication, journal title) and abstracts from MEDLINE and ClinicalTrials.gov. From each abstract, key trial outcomes (i.e., sample size, comparison groups, blinding, primary outcome, and conclusions) were extracted using the text-davinci-003 large language model interface from Open AI. This information can be loaded into Overture to generate an interactive display and discovery portal as shown on Figure 1.

Results

We have started deploying an AI system that curates, validates, filters, and summarizes clinical trial information from various sources. The system extracts key details from official databases (e.g., MEDLINE, ClinicalTrials.gov) as well as non-conventional sources (e.g., preprint servers, social media). We have preliminary results for 50 papers which demonstrate the feasibility of the approach. We will expand on this to (1) include more papers (2) expand to enable input from other existing resources (3) add a lay summary generator module to Overture.

Discussion/Conclusion

TxT, the Toronto-Technion treatment curator generates personalized summaries for doctors and lay summaries for patients, providing a user-friendly resource for comprehensive clinical trial information. We will expand on current text-based feature to support infographic generation for easy-to-use visual representation, further supporting broad dissemination and leverage of clinical information.

Supporting information

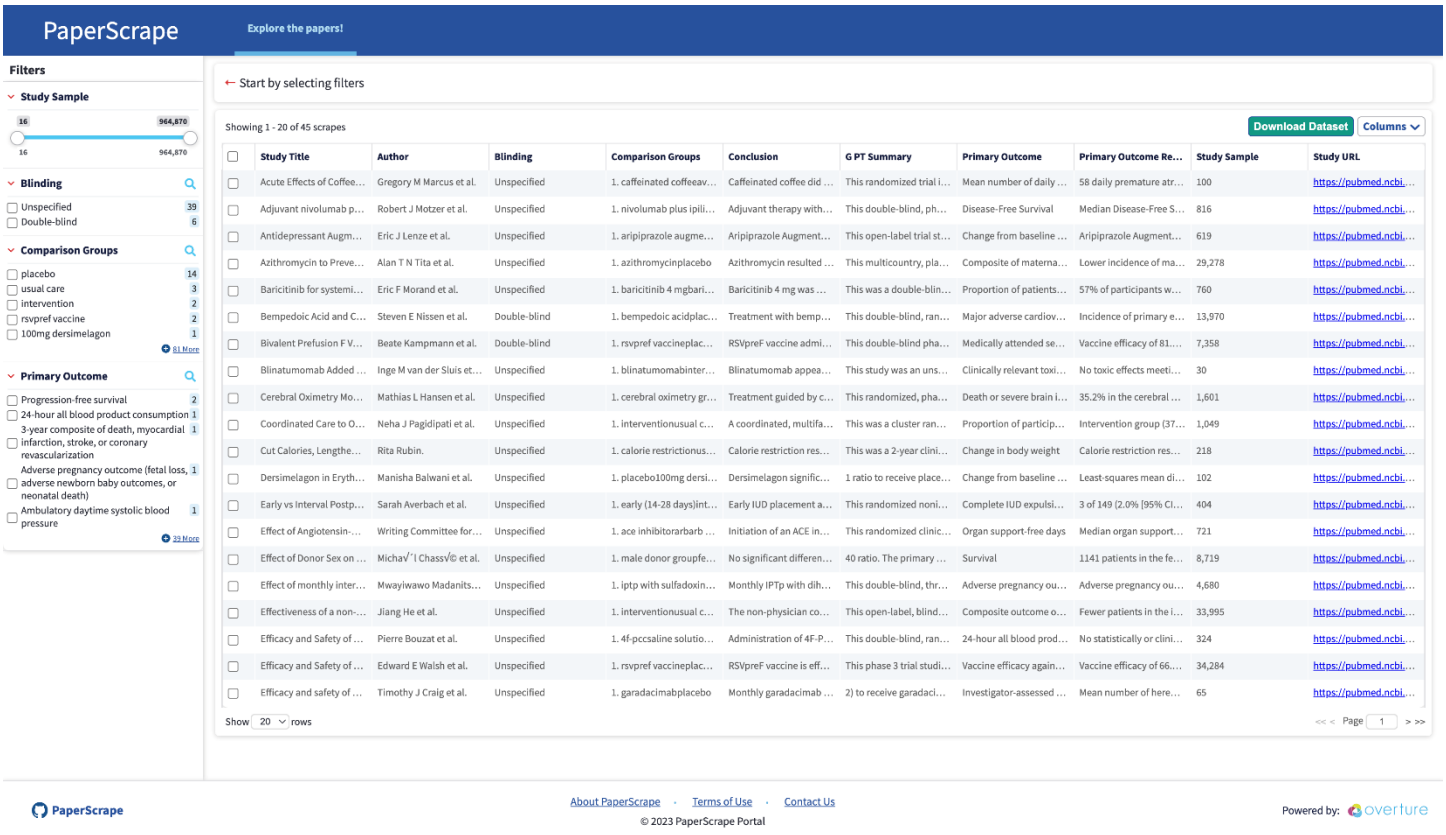


Figure 1. A screenshot of the TxT portal whowing extracted standardized information. Filters in the left-hand side enable clinicians to refine their query based on metadata attributes and/or their range. Data can be downloaded, and source study is linked

[ED01] Exploring socially accountable Artificial Intelligence for healthcare in Northern Ontario

Holly Fleming , NOSM U
Sophie Myles, Algoma OHT
Andrew Austin, NOSM U
Erin Cameron, NOSM U

Introduction

Implementing Artificial Intelligence (AI) in healthcare in Northern Ontario necessitates accommodating the region's diverse geography and demographics. AI-NORTH aims to explore Northern Ontarians' experiences to comprehend requirements for socially accountable AI system design and implementation. Key objectives are to: conceptualize socially accountable AI based on community needs, values, and priorities; identify current strengths and gaps in regional AI research; and build capacity for participatory, community-engaged AI research to elucidate the implications of AI for the people of Northern Ontario.

Methods

We created communities of practice to explore healthcare experiences across Northern Ontario, to develop a deep understanding of AI for healthcare in the region, and to understand challenges and opportunities for health service providers and users to incorporate AI in healthcare. We conducted a series of 8 interactive webinars using appreciative inquiry and deliberative dialogue with healthcare providers and community members (n=80) across Northern Ontario. Topics have included: equity, education, public health, policy. We used a narrative approach to understand the implications of AI for Northern Ontario.

Results

Key findings show Northern Ontarians should choose participation in AI systems affecting their communities. Connectivity issues in remote areas need addressing. Educating both professionals and the public on AI fundamentals—like how the technologies work, limitations, and impacts on equity, access, and privacy—is vital. This education will empower Northern Ontarians to self-advocate regarding AI implementation.

Discussion/Conclusion

Socially accountable AI requires responsible development and implementation that centers the needs, values, and welfare of Northern Ontarians. It means designing and deploying AI systems transparently, fairly, and inclusively, ensuring accountability and mitigating potential biases. Building public trust necessitates addressing concerns around privacy, security, and impacts on rural, remote, Indigenous, and other diverse communities. This includes being upfront about limitations, providing means for feedback and accountability, and ensuring local participation in shaping these technologies.

[ED02] Exploring the Potential Utility of AI Large Language Models for Medical Ethicists: An Expert Panel Evaluation of ChatGPT

Michael Balas, Temerty Faculty of Medicine, University of Toronto
Jordan Wadden, Unity Health Toronto
Philip Hébert, Temerty Faculty of Medicine, University of Toronto
Eric Mathison, Department of Philosophy, University of Toronto
Marika Warren, Department of Bioethics, Dalhousie University
Victoria Seavilleklein, Clinical Ethics Service, Alberta Health Services
Daniel Wyzynski, Office of Health Ethics, London Health Sciences Centre
Alison Callahan, Ethics Department, Ontario Shores Centre for Mental Health Services

Introduction

This study sought to evaluate the performance of an AI large language model, ChatGPT (version 4.0), in responding to a series of complex medical ethical vignettes, with the goal of understanding its potential utility and limitations in aiding medical ethicists with their ethical decision-making processes.

Methods

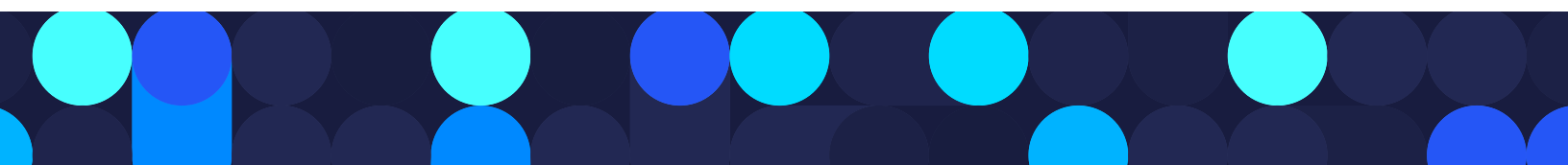
We designed a mixed-methods cross-sectional survey study. Eight ethical vignettes were created, each followed by several probing questions, to which ChatGPT-4 provided the responses. Six experienced ethicists independently evaluated each AI response using a standardized rubric consisting of six key metrics, in addition to open-text comments. Quantitative and qualitative analyses were conducted on the evaluation data. Additionally, readability metrics were computed for every AI-generated response.

Results

Of the six metrics evaluating the effectiveness of the AI responses, the overall mean score was 4.1 out of 5. ChatGPT-4 was rated highest in providing technical (4.7/5) and non-technical clarity (4.4/5), whereas the lowest-rated metrics were depth (3.8/5) and acceptability (3.8/5). There was poor-to-moderate inter-rater reliability, characterized by an intraclass-coefficient of 0.54 (95% CI: 0.30 to 0.71). Based on panelist feedback, the AI was able to identify and articulate key ethical issues, but struggled to appreciate the nuanced aspects of ethical dilemmas and misapplied certain moral principals.

Discussion/Conclusion

Large language models may hold significant potential for assisting medical ethicists in the complex task of navigating ethical dilemmas within the healthcare sector. However, it is essential to underscore that such tools should be used as supplements to, and not replacements for, human expertise. Currently, systems like ChatGPT-4 are limited in their ability to appreciate the depth and nuanced acceptability of real-world ethical dilemmas, particularly those that require a thorough understanding of relational complexities and context-specific values. As this field continues developing, the necessity for ongoing, critical evaluation of its capabilities, limitations, and impacts within the sphere of medical ethics remains paramount.



[ED03] The use of artificial intelligence in evaluating medical trainees: A literature review

Shaoyuan Wang, Department of Psychiatry, Temerty Faculty of Medicine, University of Toronto
Neil Patel , Leslie Dan Faculty of Pharmacy, University of Toronto
Tahani Dakkak, Leslie Dan Faculty of Pharmacy, University of Toronto
Sanjeev Sockalingam , Centre for Addiction and Mental Health
Pamela C.. Molina , University of Toronto
Samir Kanji, Leslie Dan Faculty of Pharmacy, University of Toronto
Certina Ho , Department of Psychiatry, Temerty Faculty of Medicine, University of Toronto

Introduction

Artificial intelligence (AI) has long been integrated into medicine for the purposes of clinical decision-making and research. More recently, it has been used in medical education to provide individualized feedback, assess education programs, and identify trainees' learning needs. We aimed to conduct a literature review that explores how medical schools and residency programs have implemented AI in evaluating trainees.

Methods

Search terms that captured the three key concepts of learners (medical students, residents, fellows, medical education), AI (machine learning, deep learning, neural networks, natural language processing), and evaluation (clinical competence, progression, success, remediation, EPAs) were employed. We searched Medline and Embase for original research papers published in English (inclusion criteria) from 2012 to 2022. Exclusion criteria were reviews, opinion pieces, conference abstracts/proceedings, expert panels, and focus groups. This initially yielded 775 results but after duplicate removal and abstract screening, 30 papers were left. In the end,18 papers remained for data extraction after full-text screening.

Results

Six papers(1-6) indicated that neural networks, machine learning, and natural language processing were able to differentiate between levels of training by analyzing surgical skills and/or clinical reasoning notes. Four papers(7-10) reported that machine learning and natural language processing could predict competence outcomes by real-life assessors. Three(11-13) papers found that robot trainers, natural language processing, and machine learning provided useful feedback that improved learner performance. Finally, three papers(14-16) used natural language processing to delineate features of narrative feedback in EPAs that were associated with entrustment.

Discussion/Conclusion

Although the new application of AI in educational assessment seems relatively new based on this literature review, existing research emphasizes its potential utility. We hope our findings can inform how AI can play a role in assessing psychiatry trainees.

Supporting information

https://docs.google.com/document/d/1YIx4RecXrTR3hPlNdq8uYTMGxvC5g7v50KXCOcCgp6Q/edit?usp=sharing

[ED04] Rethinking Vascular Surgery Training: A Comparative Study of ChatGPT 3.5 and ChatGPT 4 in Navigating VESAP 5 Modules

Tiam Feridooni, University of Toronto
Arshia Javidan, University of Toronto
Lauren Gordon, University of Toronto
Sean Crawford, University of Toronto

Introduction

Artificial intelligence (AI) is increasingly becoming as a viable resource in medical education and training, with the potential to enhance learning outcomes. This study aims to evaluate the proficiency of readily available online Generative Pre-trained Transformer models, ChatGPT 3.5 and ChatGPT 4, in navigating the Vascular Education and Self-Assessment Program (VESAP) 5 Modules, a key resource for vascular surgery training.

Methods

We conducted a comparative analysis of the performance of ChatGPT 3.5 and 4 in navigating the VESAP 5 in text-only multiple-choice questionnaire. The incorrect output of each was evaluated and labelled for the reason for the incorrect answer which included logical, information and statistical error.

Results

ChatGPT 4 performed better than ChatGPT 3.5 in aortoiliac disease (51% vs 43%, n=35), cerebrovascular disease (67% vs 50%, n =48), renal and mesenteric disease (49% vs 78%, n=41) vascular medicine (55% vs 80%, n=49) and venous disease (72% vs 79%, n=39). Overall, ChatGPT4 performed significantly better in all sections (71.0% ± 12.2% vs. 53.7 ± 11%, $p = 0.026$). The absolute difference in overall accuracy was 17.9% ($p = 0.0001$). With regards to the incorrect responses, we noted that compared to ChatGPT 4, ChatGPT3.5 had a significantly higher percentage of logical errors (31.6 ± 13.1% vs. 18.2 ± 5.3%, $p = 0.015$), while ChatGPT 4 had a significantly higher percentage of information errors (80.7 ± 7.1% vs. 68.4 ± 5.3%, $p = 0.015$).

Discussion/Conclusion

ChatGPT 4 performed significantly better than ChatGPT3.5 in the five selected section of VESAP 5. Our study demonstrated that there has been a significant improvement in the logical capabilities of ChatGPT, as ChatGPT 4 had a significantly lower number of logical errors. Our study highlights the potential of AI-powered learning platforms, such as ChatGPT, to revolutionize medical education by offering personalized, efficient, and scalable learning solutions.

[P01] 3D Pose Estimation Using RGB-D Data for Rehabilitation

Gloria-Edith Boudreault-Morales, University of Toronto

José Zariffa, KITE - Toronto Rehabilitation Institute - University Health Network; Institute of Biomedical Engineering, University of Toronto; Rehabilitation Sciences Institute, University of Toronto; Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto

Christopher Witiw, Division of Neurosurgery, St Michael's Hospital

Introduction

The rise of new neuromodulation therapies has created a need to track motor performance to evaluate and personalize new approaches. Pose Estimation (PE) neural networks can be used for this purpose. Most current published approaches use color (Red-Green-Blue, or RGB) data as an input, and are trained and validated using data from healthy individuals.

We wish to develop a self-contained system that can collect data and perform 3D PE. The aims are: 1) Determine how depth (D) data affects a lightweight monocular RGB PE model (accuracy and speed). 2) Validate the model using data from individuals with physical impairments to ensure that accuracy is not affected by the presence of disabilities.

Methods

1) State-of-the-art lightweight monocular RGB PE models (MobileHumanPose and Dite-HRNet) will be altered to include depth as an input. They will be retrained using RGB-D data from a public dataset. Model accuracies and computational efficiency will be compared to their RGB-only versions.

2) Participants, uninjured or with physical impairments due to a stroke, will be recorded using a RGB-D camera and a marker-based motion capture system (for ground truth). The RGB-D data will be fed into models from Aim 1 to generate joint location predictions. Accuracies between groups (uninjured and injured) will be compared using t-tests to determine whether there is a significant difference between them.

Results

We hypothesize that adding depth information as an input will slow down the model but increase its accuracy, and that the presence of physical impairments will not have a significant effect on the model's accuracy.

Discussion/Conclusion

Implementing PE models into rehabilitation processes could lead to more precise ongoing motor assessment and support personalized interventions. It could also provide a scalable approach to capturing data about recovery trajectories and play a key role in improving the evidence base for interventions.

[P02] A Novel AI Clinician: Training an Intelligent System to Predict COVID-19 Pneumonia Hospital Outcomes

George Chen, Centre for Heart Lung Innovation (HLI)

James Russell, Centre for Heart Lung Innovation (HLI)

Introduction

As COVID-19 erupted, healthcare systems became overwhelmed by sheer uncertainty. However, many of the answers we sought lay within data amassed during the pandemic, just out of reach.

Our goal is to support future pandemic responses by using artificial intelligence (AI) to harness this data. Our model analyzes day 0 factors of hospitalized COVID-19 pneumonia patients to predict the eventual need for ventilation and vasopressors.

Methods

1210 patients with COVID-19 pneumonia from the CAPtivate network were used to train a random forest classifier.

Post-day-0 variables and COVID-19-negative patients were removed. The data was normalized. Missing values were imputed using MICE, column-wise mean, and KNN imputation. The dataset was balanced using SMOTE. Data was split 70% for training and 30% for testing. The model was tuned with grid-search hyperparameter optimization.

An additional 1210 patients from the same population were fed into the model to simulate clinical use.

Results

Ventilator use was predicted with an accuracy of 0.95, sensitivity of 0.91, and specificity of 0.98. Vasopressor use, acute respiratory distress syndrome (ARDS), and intensive care unit (ICU) admission were the most important predictors with relative importance of 0.222, 0.109, and 0.088, respectively.

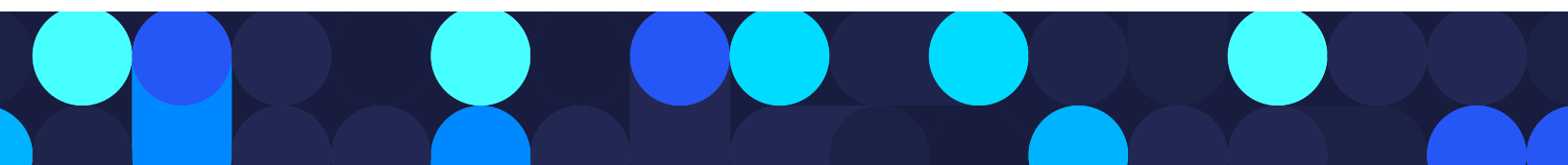
Vasopressor use was predicted with an accuracy of 0.88, sensitivity of 0.75, and specificity of 0.94. ARDS, ICU admission, and shock were the most important predictors with relative importance of 0.123, 0.121, and 0.063, respectively.

Discussion/Conclusion

Our system has shown promising performances in predicting outcomes, especially the need for ventilation.

Hopefully, our program will be used in supporting decisions regarding the need for important interventions. By interpreting patient data and providing care suggestions, we want to help clinicians save time and feel more informed. This does not replace clinicians but rather acts as another component in building a treatment plan.

Next, we will test the generalizability of this model to non-COVID pneumonia patients.



[P03] Accelerating Personalized Breast Cancer Treatment: Validating an AI-Driven Quantitative Ultrasound Model for Predicting Neoadjuvant Chemotherapy Response

Matthew Shammas-Toma, Sunnybrook Health Sciences Centre
David Alberico, Sunnybrook Health Sciences Centre
Lakshamanan Sannachi, Sunnybrook Health Sciences Centre
Schontal Halstead, Sunnybrook Health Sciences Centre
Gregory Czarnota, Sunnybrook Health Sciences Centre
Katarzyna Jerzak, Sunnybrook Health Sciences Centre

Introduction

Background: Locally advanced breast cancer (LABC) is advanced breast cancer without distant metastases. Quantitative ultrasound (QUS) has emerged as a novel technique for assessing tumour microstructure changes. Previously, we developed a machine learning model combining QUS, texture analysis and molecular subtypes which predicts LABC tumor response before neo-adjuvant chemotherapy (NAC) treatment. This study aims to quantitatively evaluate the performance of our LABC machine learning model using an independent patient cohort.

Methods

We enrolled 92 LABC patients with a median age of 50 years. Their primary tumours were imaged using ultrasound (Ultrasonix Sonix RP system) before the commencement of NAC. For each patient, the primary tumour and peritumoral region were manually contoured at five distinct ultrasound frames. A set of QUS parameters were derived using spectral analysis methods. Next, texture and texture-derivative features were obtained via image analysis using the gray-level co-occurrence matrix method. Tumour molecular subtypes were defined based on hormone receptor status. Our response prediction model was previously trained on a separate cohort of 208 patients. It was developed using a support vector machine algorithm with an accuracy of 83% and an area under the curve (AUC) of 0.87. The QUS, texture, texture-derivate features and molecular subtypes were used as the input to the model.

Results

Responders and non-responders to NAC treatment accounted for 88 patients and 4 patients, respectively. Among 92 patients, our model correctly predicted 81 responders and 4 non-responders. There were 7 false positives and 0 false negatives. The model exhibited a sensitivity of 100%, specificity of 92.0% and an accuracy of 92.3%. In addition, the model demonstrated an AUC of 0.83.

Discussion/Conclusion

The ability of a QUS-texture based model in the prediction of treatment response was prospectively validated in the current study. The study demonstrates the feasibility of QUS-guided NAC treatment for patients with breast cancer.

[P04] An AI-assisted Chatbot for Patient Education and Care in Radiotherapy

James C. L.. Chow , UHN
Leslie Sanders , York University
Kay Li , University of Toronto

Introduction

Our team developed the RT Bot, an AI-chatbot, to provide educational information about radiotherapy. It caters to patients, the general public, and radiation staff. The Bot uses machine learning to personalize responses and offer human-like guidance. It detects user backgrounds and needs, utilizing different datasets for interaction.

Methods

The Bot was created using IBM Watson Assistant functionalities deployed on the IBM cloud. To provide comprehensive information to users, we prepared specific datasets tailored to different user groups. These datasets encompass descriptions of radiotherapy processes, promote cancer screening, and provide basic cancer prevention measures. Through machine learning, the Bot was trained using user questions to enhance its accuracy. Personalization of the Bot's character was achieved by leveraging AI features such as natural language understanding. This enabled the Bot to precisely identify user intentions and provide information that aligns with their concerns.

Results

The Bot can be easily accessed through a front-end window on various Internet-of-Things devices. When users interact with the Bot, it initially engages them with an introduction, establishing a connection. Users can input their inquiries, with the Bot responding to questions and providing relevant information. When understanding is unclear, guidance is offered. For example, the Bot determines the user's background to assign them to a specific user group, and subsequently presents a list of options for the user to select from, followed by a more detailed breakdown of the selected option. This user-centric approach allows the bot to accurately and efficiently address the user's needs.

Discussion/Conclusion

We developed an AI-powered chatbot to educate patients, the general public, and radiation staff about radiotherapy. It finds applications in cancer centers, schools, community centers, and charities promoting cancer prevention and screening. The Bot reduces patient concerns, provides accurate information, and educates the public on preventive measures and screening programs.

[P05] Applying mechanistic in-silico EEG biomarkers for improved diagnosis in depression

Frank Mazza, Krembil Centre for Neuroinformatics, CAMH

Muhammad Hassan , McMaster University

Alexandre Guet-McCreight , Krembil Centre for Neuroinformatics, CAMH

Etay Hay , Krembil Centre for Neuroinformatics, CAMH

Introduction

Recent post-mortem studies indicate that one mechanism of depression may be reduced inhibition from somatostatin-expressing interneurons (SST+INs). However, the link between SST+IN inhibition and quantitative clinical measures of brain activity (electroencephalogram, EEG) remains unknown. As it is currently impossible to test cell-specific mechanisms in living patients, we leveraged simulations of detailed computational models to identify mechanistic EEG biomarkers and used ANNs trained on these biomarkers to predict patient inhibition level.

Methods

We used our previous biophysically detailed models of human cortex. We generated depression models of different severities by reducing SST+IN inhibition by 10% - 60%, as estimated from post-mortem gene expression changes in depression. We identified EEG biomarkers of reduced inhibition using standard methods EEG power spectral analysis.

We trained ANNs on in-silico EEG to predict microcircuit group and inhibition level. We then used the highest performing ANN to predict patient group and inhibition level. Resting-state patient EEG, depression severity, and symptoms scores were used from the EMBARC dataset (n=300 depression, n=75 healthy controls).

Results

ANNs predicted microcircuit group (96% accuracy, CI 92 – 100) and inhibition level with high performance (0.91 R2, CI 0.87 – 0.94; 5.1 RMSE, CI 5.1 – 7.1). ANNs predicted patient group with highest accuracy at anterior-frontal electrodes (AF7, 65% CI: 60 – 71), with moderate predictive power of inhibition level (severity) (0.55 R2, CI 0.45 – 0.6; 12.5 RMSE, CI 9.2– 18.1).

Discussion/Conclusion

Anterior-frontal electrodes showed highest discriminatory power, following localized SST expression reduction in post-mortem studies. Our findings suggest detailed microcircuit models may be used to predict patient disease for improved mechanistic diagnosis.

[P06] Beyond Hand-Crafted Features For Pretherapeutic Molecular Status Identification Of Pediatric Low-Grade Neuroepithelial Tumors

Kareem Kudus, University of Toronto
Khashayar Namdar, University of Toronto
Matthias Wagner, The Hospital for Sick Children
Birgit Ertl-Wagner, The Hospital for Sick Children
Farzad Khalvati, University of Toronto

Introduction

The use of targeted agents in the treatment of pediatric low-grade neuroepithelial tumors (PLGNT) currently relies on determining molecular status through biopsy. However, it was recently shown that the two most common types of gene alterations can be identified non-invasively using MRI-based radiomic features. We used Convolutional Neural Networks (CNN) along with radiomics to build a model that, in addition to the two most common types, can differentiate all remaining gene alterations observed in PLGNT as a third type.

Methods

This study used the tumor region, manually segmented from T2-FLAIR MR images, of 339 patients treated for PLGNT between 1999 and 2018. Both a CNN and Random Forest (RF) radiomics model, along with another RF model trained on a combination of CNN and radiomic features, were trained to predict the genetic status of PLGNT. Additionally, we investigated whether CNNs could predict radiomic feature values from MR images.

Results

The combined model (mean AUC: 0.824) outperformed the individual radiomics (0.802) and the CNN (0.764) models. The difference in the performance of the models was statistically significant, (p-values under 0.05). The CNN was able to learn predictive radiomic features such as surface-to-volume ratio well (average correlation: 0.864) but was unable to learn others such as run-length matrix variance (-0.017).

Discussion/Conclusion

Theoretically, handcrafted features represent a subset of the features CNNs can capture. Thus, it is commonly believed that radiomic features are redundant to those discovered by CNNs. We presented evidence contrary to this belief, along with the first model that can accurately distinguish between all pLGG gene alterations non-invasively. We showed that a model relying on both CNN and radiomic features performs better than either approach separately. Additionally, we showed that a CNN has a limited ability to represent certain handcrafted features in this dataset, which explains why they work well in combination with radiomics.

[P07] Deep Learning Architectures for 3D Reconstruction of Oral Cancer Models from Intraoperative Spatial-Frequency Fluorescence Images

Natalie J. Won , Princess Margaret Cancer Centre
Anjolaoluwa Adewale , Princess Margaret Cancer Centre
Jerry Wan , Princess Margaret Cancer Centre
Mandolin Bartling , Department of Otolaryngology – Head & Neck Surgery, University of Toronto
Alon Pener-Tessler , Department of Otolaryngology – Head & Neck Surgery, University of Toronto
Esmat Najjar , Department of Otolaryngology – Head & Neck Surgery, University of Toronto
Brian C. Wilson , Princess Margaret Cancer Centre
Jonathan C. Irish , Department of Otolaryngology – Head & Neck Surgery, University of Toronto
Michael J. Daly , Princess Margaret Cancer Centre

Introduction

Fluorescence-guided surgery delineates tumors at the tissue surface during cancer interventions yet fails to illustrate overall 3D geometry. To address this limitation, our group is developing a deep learning (DL)-enabled fluorescence imaging system that predicts fluorescence concentration and the deep margin of tumors from structured illumination images.

Methods

To capture 3D information and tissue optical properties (OP), we use reflectance (R) and fluorescence (F) spatial frequency domain (Fx = 0, 0.05, 0.1, 0.15) imaging to obtain R/F/OP images for DL. Here, we compare three 3D convolutional neural network (CNN) architectures (Figure 1a): i) F + OP as inputs for a Siamese CNN (parameters = 900,802); ii) F + R as inputs for a symmetric Siamese CNN (parameters = 934,018); and iii) F/R as a single input for a CNN (parameters = 707,330). Models were trained using the Adam optimizer in AWS with synthetic images of tumor models (N = 10,000) generated from a numerical light propagation simulator. MRI contours of tongue tumors are used to assess model performance both in silico and experimentally with optical phantoms.

Results

There is no significant difference (F=1.16 < F crit=2.46, P=0.33) across models when testing with in silico data. The F/R model has the best performance when looking at depth predictions from experimental phantom data (Figure 1b).

Discussion/Conclusion

These preliminary results indicate that the F/R model shows the most promise when dealing with images collected experimentally. Future studies that assess model performance with in vivo tumors are required.

Supporting information

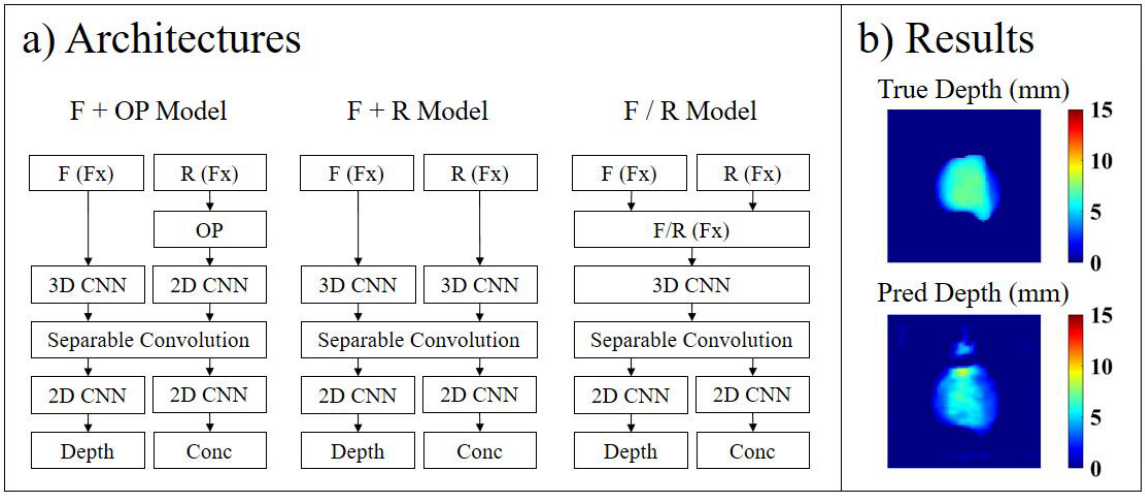


Figure 1. DL models. (a) Architectures (b) True and predicted depth (F/R model) from experimental images of optical phantoms.

[P08] Deep Learning-Enabled 3D Fluorescence Imaging for Surgery: A Simulation Study in the Second Near-Infrared Window

Jerry Wan , Princess Margaret Cancer Centre
Anjolaoluwa Adewale , Princess Margaret Cancer Centre
Natalie J. Won , Princess Margaret Cancer Centre
Mandolin Bartling , Department of Otolaryngology – Head & Neck Surgery, University of Toronto
Alon Pener-Tessler , Department of Otolaryngology – Head & Neck Surgery, University of Toronto
Esmat Najjar , Department of Otolaryngology – Head & Neck Surgery, University of Toronto
Brian C. Wilson , Princess Margaret Cancer Centre
Jonathan C. Irish , Department of Otolaryngology – Head & Neck Surgery, University of Toronto
Michael J. Daly , Princess Margaret Cancer Centre
Shaf Keshavjee, Latner Thoracic Research Laboratories, Toronto General Hospital Research Institute, University Health Network

Introduction

Fluorescence imaging is a promising optical technology for cancer surgery. Existing systems, however, have limited abilities to quantify tumor depth, posing challenges for surgeons. Thus, we are developing a deep learning-enabled 3D fluorescence imaging system. To date, the prototype system measures light emitted in the first near-infrared window (NIR-I), from 700-900 nm, to determine tumor depth. Newer camera technology operating in the second near-infrared window (NIR-II), above 1000 nm, may be less sensitive to tissue absorption and scattering and permit deeper light penetration. Hence, this simulation project assesses performance of our existing neural network with NIR-II window optical properties.

Methods

The prototype structured illumination system (Fig. a) collects reflectance and fluorescence images using a NIR-I camera. A Siamese, convolutional neural network converts fluorescence images and optical properties to maps of tumor depth and fluorescence concentration using the Adam optimizer (Fig. b). The convolutional model uses 3x3 filters for 2D convolution blocks and 3x3x6 filters for 3D convolution blocks. In silico training data (N&#3f10,000) is generated using numerical simulations of near-infrared light propagation based on tumor shapes comprised of composite spherical harmonics.

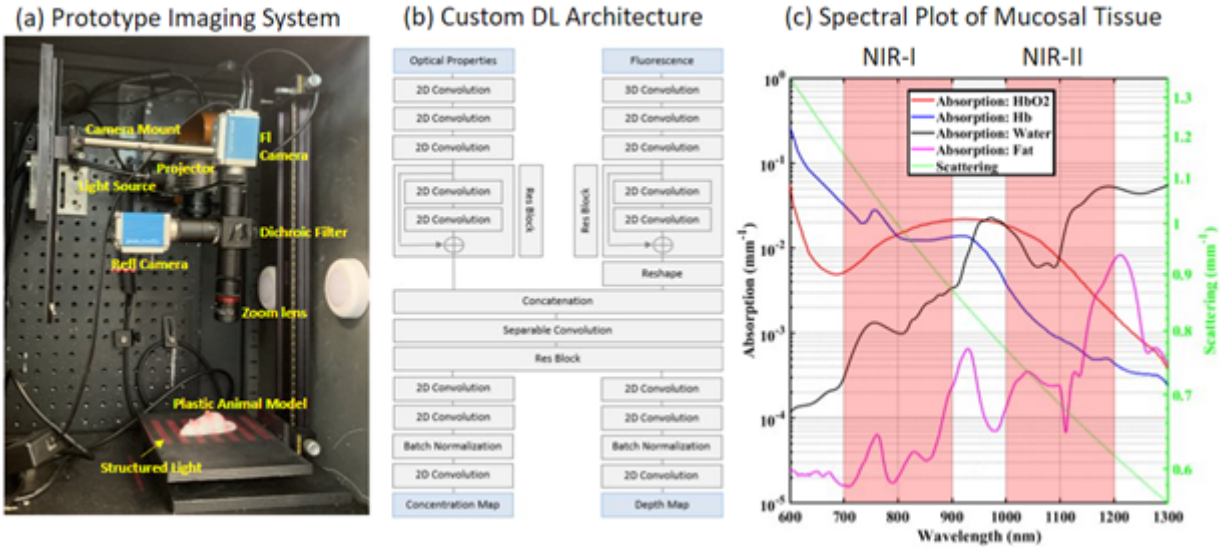
Results

A MATLAB simulation of optical absorption and scattering properties in the NIR-I and NIR-II (Fig. c) was based on known spectral properties of constituent components including oxyhemoglobin (HbO2), deoxyhemoglobin (Hb), water, and fat (lipid). Diffusion theory calculations of optical penetration depth are compared across clinically realistic ranges of optical properties. Ongoing deep learning simulations will compare performance for NIR-I and NIR-II estimates of tumor depth.

Discussion/Conclusion

This simulation study will assess the benefits of NIR-II camera technology for potential use in a prototype 3D fluorescence imaging system. Future studies will focus on optimizing our hardware system to integrate a NIR-II camera system.

Supporting information



[P09] Designing a healthbot for varenicline adherence using healthcare provider perspectives

Mackenzie V. Earle , CAMH
Sowsan Hafuth, CAMH
Kamna Mehra, CAMH
Jodi Wolff, CAMH
Matt Ratto, University of Toronto
Jonathan Rose, University of Toronto
Scott Veldhuizen, CAMH
Laurie Zawertailo, CAMH
Peter Selby, CAMH
Gary Bader, The Donnelly Centre, University of Toronto

Introduction

Healthbots - artificial intelligence programs designed to simulate human conversation via text - are being increasingly used to meet healthcare needs. Healthbots have the capability to act as supplemental healthcare agents to increase patient education, improve treatment compliance, and provide tailored medication support. In order to implement healthbots successfully into the healthcare system, healthcare providers must be willing to recommend and prescribe them to patients. Unfortunately, healthcare providers do not commonly prescribe healthbots or other artificial intelligence supports to patients, and there is little known reason why.

Methods

In this qualitative study, we explored the barriers and facilitators influencing healthcare providers' decision to prescribe healthbots to their patients. Nineteen different healthcare providers (e.g., physicians, nurses, etc.) from varying population centres (e.g., rural, urban) and locations across Ontario, Canada were interviewed on their perspectives of healthbots. The interviews were based on the Capability, Opportunity, Motivation, and Behaviour (COM-B) model of behaviour change and the Theoretical Domains Framework, a guide made up of the social, mental, and environmental influences known to elicit behavioural change.

Results

The most common barriers included healthcare providers' lack of motivation and capability to recommend healthbots because of technological barriers, fear for the future of artificial intelligence agents, and concerns about the impact on the patient-physician relationship. The most common facilitator was motivation, as healthcare providers believe healthbots can reduce workloads and increase patient follow-up.

Discussion/Conclusion

We are using this information to select appropriate behaviour change techniques for a healthbot to improve varenicline adherence so that healthcare providers will be comfortable recommending it to their patients. This project will help develop a healthbot that can increase patient compliance with varenicline and one that healthcare providers may willingly recommend.

[P10] Detection and recognition of hand impairment level in stroke survivors using egocentric video of activities of daily living

Anne Mei, KITE - Toronto Rehabilitation Institute - University Health Network; Institute of Biomedical Engineering, University of Toronto

Meng-Fen Tsai, KITE - Toronto Rehabilitation Institute - University Health Network; Institute of Biomedical Engineering, University of Toronto

José Zariffa, KITE - Toronto Rehabilitation Institute - University Health Network; Institute of Biomedical Engineering, University of Toronto; Rehabilitation Sciences Institute, University of Toronto; Edward S. Rogers Sr. Department of Electrical and Comp

Introduction

Stroke can impair upper-extremity motor control which in turn negatively impacts survivors' independence and quality of life. Current outcome measures evaluate motor function and the effectiveness of rehabilitation therapies with structured clinical assessments, which do not reflect true behaviour in activities of daily living (ADLs) at home. Wearable (egocentric) cameras provide a way to capture hand function information in natural environments. Automated analysis of egocentric video has previously been used to quantify at-home hand-use in stroke survivors, but not to describe impairment level. In this study, we investigated deep learning approaches to estimate the level of hand impairment at home after stroke. We aim to develop an approach that considers information from multiple ADLs, as clinical hand function assessments rely on observing hands in several functional tasks.

Methods

An egocentric video dataset of 16 ADLs performed by 4 stroke survivors in a home simulation lab was used to train a video recognition model (SlowFast) to predict hand impairment levels. Models were trained for binary impairment classification using 2 types of input: full-frame videos with corresponding bounding boxes to indicate hand regions; and videos cropped around each hand region. Predictions from sets of 3 different ADLs were concatenated and passed through a fully-connected neural network to estimate subject-level hand impairment. All models were evaluated with Leave-One-Subject-Out-Cross-Validation.

Results

SlowFast had an ADL-level F1-score of 0.638 ± 0.105 for classification on full-frame inputs and an F1-score of 0.777 ± 0.112 on cropped inputs. Trained on SlowFast predictions made from cropped inputs, the fully-connected network had a subject-level F1-score of 0.846 ± 0.113.

Discussion/Conclusion

SlowFast performs better on cropped inputs than full-frame inputs, likely because background and information less relevant to impairment classification were removed. Performance is improved when ADL-wise predictions are combined to the subject-level, indicating the benefit of considering multiple ADLs.

Supporting information

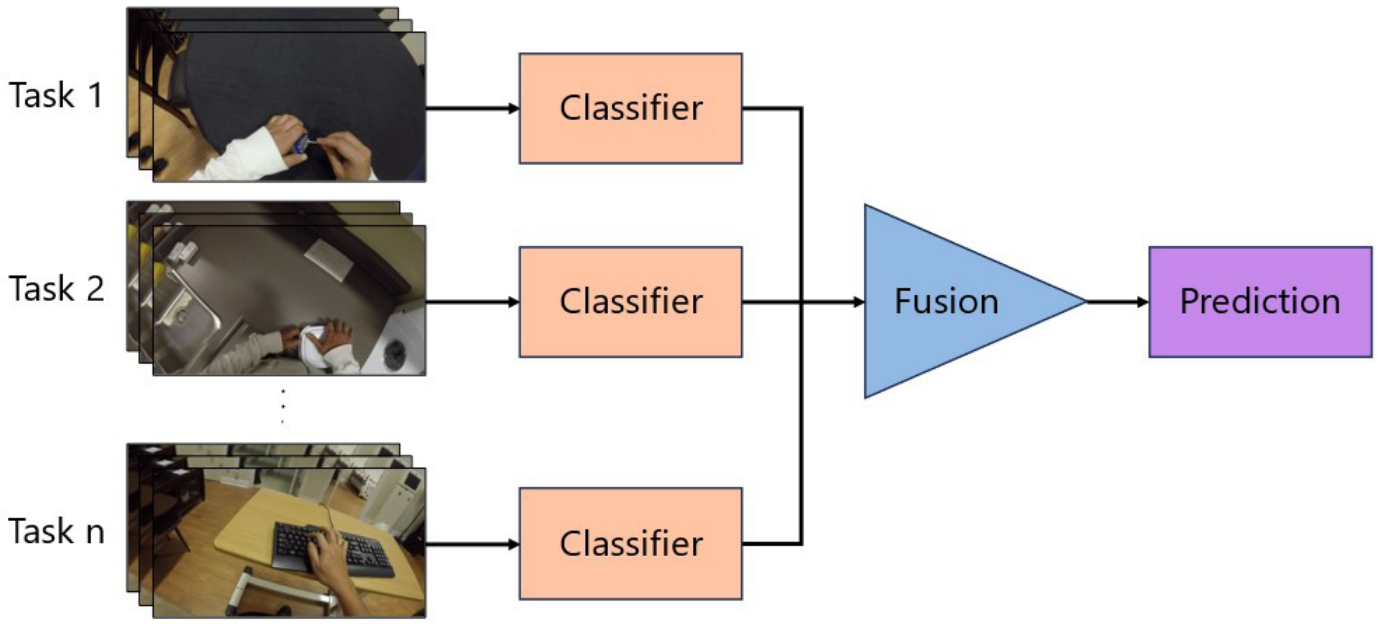


Figure 1: Multi-video deep learning pipeline for hand impairment classification

[P11] Evaluating the Effectiveness of Early Survival Prediction for Coronary Artery Disease Patients: Pre-Treatment vs. Combined Pre and Post Treatment Features

Anita Khalafbeigi, University of Alberta
Sunil Kalmady, University of Alberta
Kevin Bainey, University of Alberta
Robert Welsh , University of Alberta
Padma Kaul , University of Alberta
Russell Greiner, University of Alberta

Introduction

Coronary artery disease (CAD) is commonly treated with coronary artery bypass graft (CABG) or percutaneous coronary intervention (PCI), each with its associated risks and benefits. Our study aims to assess the effectiveness of early survival prediction, utilizing pre-treatment features, compared to using both pre-treatment and post-treatment features. Early survival prediction is crucial for clinicians, offering insights into individual survival distribution (ISD) before interventions.

Methods

We learned survival models from a cohort of CAD patients with stable angina who underwent either PCI or CABG, using several survival models: neural multi-task logistic regression (N-MTLR), deepHit, cox proportional hazards (coxPH), and random survival forest (RSF). In the first experiment, the survival models were trained using patient features measured before receiving the treatment, during cardiac catheterization (CATH), to predict the patient's specific ISD. The second experiment followed a similar approach but utilized both pre-treatment and post-treatment features.

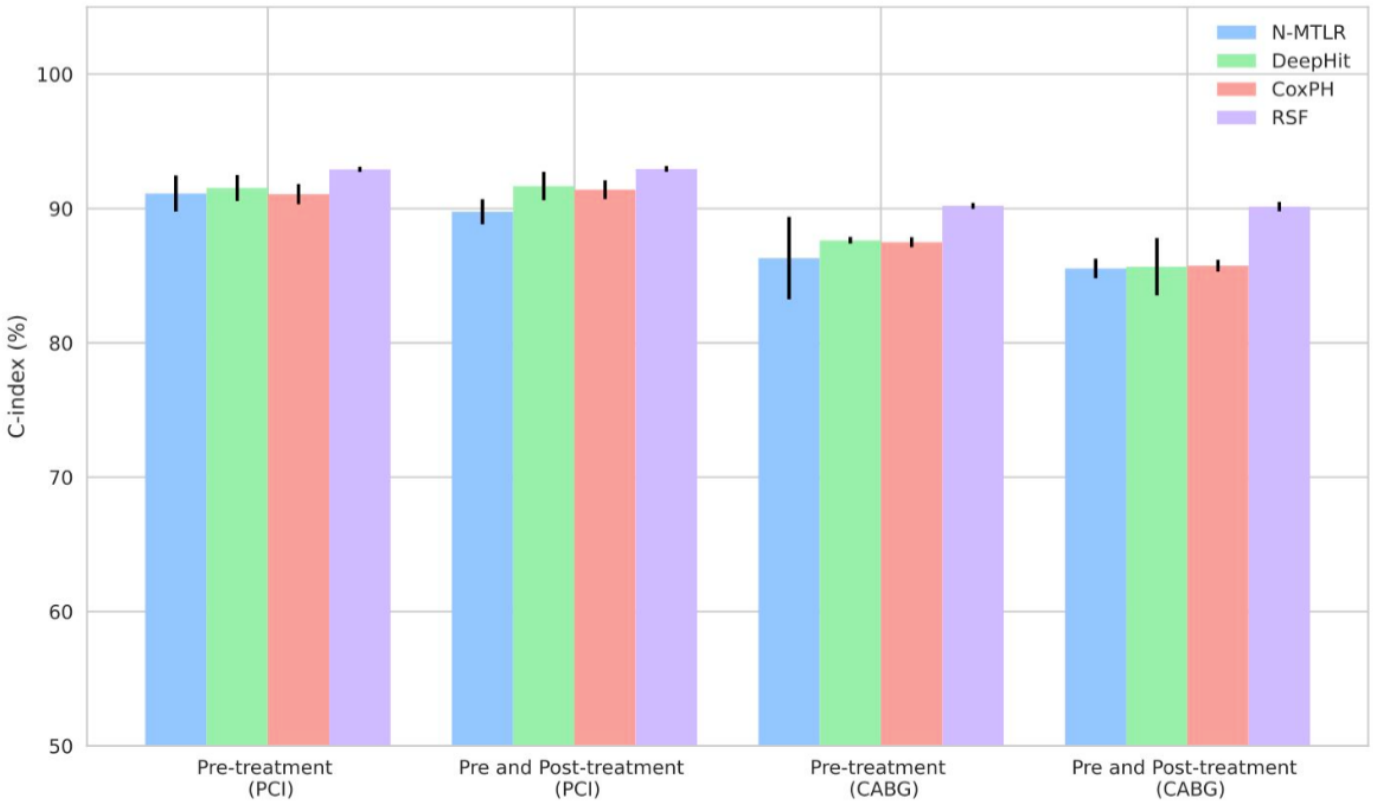
Results

The PCI and CABG datasets used in this study consist of 84,954 and 20,996 records, respectively, with a division of 70% for training and 30% for testing, at random. The performance of the survival models was evaluated using the c-index, which ranged from 84.92% to 92.94%, varying depending on the specific survival model and the dataset used. For CABG patients, the most performant model, RSF, achieved a c-index of 90.14% using only pre-treatment features and 90.19% when using both pre-treatment and post-treatment features. Similarly, for patients undergoing PCI, RSF was the top-performing model, attaining a c-index of 92.91% with pre-treatment features alone, and 92.94% when leveraging both pre-treatment and post-treatment features.

Discussion/Conclusion

The negligible difference in the models' performance with and without post-treatment features demonstrates the effectiveness of early survival prediction for CAD patients. Using this approach, clinicians can reliably predict survival at the time of diagnostic investigation without waiting for the patient to receive treatment.

Supporting information



Comparison of C-Index (%) for different survival models using pre-treatment or combined pre and post-treatment features with standard error

[P13] Evaluation of Synthetic Data Augmentation for Mitigating Covariate Bias in Real World Health Data

Lamin Juwara, University of Ottawa
Khaled El Emam, University of Ottawa

Introduction

Covariate imbalance (data bias) is pervasive in biomedical research, especially when evaluating large-scale observational datasets. In regression modeling, the presence of bias in the training dataset produces imprecise predictions and inconsistent estimations. Our primary objective is to evaluate synthetic data augmentation (SDA) to mitigate bias and compare its performance to commonly used bias-mitigating approaches.

Methods

SDA (boosted decision tree sequential synthesis) was compared to three other types of bias-mitigating methods (rebalancing, algorithmic, and post hoc approaches) using extensive simulations and evaluation on four diverse real datasets. Different types of marginal and conditional bias were induced in the original data, and mitigation approaches evaluated relative to the ground truth on a logistic regression estimation and prediction analytic workload.

Results

In low to medium bias severity (less than 50% missing proportion), SDA produces the results with the least bias (difference between the model estimate and ground truth) and the best precision in estimating the regression coefficient. SDA also produces AUCs comparable to the best approaches while ensuring the best fairness in the training data. In high bias cases (more than 80% missing proportion), the results are not always conclusive: in some cases, SDA suitably mitigates the bias, and among the traditional approaches, rebalancing using random oversampling yields results that are comparable to the original cohort in the high bias scenario.

Discussion/Conclusion

The proposed synthetic augmentation method produces results that are most comparable to the ground truth.

[P14] Identifying Trusted and Ambiguous Regions in Neural Network Predictions: High-Fidelity AI For Image Pathology

Kenneth Wenger, Toronto Metropolitan University
Katayoun Hossein. Abadi, Squint AI Inc
Damian Fozard, Squint AI Inc
Kayvan Tirdad, Toronto Metropolitan University
Alex Dela Cruz, Toronto Metropolitan University
Alireza Sadeghian, Toronto Metropolitan University

Introduction

Recent years have seen a rise in Machine Learning research being applied to the medical domain. Yet, to date we have not seen a massive rollout of ML applications in this domain. The reason is that ML algorithms don't always generalize well beyond the training/testing datasets, and high-confidence mistakes happen too often in production-like environments. Furthermore, the quality of the mistakes is often such that a human pathologist would never make those same mistakes. In this work we present a framework grounded in explainable AI algorithms that quantifies the information gain in a ML model to generate a map of trusted regions and ambiguous regions in the data manifold of the model's latent representations.

Methods

We used the XAI algorithms PacMap and T-SNE to generate a 2D map of internal representations for a state-of-the-art model trained to grade WSI images of bladder cancer. We identified regions in the map with correct predictions, and regions where mistakes were clustered. We called these regions: Trusted, and Ambiguous. We monitored runtime predictions in a simulated production environment where human pathologists were alerted for predictions originating in the ambiguous region, while predictions originating in the trusted regions where accepted without involving humans.

The research involved researchers from Toronto Metropolitan University, and Squint AI Inc.

Results

in our 2D map the trusted region encompassed 75% of the model's predictions.

The ambiguous region encompassed 25% of the predictions.

83% of the model's mistakes were located in the ambiguous region.

Overall accuracy improved from 93% to 97%.

Discussion/Conclusion

By generating a map that localizes 83% of a model's mistakes in a region encompassing 25% of the predictions we can create a hybrid (human-AI) pipeline that automates grading on 75% of the data, reducing cost (in human hours) to 25% of effort, while eliminating 83% of mistakes in a state-of-the-art model

[P15] Iterative XAI Frameworks for Oncology Decision Making: Integrating Expert Feedback to Enhance Cancer Diagnosis

Shermineh Ghasemi , Toronto Metroploitan University

Alireza Sadeghian, Toronto Metropolitan University

Introduction

Addressing the global health burden of cancer requires innovative solutions. Artificial Intelligence (AI) and Deep Learning (DL) hold transformative potential for precision oncology. However, challenges such as data quality, model opacity, and potential biases have limited their full-scale integration. Our research introduces an Iterative Explainable Artificial Intelligence (XAI) framework to overcome these hurdles and enhance early cancer diagnosis.

Methods

Our work utilizes deep learning algorithms trained on diverse patient data including CT and MRI images and Electronic Health Records (EHR). We incorporate patient selection criteria that enhance the predictive accuracy of our models. These include survival data and those without death events during follow-up.

The cornerstone of our research is an innovative, end-to-end explainable framework integrated seamlessly with a deep learning model for cancer screening. This flexible model operates under various learning paradigms, tailoring explanations to a broad spectrum of healthcare stakeholders.

Recognizing the critical role of human expertise in clinical decision-making, we've implemented a feedback loop. This feature enables clinicians to input their expertise, guiding the model based on patient-specific information (Figure. 1).

Our dynamic system perpetually re-evaluates and updates the model's performance as new data or changes in patient conditions occur. The model's performance is evaluated on fidelity, consistency, and transparency metrics, ensuring its reliability.

Results

Our research successfully applied the SSL algorithm based on the VGG-16 neural network to identify cancer cells in images. Now, we are augmenting the dataset with textual reports, patient demographics, and clinical annotations, aiming to integrate an XAI framework into the model. This integration will provide interpretable explanations for clinicians.

Discussion/Conclusion

The impacts of this work are manifold: improved diagnostic accuracy, personalized treatment selection, healthcare efficiency, enhanced patient outcomes, and increased trust in AI-based healthcare systems. Our research ultimately aims to bridge the gap between AI technology and its meaningful adoption in oncology.

[P16] Machine Learning for Assessment of Capsulorhexis Performance in Cataract Surgery

Jonathan ZL. Zhao , University of Toronto
Niveditha Pattathil, Queen's University
Olapeju Sam-Oyerinde, Institute of Ophthalmology, University College London
Amrit Rai, Department of Ophthalmology and Vision Sciences, University of Toronto
Shuja Khalid, Surgical Safety Technologies
Frank Rudzicz, Faculty of Computer Science, Dalhousie University
Tina Felfeli, Department of Ophthalmology and Vision Sciences, University of Toronto; The Institute of Health Policy, Management and Evaluation (IHPME), University of Toronto
Jonathan Rose , University of Toronto

Introduction

Cataracts are the leading cause of blindness worldwide and cataract surgery is one of the most frequently performed surgeries globally. Ophthalmologists are traditionally trained via apprenticeship which entails direct observation and subjective feedback from senior surgeons. The efficacy of training curricula is constrained by a lack of practical and reliable measures to objectively and accurately assess surgical performance. This study aims to evaluate the use of machine learning as an objective means for assessment of cataract surgical skills during capsulorhexis.

Methods

This cross-sectional study uses a sample of 308 videos from a database of deidentified cataract surgery videos performed by faculty and trainee surgeons in the University of Toronto ophthalmology residency program from January 2022 to March 2022. Procedures were graded using a validated rubric for assessing capsulorhexis (scale from 1-5) by three authors and were manually annotated, which served as the ground truth. Videos were segmented using a detectron2 model. Bounding box features that track the locations of the eye and tools as well as capsulorhexis duration were used to train XGBoost models to predict rubric parameters. Performance was measured using F1 score. SHAP was used to determine feature importance.

Results

Intraclass correlations of rubric parameters were 0.83 (motion), 0.86 (regrasp), 0.72 (flap commencement), and 0.50 (formation and circular completion). Average F1 scores of 0.79 (motion; regrasp) and 0.81 (flap commencement; formation and circular completion) were achieved via 5-fold validation by oversampling underrepresented classes. Capsulorhexis duration was the most predictive feature for all parameters except formation. Location features contributed (approximately) equally to all predictions.

Discussion/Conclusion

Machine learning algorithms may yield useful tools for automated and objective assessment of capsulorhexis surgical performance in ophthalmological training. These findings, in a broader context, reveal the promise of machine learning for enhancing surgical training and assessments for other surgical fields.

Supporting information

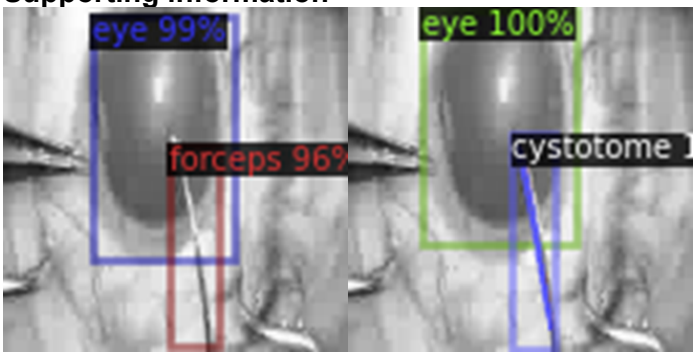


Figure 1: Segmentations of two sample video frames by detectron2 model. These bounding box features, in addition to capsulorhexis duration, are the input features for our XGBoost models.

[P17] Multi-Document Summarization of Patient Neurovascular Radiology Reports

Heet Sheth, St. Michael's Hospital

Introduction

Radiology reports play an essential role in patient care, diagnosis, and treatment. However, reviewing radiology reports can be a time-consuming, error-prone, and overall burdensome process for radiologists. The aim of this research is to present a chronological summary of a patient's neurovascular radiology report history. Based on a thorough search of relevant literature, multi-document summarization for radiology reports has not been attempted yet.

Methods

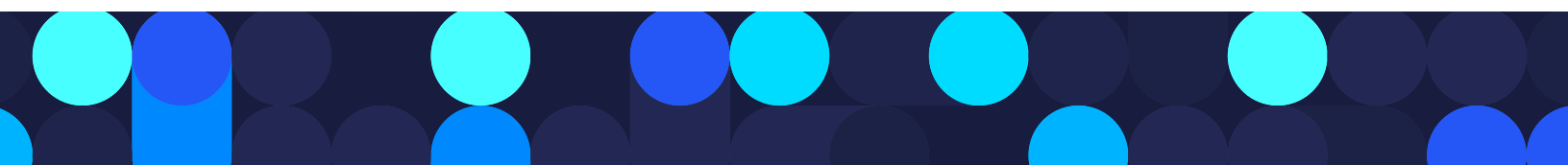
A total of 182 neurovascular radiology reports across 14 patients were examined. In each report, the "Impressions"/"Interpretations" section was designated as the gold standard. Four extractive summarization methods—Frequency-Based, TF-IDF, LexRank, and LSA—were applied across the reports and compared. Summaries were evaluated using the ROUGE metric which incorporates recall, precision, and F1 scores. Further plans for this research include increasing the size the dataset and attempting abstractive summarization approaches.

Results

It was found that the LexRank algorithm outperformed all other algorithms in both recall and F1 scores. The average 3-sentence LexRank summary recall scores for ROUGE-1, ROUGE-2, and ROUGE-L were 0.20, 0.08, and 0.19, respectively. The average 5-sentence LexRank summary F1 scores for ROUGE-1, ROUGE-2, and ROUGE-L were 0.23, 0.10, and 0.22. The average 5-sentence TF-IDF summary had the highest precision scores at 0.46, 0.21, and 0.43 for ROUGE-1, ROUGE-2, and ROUGE-L, respectively. Further plans for this research include performing evaluation on abstractive summarization approaches.

Discussion/Conclusion

Overall, the LexRank algorithm had the best performance. The higher recall and F1 scores can be attributed to the algorithm providing descriptive summaries and including sentences that are negative in nature (often overlooked by other extractive summarization algorithms). TF-IDF performs best for all precision scores as it accounts for a variety of key terms from the report text, ensuring the inclusion of relevant information. The next stage in our research is to evaluate the results of using abstractive summarization methods.



[P18] Predicting self and care staff's evaluation of resident's rehabilitation potential in post-acute care setting using machine learning algorithms

Bonaventure A. Egbujie, University of Waterloo

Anastasiia Avksientieva, University of Waterloo

John Hirdes, University of Waterloo

Introduction

Self-efficacy significantly affects rehabilitation outcome and a positive self and, or care provider belief in the patient's ability to improve in function correlates with eventual good clinical outcomes. Patient-level factors are associated with positive belief in rehabilitation potential are not known. Such information could be used to target rehabilitative care in post-acute care setting.

Methods

Methods: We trained and validated different machine learning algorithms on MDS 2.0 dataset of post-acute care patients in Canada. Validated algorithms were then applied to previously unseen data to predict whether an individual or their care provider believes they could improve in function at the time of admission. Model performance was checked against metrics.

Results

XGBoost algorithm achieved the best overall performance with AUC: 80.7% vs 72.9%, F1: 72.3 vs 72.3%, Sensitivity; 73.5% vs 73.4%, PPV; 71.1% vs 71.2% on training vs unseen data for patient's self-belief, and AUC: 79.6% vs 71.9%, F1: 74.2 vs 74.5%, Sensitivity; 76.7% vs 76.9%, PPV; 71.9% vs 72.3% on training vs unseen dataset for care staff's belief (Table 1). Admission activities of daily living (ADL) score, body mass index (BMI), frailty index and age, social engagement score and having a preference to go home were associated with self and care staff's belief in rehabilitation potential.

Discussion/Conclusion

The findings suggest that machine learning algorithms can predict self-efficacy for rehabilitation potential in post-acute settings. The technique could become a useful tool for triaging patients waiting for rehabilitation, with the potential for better outcomes and reduced wastage.

Supporting information

	Model	AUC	Accuracy	F1	Sensitivity	Specificity	PPV	NPV
All Variables	**Logistic Regression**		69.0%		74.0%	63.3%		
	AdaBoost		69.3%		76.2%	61.6%		
	Random Forest		72.2%		75.8%	68.2%		
	XGBoost		72.2%		76.9%	67.0%		
	Ex-		72.1%		73.6%	70.5%		

[P19] Prediction Algorithms Driving the Making of the National Early Warning System Two Plus Capacity of the Electronic Casualty Card System in Any Disaster Situations – A Component of Vimy Multi-System

Ons Loukil, Ecole Nationale d'Ingénieurs de Carthage, Applicare-AI Inc. and Medical Intelligence CBRNE Inc., Member esearch Regroupment in Artificial Intelligence Apllied to Critically Ill Children CDSS LabCHU Ste-Justine
Wala Bahri, Institut National des Sciences Appliquées et de Technologie, Applicare-Ai Inc. and Medical Intelligence CBRNE Inc.
Olfa Lamouchi, Ecole Nationale d'Ingénieurs de Carthage
Abderrazak Jemai, Institut National des Sciences Appliquées et de Technologie
Philippe Jouvet, Université de Montréal (Faculté de médecine), Head of Research Regroupment in Artificial Intelligence Apllied to Critically Ill Children CDSS LabCHU Ste-Justine
Mariem Abid, Head of Applicare-AI Inc. and member of esearch Regroupment in Artificial Intelligence Apllied to Critically Ill Children CDSS LabCHU Ste-Justine
Stephane Bourassa, Ecole de Technologie Supérieure (Université du Québec à Montréal), Université de Montréal & Research Regroupment in AI Applied to Critically Ill Children CDSS Lab CHU Ste-Justine, Medical Intelligence CBRNE Inc. & National Defence Department

Introduction

VIMY Multi-System Research Program, for the mass casualty management in any disaster situations, started with the development of the Electronic Casualty Card System (ECCS). This study aims to upgrade the triage component of the ECCS with new triage terms and prediction algorithms in real-time within an informatic laboratory setting.

Methods

This prospective design study focuses on the pre-processing step of the machine learning lifecycle. The triage system named NEWS2+ is configured with five triage terms, and prediction algorithms applied on time series of vital signs at one-second precision obtained by applying linear interpolation (Figure 1) . For each vital sign, NEWS2+ alerts were incorporated, along with hypoxemia and acute respiratory distress severities. Clinical data from MIMIC III and IV databases were processed, with a focus on handling missing values using two distinct imputation methods (K-nearest neighbors (KNN), ImputEHR). A deep learning method based on a long short-term memory (LSTM) was developed to predict patient conditions based on the predefined triage terms, at prediction windows, starting from 600 seconds and decreasing in 60-second increments until reaching 30 seconds.

Results

As preliminary results, analyses were conducted on 222 hypoxemic and acute respiratory distress patients out of 4,048 patients from the MIMIC III and IV. The KNN imputation yielded the best results obtained, with a root mean squared error (RMSE) of 3.80 and a mean absolute error (MAE) of 1.37, compared to the ImputEHR technique with RMSE of 4.36 and MAE of 1.52. Among the other models used so far, including XGBoost and Gated Recurrent Units, LSTM achieved the highest accuracy of 0.81 for predicting 120 seconds with a gap of 120 seconds on the testing data.

Discussion/Conclusion

Preprocessing on MIMIC III & IV addresses missing data and ensures reliable results, incorporates one-second precision-based interpolation to align ECCS's innovation for real-time disaster medical responses.

Supporting information

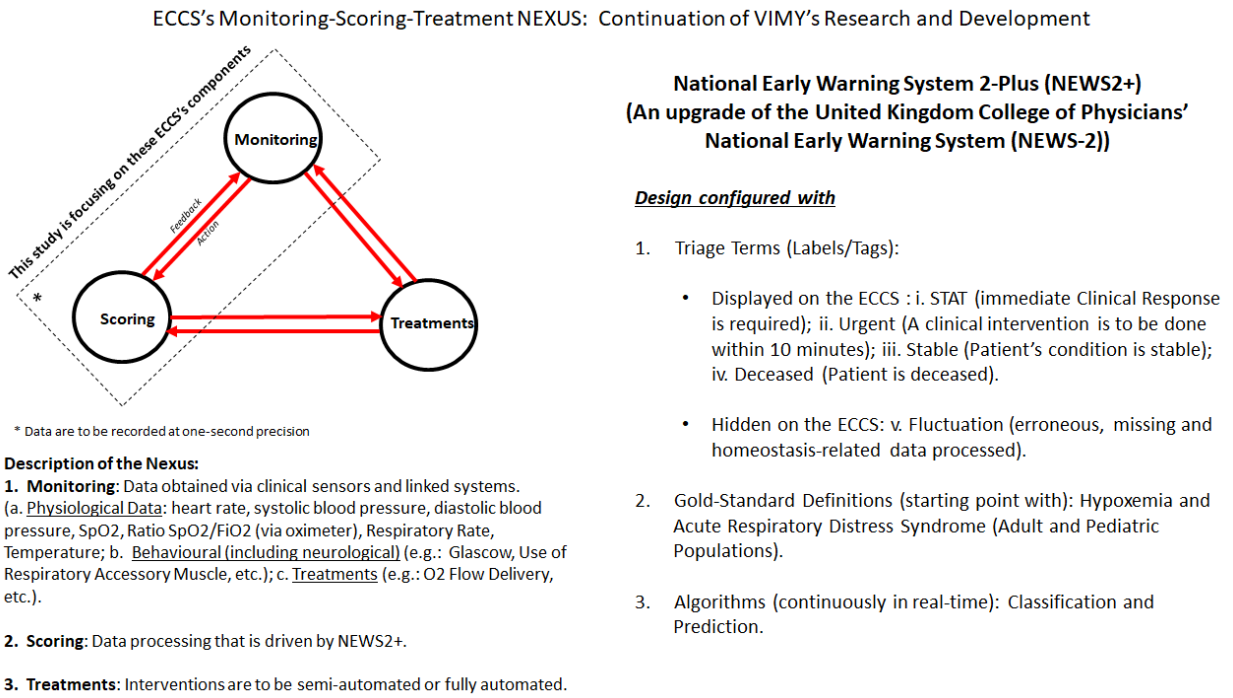


Figure 1. Illustration on Concise Details of the Electronic Casualty Card System Research and Development (VIMY Multi-System)

[P20] Prediction of Depression Relapse in Youth and Adolescence from Ecological Momentary Assessment (EMA) data through Machine Learning

Cheuk Hei Chung, Department of Human Biology, University of Toronto, Toronto, Canada
Cheuk Hei Chung, Department of Human Biology, University of Toronto, Toronto, Canada
Mai Ali, Department of Electrical and Computer Engineering, University of Toronto & Vector Institute for Artificial Intelligence, Toronto, Canada
Christopher Lucasius, Department of Electrical and Computer Engineering, University of Toronto, Toronto, Canada
Tanmay Patel, Division of Engineering Science, University of Toronto, Toronto, Canada
Deepa Kundur, Department of Electrical and Computer Engineering, University of Toronto, Toronto, Canada
Peter Szatmari, Cundill Centre for Child and Youth Depression at the Centre for Addiction and Mental Health, Toronto, Canada
John Strauss, Vancouver Island Health Authority, Vancouver, Canada
Marco Battaglia, Cundill Centre for Child and Youth Depression at the Centre for Addiction and Mental Health, Toronto, Canada

Introduction

Major depressive disorder (MDD) in youth (MDD-Y) is a leading global cause of disability. Depression in youth and adolescents causes a high disease burden and has a high relapse rate. Therefore, it is crucial to predict the possibility of relapse in youth and adolescents, providing medical personnel with the opportunity to prevent another depressive episode from happening.

Methods

Ecological Momentary Assessment (EMA) data is collected through the Depression Early Warning study conducted at the Centre for Addiction and Mental Health, Toronto, Canada. Smartphones were used to collect EMA data from one-hundred-thirty clinically depressed adolescents. Ecological questions are asked to the participants 5 times a day for seven days. The rich EMA data contains crucial physiological and behavioural information allowing for analysis and prediction of possible relapse through cues of depression, including increased mood and sleep changes.

Results

The gold standard clinical survey is used as a baseline for depression relapse. Different physiological and behavioural information types in the EMA data are then used to predict depression relapse through different machine learning methods. When using the sleep quality (one aspect of the EMA data) reported by the patients to predict depression relapse with logistic regression, an accuracy of 90.1% was achieved. Combining all elements of the EMA data will further strengthen the prediction result.

Discussion/Conclusion

The high percentage of prediction accuracy might be due to the imbalance of patient relapse in the data. Therefore, a more balanced set of data could lower the accuracy. Nonetheless, with the vast amount of information contained in EMA data, the implementation of machine learning in the analysis of clinically depressed patients could provide a crucial forecast for psychiatrists, allowing them to improve the health of patients by preventing relapse and medicine as a whole.

Supporting information

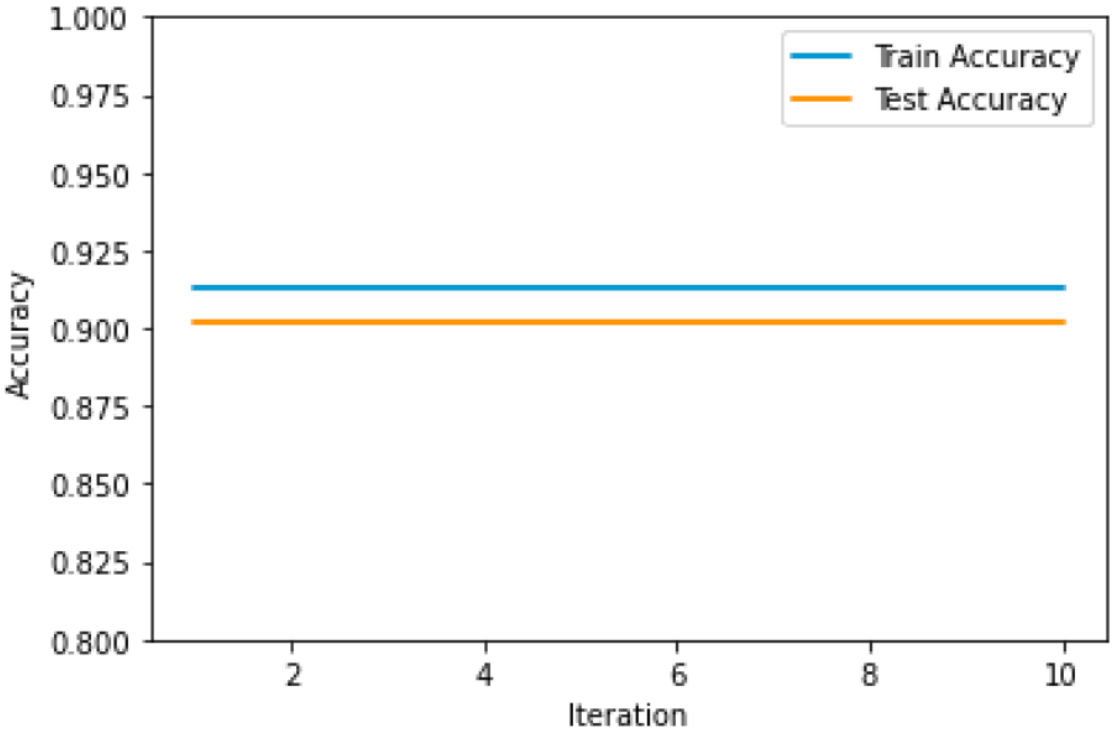


Figure 1: The Logistic Regression accuracy from using the EMA sleep quality data and baseline relapse.

[P21] Robust Detection of Seizure Onset Zone in Patients with Epilepsy using a Novel Graph-based Neural Network

Milad Lankarany , University of Toronto

Andrew T. Sage, Institute of Medical Science, Temerty Faculty of Medicine, University of Toronto

Introduction

Epilepsy is the most common serious neurological disorder in the world, affecting over 50 million individuals worldwide. Today, the gold-standard treatment for those who are medically refractory (failed medical treatment) is to surgically remove the seizure onset zone (SOZ), the area of the brain believed to cause seizures: the main symptom of epilepsy. Unfortunately, only 50-70% of resective surgery outcomes are successful, which is partly attributed to poor SOZ localization (SOZ-L). The current methods for SOZ localization mainly suffer from the lack of scalability and interpretability that limit their usability for clinical translation and practices.

Methods

My postdoctoral fellow, Dr. Alan Diaz, and I have recently developed a data pipeline for processing and managing iEEG data collected from patients with Epilepsy. Our work can be found here: https://www.biorxiv.org/content/10.1101/2023.06.02.543277v1.full.pdf.

All codes are available in Python, we developed a novel graph neural network (GNN) with a self-supervising algorithm, and extend it to account

for node and edge features pertinent to the iEEG electrodes and functional connectivity measurements.

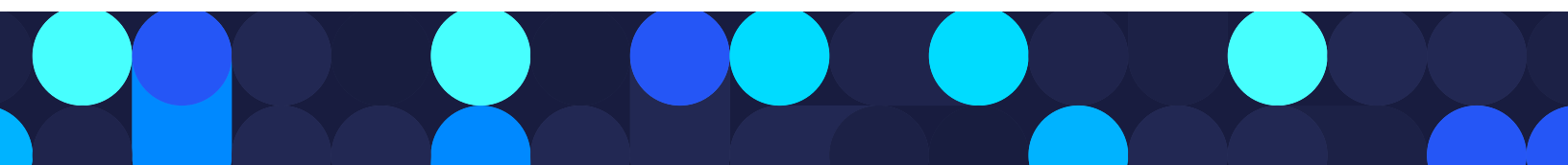
Results

We assessed the performance of the GNN on iEEG data using 9 different graph representations. We

consider 4 metrics to assess the performance of the GNN model: accuracy, F1-score, AUC, and loss. Using the proposed features from the iEEG data (e.g., functional connectivity) in graphs with both nodes and edge embeddings, we improved the performance of the GNN model with a median near 90% (compared to GRs without edge embeddings).

Discussion/Conclusion

To assist human experts, we propose an AI-driven solution that automates seizure detection and SOZ-L. We have developed a data pipeline for processing iEEG data from patients with epilepsy and tested its performance on publicly available data for the proof-of-concept phase. We aim to deploy our solution in the Epilepsy Program at the Krembil Brain Institute at the Toronto Western Hospital.



[P22] Sex-Based Differences in Speech Coherence for classifying Schizophrenia using BERT similarity

Alban Voppel , McGill university

Paulina Dzialoszynski , PEPP - London healthy Sciences centre

Sabrina Ford , Western University

Betsy Schaefer , LHSC - University Hospital

Lena Palaniyappan , Department of Psychiatry, McGill University

Introduction

Speech and language impairments are prominent in schizophrenia-spectrum disorders, and advances in natural language processing (NLP) enable quantification of these characteristics in these psychiatric disorders. Speech NLP has the potential to make AI-applications for disease classification and prognostication more accessible now than ever before. However, sex differences in symptomatology and demographics in schizophrenia remain unexplored in NLP studies, potentially leading to lower generalizability or bias. This study investigates sex-based differences in speech coherence in schizophrenia to clarify sex-based biases that need to be addressed for potential clinical usage of NLP.

Methods

We included 58 participants, comprising schizophrenia-spectrum diagnosed individuals (n=30) and healthy controls (n=28), age and sex-matched. Participants completed a standardized interview task to record speech, using the DISCOURSE protocol. We used BERT, a large language model, to calculate sentence-to-sentence predictability measures of transcribed interviews, resulting in per-participant sentence predictability measures. Mean, minimal, maximum, and standard deviation of predictability were extracted per participant and compared, splitting for sex.

Results

Significant differences in sentence predictability were found between schizophrenia and control groups ($p = 0.011$) for minimum sentence predictability. Upon sex-based comparison, the difference in minimum sentence predictability between healthy men and men with schizophrenia was non-significant ($p = 0.227$). However, among women, a significant difference in minimum sentence predictability was observed ($p = 0.014$) between healthy controls and subjects with schizophrenia.

Discussion/Conclusion

Our study revealed sex-based differences in sentence predictability between individuals with schizophrenia and healthy controls. Schizophrenia patients showed impaired speech compared to controls, consistent with prior findings. However, when comparing men with schizophrenia to healthy men, the difference in coherence was non-significant, while a significant difference persisted in women. This underscores the importance of considering sex-specific characteristics in NLP studies for schizophrenia. Understanding these differences would improve NLP's clinical utility for predicting and assessing symptoms, reducing bias, and enhancing specificity.

[P23] Smoking Cessation Interventions in South Asian Region: a systematic scoping review

Rameesha Rehmani , University Health Network

Introduction

Tobacco use is one of the most averted causes of morbidity and mortality. Since 2005, World Health Organization Framework Convention on Tobacco Control (WHO-FCTC) has provided an effective global strategy for tobacco control. Many countries throughout the world have successfully lowered cigarette smoking rates. However, the frequency of cigarette smoking is increasing in emerging countries, emphasizing the need for immediate intervention. By systematically examining relevant recently published and unpublished material, this scoping review aims to investigate the extent and type of Smoking Cessation (SmC) treatments and associated factors in the South Asian Region (SAR).

Methods

The Joanna Briggs Institute (JBI) framework guides the scope of this review.

PubMed, EBSCO CINAHL Complete, Cochrane Library, ProQuest Dissertation and Theses, local internet pages, and other grey literature sources were scoured for relevant literature. A total of 573 literature sources were reviewed. Finally, 48 data sources were included for data extraction and analysis, following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram.

Results

Most FCTC-recommended smoking cessation tactics (articles) were either ignored or treated in a discordant manner by various anti-smoking groups in SAR. Lack of awareness, inadequate enforcement of anti-smoking laws, and socio-cultural acceptance of tobacco use were identified as major hurdles to the success of smoking cessation initiatives. A greater degree of awareness about the dangers of smoking and the benefits of quitting, effective implementation of anti-smoking laws, smoking cessation trained healthcare professionals, networks of support, and community aversion to cigarette smoking were identified as facilitators of smoking cessation interventions.

Discussion/Conclusion

The neglected or uncoordinated guidance of the FCTC on smoking cessation strategies has led to continued increases in smoking prevalence in developing countries. There is a need for considerations of local barriers and a reorientation of smoking cessation strategies in SAR.

[P24] Summarizing Clinical Trials with Large Language Models

Bryant Lim , Department of Medicine, University of Toronto
Emily Bartsch , Department of Medicine, University of Toronto
Katarina Zorcic , Department of Medicine, Sinai Health System
Tamara Van Bakel , Department of Medicine, Sinai Health System
Michael Fralick , Department of Medicine, Sinai Health System

Introduction

For physicians, keeping up with emerging clinical trials is essential for practicing evidence-based medicine, but also challenging given their academic and clinical responsibilities. Here, we explored the potential of open-access large language models to generate concise and reliable summaries of clinical trials to help medical professionals digest the medical literature more efficiently.

Methods

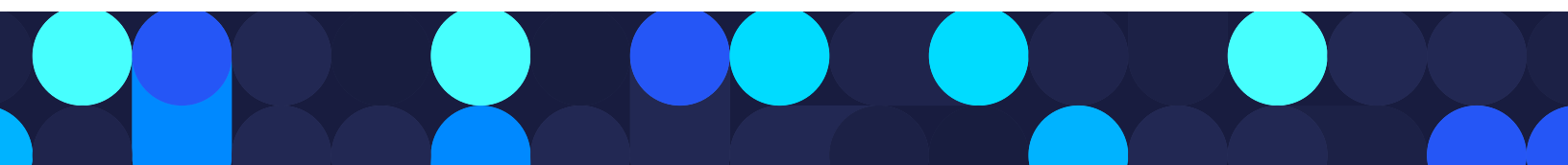
Publications of randomized clinical trials from the New England Journal of Medicine, Annals of Internal Medicine, Journal of the American Medical Association (JAMA), JAMA Internal Medicine, and Lancet were extracted from the MEDLINE database (accessed 28/05/2023). Using the text-davinci-003 application programming interface (API; OpenAI), input prompts were optimized to generate abstract summaries. Summaries were then published in a biweekly newsletter targeted at internal medicine physicians. Accurate capture of key trial information was manually evaluated. Newsletter readers (n = 11) were surveyed to assess the quality of the newsletter.

Results

Summaries were generated from 96 randomized clinical trial abstracts. The accuracy for reporting the correct information was 97.0% for study phase, 92.1% for blinding, 84.4% for sample size, 96.7% for patient population, 93.7% for comparison groups, and 94.8% for primary outcome. Four summaries (4.2%) included information that was no present in the abstract. The average score out of five for reader-assessed summary quality and its utility as an aid for keeping up with the medical literature were 4.57 and 4.69, respectively.

Discussion/Conclusion

In the 96 clinical trial abstract summaries generated by the text-davinci-003 API, accurate inclusion of key trial information ranged from 84.4% to 97.0%. Hallucinations occurred in less than 5% of generated summaries. The high ratings of summary quality and utility by readers underscore the potential for implementing automated summaries in newsletters to disseminate clinical trial findings.



[P25] Supervised Machine Learning Pipeline to Classify Pain using sEMG and MMG during Neuromuscular Electrical Stimulation to Combat Intensive Care Unit Acquired Weakness

Meg Sharma, Cleveland Clinic

Alireza Sadeghian, Toronto Metropolitan University

Introduction

Up to 1 million patients worldwide may develop the syndrome of weakness termed ICU-acquired weakness (ICUAW), which can lead to prolonged mechanical ventilation, hospital length of stay and mortality. NMES has been clinically demonstrated to improve muscle strength and endurance in immobilized patients and revert muscle wastage in ICU long-term patients. Surface electromyography (sEMG) and mechanomyography (MMG) signals can be collected during NMES delivery to automate the detection of pain during therapy delivery.

Methods

Rectus femoris, vastus lateralis and tibialis anterior muscle groups were stimulated in a virtual, self-administered setting using custom hardware on 12 healthy participants using a virtual, REB-approved data collection protocol by University of Toronto physical therapy students. sEMG and MMG signals were preprocessed, manually and auto-feature extracted, and trained using classic and deep learning algorithms in MATLAB and Python. For model training, a Leave-One-Subject-Out cross-validation strategy was used to avoid overfitting and a binary cross entropy loss function was selected for the binary classification task and the models were evaluated using the F1 score due to the unbalanced number of no pain to pain labels in the dataset.

Results

A novel muscle contraction onset approach was implemented. The performant model was a L2-regularized Long-Term Short Memory neural network with dropout and batch normalization. This tuned-LSTM outperformed fully connected neural networks and convolutional neural networks, with a F1 score of 84.96%.

Discussion/Conclusion

This novel machine learning pipeline demonstrates the feasibility of using the sEMG and MMG biosignals to characterize and predict pain using a low-cost, low-runtime computational approach that is scalable to production. The results show promise in further automating NMES delivery in the ICU and can be extended to non-healthy subjects in the future. Future work will explore additional architectures to improve performance, train using nonhealthy subject data and leverage data augmentation strategies to reduce overfitting.

[P26] Using Artificial Intelligence To Label Free-Text Operative And Ultrasound Reports For Grading Pediatric Appendicitis

Waseem Abu-Ashour, McGill University Health Center – Research Institute (MUHC-RI)

Dan Poenaru, McGill University & McGill University Health Center – Research Institute (MUHC-RI)

Sherif Emil, McGill University & McGill University Health Center – Research Institute (MUHC-RI)

Introduction

Small sample sizes and unstructured electronic medical records (EMRs) can hinder personalized data science methods in managing pediatric appendicitis. Chatbots utilizing artificial intelligence (AI) and large language models (LLMs) can help restructure free-text EHRs. This study aims to evaluate the quality of data extraction by ChatGPT-4 in comparison to human data collectors.

Methods

AI models were trained to preoperatively determine pediatric appendicitis severity using detailed preoperative and operative data from 2100 children who underwent surgery for acute appendicitis between 2014-2022. This data was extracted by trained data collectors (medical students and research assistants) who achieved satisfactory Kappa scores. We directed ChatGPT-4 to reorganize free text from 103 randomly selected anonymized ultrasound and surgical records in the dataset according to established variables and coding options. It was also asked to calculate the Pediatric Appendicitis Grade (PAG) from the surgical report. A pediatric surgeon then analyzed all data, identifying inaccuracies in each method.

Results

Among the 44 ultrasound (42.7%) and 32 surgical reports (31.1%) that exhibited discrepancies in at least one aspect, 98% of errors were associated with manual data collection. The PAG was incorrectly determined manually in 29 cases (28.2%), and by ChatGPT-4 in only 3 cases (2.9%). Overall, the AI chatbot prevented misclassification in 59.2% of records including both types of reports, and approximately extracted the required data.

Discussion/Conclusion

The AI chatbot, ChatGPT-4, demonstrated superior performance in data extraction accuracy from ultrasound and surgical reports compared to manual methods, and correctly determined the PAG score. Despite the need for broader validation and thorough assessment of data security concerns, these innovative AI tools present significant potential for enhancing the precision and efficiency of research data collection.

[P28] Utilization of unsupervised image feature-based clustering to scale classifier design in histopathology

Minli Chen, Department of Medical Biophysics, University of Toronto
Kevin Faust, University of Toronto
Alberto Leon, University Health Network
Michael Lee, Univerisity of British Columbia
Marly Mikhail, McMaster University
Dimitrios Oreopoulos, Univerisity of Western Ontario
Phedias Diamandis, Department of Laboratory Medicine and Pathobiology, University of Toronto

Introduction

Histopathological analysis of patient tissue is a powerful clinical tool for diagnosis and study of human disease but is challenged by inter-subjective variation and need for sub-specialized pathologists. While there is much excitement surrounding use of artificial intelligence (AI) to automate and increase objectivity of microscopic examination, current use cases represent largely proof-of-concepts limiting widespread adoption. To address this, I describe a prototype AI workflow that empowers pathologists to design and share tissue classifiers without need for any sophisticated coding or complex collaborations.

Methods

An image feature-based clustering approach that was trained on nearly 1 million pathology annotated image patches extracted from over 1000 brain tumours was re-deployed to recognize tumour heterogeneity in tumours of various organs in the cancer genome atlas (TCGA). This image feature-based clustering workflow was then used to develop 18 organ specific classifiers that identifies tumor/non-tumor patterns in various organs with annotations of 10 participants with pathology training. In this workflow, an image cohort was selected by pathologists to train their classifier. This cohort was run through convolutional neural network and label for each clustered region is generated. Pathologists then modified the automatic label with organ-specific annotations, and merged annotations of each class is pooled to make new classifier.

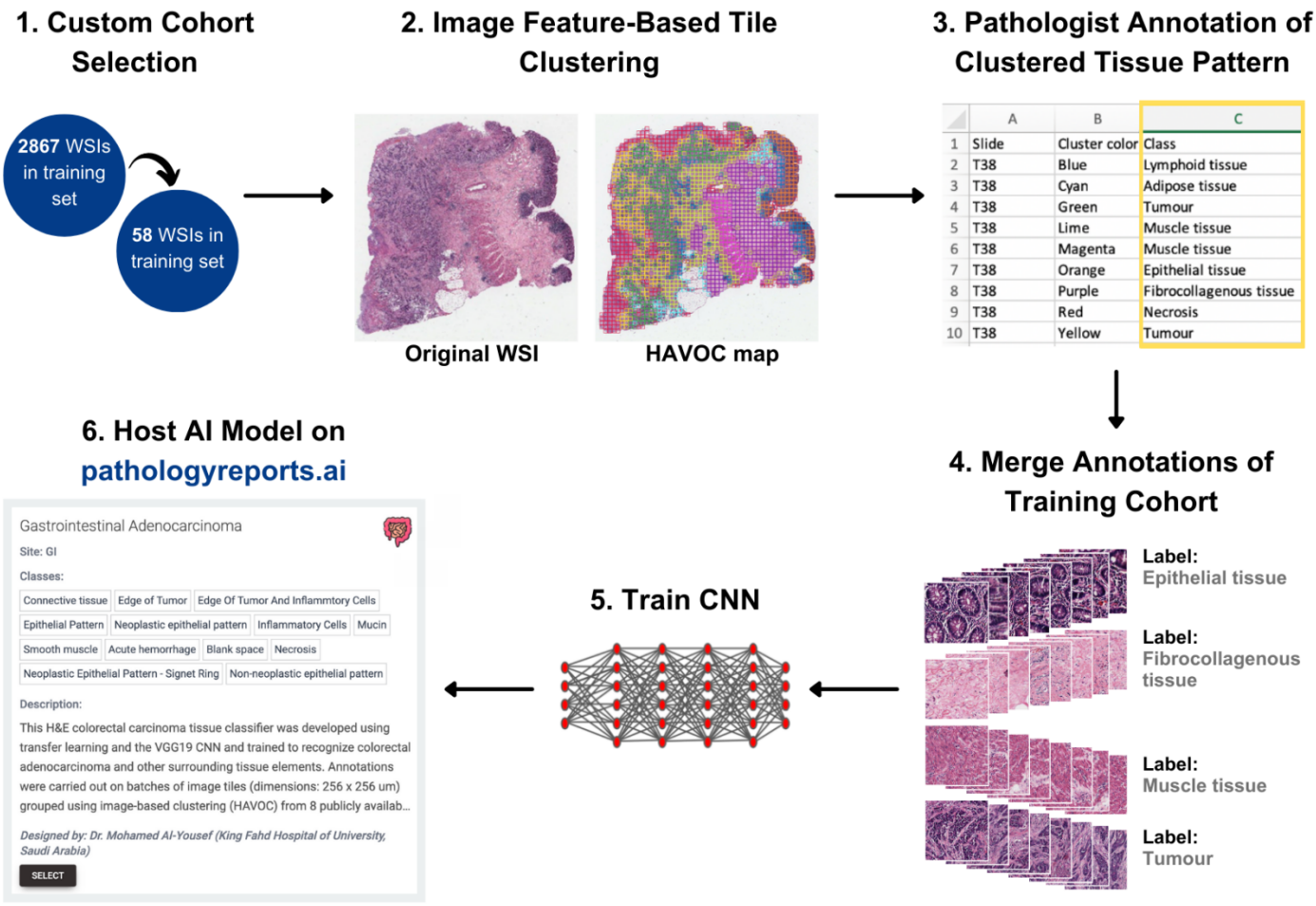
Results

The image feature-based clustering workflow generated tumour heterogeneity map, cellularity counting in different regions and Pearson correlation value of regional morphology similarity. This approach is tested across 33 tumor types and is found to be generalizable on tumors of other organs. 6 organ-specific classifiers were trained following the workflow and 12 classifiers are under development. Notably, the ovarian cancer classifier reached above 83% accuracy between its true labels and predicted labels in all of its 8 tissue classes.

Discussion/Conclusion

Scaling this workflow through crowd-sourcing could allow for large-scale development of custom classifiers across all relevant histopathology.

Supporting information



[P30] Whole-Person Biopsychosocial Subtyping of Mental Illnesses in Treatment-Seeking Youth

Denise Sabac, CAMH/Institute of Medical Science
Mu Yang, CAMH/Dalla Lana School of Public Health
Jimmy Wong, Centre for Addiction and Mental Health
Daniel Felsky, CAMH/Dalla Lana School of Public Health

Introduction

Current categorical classifications used to diagnose mental illness in youth are unreliable due to high rates of comorbidities and similarities in molecular mechanisms across diagnoses. Psychosis Spectrum Symptoms (PSS), which may predispose youth to serious future illness, are particularly challenging to predict or treat, since their biological and environmental determinants are poorly understood. Existing data-driven, biopsychosocial studies aiming to characterize PSS are limited by non-clinical populations, small sample sizes, and few data types. Here, we use sophisticated network based methods to cluster treatment-seeking youth using several biopsychosocial data types.

Methods

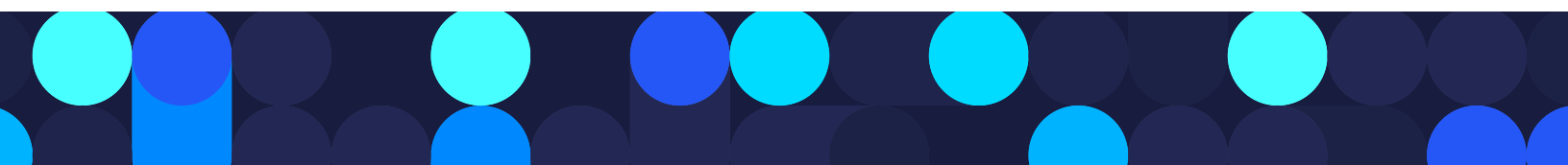
We analyzed 202 participants from the Toronto Adolescence and Youth CAMH Cohort Study (mean age=18.6y) who underwent clinical, cognitive, and neuroimaging-based assessments at baseline. Association-signal-annotation boosted similarity network fusion (ab-SNF) was used to cluster participants by integrating k=246 features across four data types: sociodemographics (e.g. age, sex assigned at birth, educational attainment – k=17); cognitive performance (NIH Toolbox Battery – k=17); regional cortical thickness and subcortical volumes derived from T1-weighted MRI (k=86); and white matter microstructural integrity (i.e. tract-wise fractional anisotropy and mean diffusivity – k=126). Feature weights were assigned based on association strength with PSS.

Results

Ab-SNF identified three distinct participant clusters (cluster 1 n=106, cluster 2 n=26, cluster 3 n=70), with mean ages of 19.5y, 15.5y, and 18.5y, respectively. The top-contributing-feature to this clustering solution was biological sex at birth, with cluster 1 consisting mostly of females (99.1%). Other top features were predominantly cortical thickness measures, with cluster 2 having the greatest average thickness scores. Cluster assignment was significantly associated with the presence of PSS (p=0.0019), albeit this difference was largely due to the strong weighting of sex-specific factors.

Discussion/Conclusion

Preliminary analyses demonstrate the potential for biopsychosocial subtyping of mental illness in treatment-seeking youth; however, ongoing work must account for observed sex differences which are most salient to the clustering algorithm.



[P31] A Comprehensive Study of Radiomics-based Machine Learning for Liver Fibrosis Detection in CT Images

Jay Yoo, University of Toronto
Khashayar Namdar, University of Toronto
Sean Carey, University Health Network
Sandra Fischer, University Health Network
Chris McIntosh, University of Toronto
Farzad Khalvati, The Hospital for Sick Children
Patrik Rogalla, University Health Network

Introduction

Liver fibrosis is a key predictor of mortality from nonalcoholic fatty liver disease. Early detection of liver fibrosis can help cure or prevent disease progression as the process leading to end-stage liver disease is reversible when detected early. We performed a comprehensive study of machine learning (ML)-based approaches to fibrosis detection on computed tomography (CT) images using radiomics.

Methods

For this REB-approved retrospective study, radiomic features were extracted from spherical Regions of Interest (ROIs) on CT images of 169 patients (mean age, 51.19 ± 14.70, 101 men) who underwent simultaneous liver biopsy and CT examinations from October 2019 to April 2021. Models were trained 100 times on ROIs corresponding to biopsy locations and evaluated on ROIs distant from biopsy locations. Different combinations of image contrast, image normalization, and ML model were evaluated based on their mean test Area Under the Receiver Operating Characteristic curve (AUC). The top features were determined based on their frequency among the best settings after Boruta feature selection.

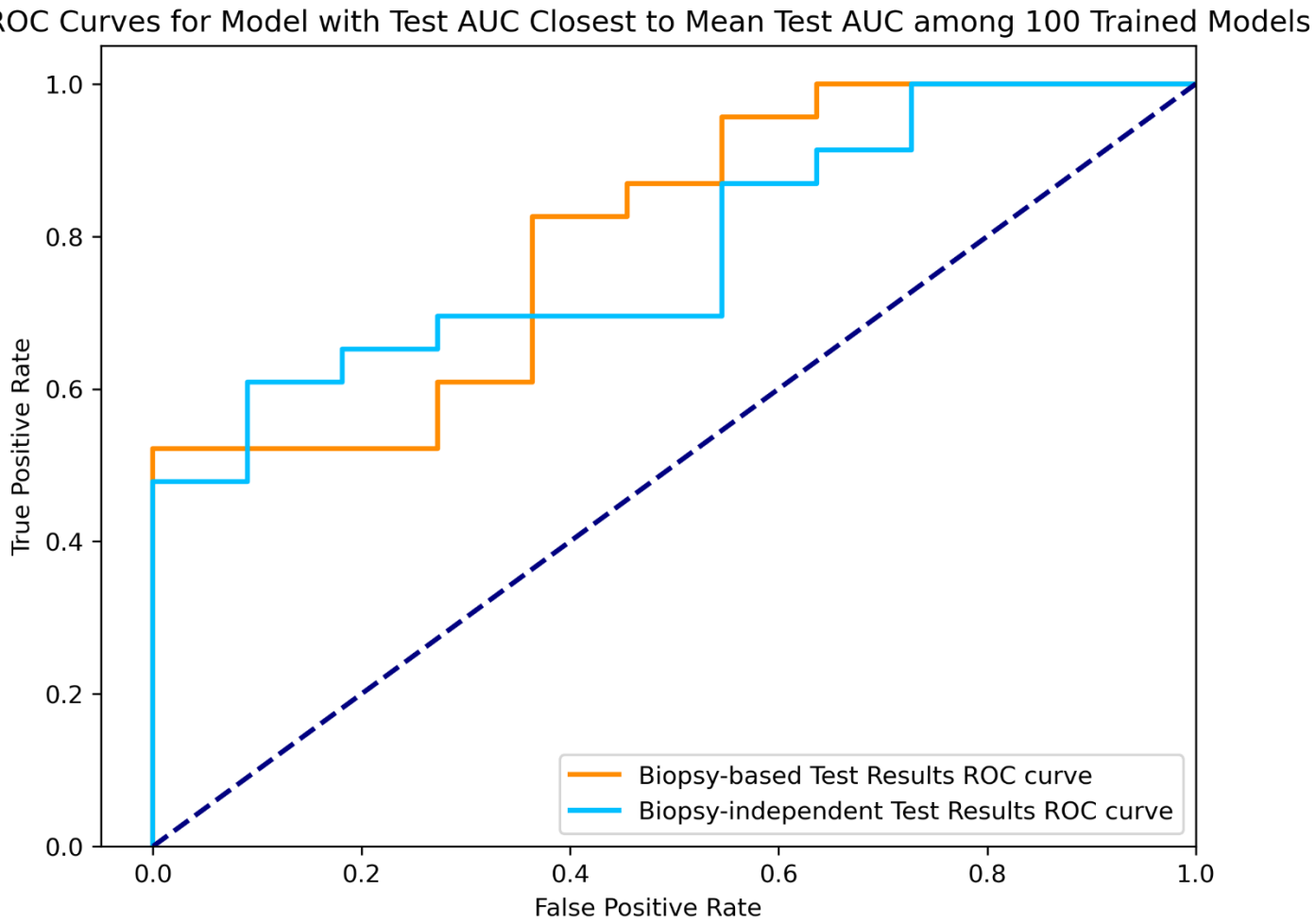
Results

Using maximum, energy, kurtosis, skewness, and small area high gray level emphasis features extracted from non-contrast enhanced (NC) CT normalized using gamma correction with $\gamma = 1.5$ to train logistic regression models performed best. These models were effective for liver fibrosis detection on both biopsy-based (AUC, 0.8041; 95% confidence interval (CI): 0.8032, 0.8049) and biopsy-independent (AUC, 0.7833; 95% CI: 0.7821, 0.7845) ROIs.

Discussion/Conclusion

Logistic regression models trained on maximum, energy, kurtosis, skewness, and small area high gray level emphasis features extracted from NC images normalized using gamma correction ($\gamma = 1.5$) are effective for liver fibrosis detection in CT images. Enabling radiomics-based liver fibrosis detection in CT images may contribute to early non-invasive detection of liver fibrosis with the potential to be seamlessly incorporated into multi-phase liver protocols, which obviates the need for additional dedicated acquisitions.

Supporting information



[P32] An Automatic Retinal Image Analysis Method to Estimate Comprehensive Glaucomatous OCT Parameters Using Two-dimensional Images

Chuying Shi, The Chinese University of Hong Kong
Jack Lee, The Chinese University of Hong Kong
Di Shi, Fudan University Shanghai Cancer Center
Gechun Wang, Zhongshan Hospital, Fudan University
Fei Yuan, Zhongshan Hospital, Fudan University
Benny Zee, The University of Hong Kong

Introduction

Glaucomatous parameters including optic nerve head (ONH) parameters (rim area, disc area, average cup-to-disc-ratio (C/D), vertical C/D, and cup volume), average retinal nerve fibre layer (RNFL) thickness, and average and minimum ganglion cell-inner plexiform layer (GCIPL) thickness can be measured by OCT on two-dimensional (2D) non-mydriatic retinal images.

Methods

To generate exact features on the region of interests (ROIs) of the optic disc and macula, we localized and segmented them using the intensity-based method. Then, we applied transfer net ResNet-50 deep network and the ARIA automatic features generation approach to generate features and used the Glmnet approach to select the associated features. Ten-fold cross-validation with random forest (RF) bagtree models was applied in the primary dataset. Finally, we confirmed the estimation performance of our RF models in the validation dataset.

Results

In both the primary dataset and validation datasets, there were significant correlation between estimated and true values in all OCT parameters ($p < 0.001$) and the three parameters with the best performance are RNFL thickness (correlation coefficient r=0.640; RMSE=11.998; MAE=9.096, and r=0.466; RMSE=13.834; MAE=10.783), average C/D (r=0.648; RMSE=0.128; MAE=0.093, and r=0.511; RMSE=0.130; MAE=0.100), and vertical C/D (r=0.678; RMSE=0.127; MAE=0.092, and r=0.529; RMSE=0.130; MAE=0.098).

Discussion/Conclusion

It has the potential to be a convenient, cost-effective, and accurate tool to assist clinicians to diagnose, monitor, and indicate the severity of glaucoma and other diseases related to OCT parameters.

[P33] An integrated toolkit for measuring fairness of risk predictive models for healthcare

Shaina Raza , Vector Institute for Artificial Intelligence
Franklin Ogidi, Vector Institute
Vaakesan Sundrelingam, Unity Health Toronto
Amol Verma, Unity Health Toronto
Fahad Razak, Unity Health Toronto
Janice Da Silva , Vector Institute
Amrit Krishnan , Vector Institute

Introduction

Healthcare is undergoing a significant transformation, powered by the integration of data-driven systems. Machine learning (ML) and predictive analytics provide healthcare professionals with valuable decision-making tools, however guarding against unintentional bias is vital to ensure fair outcomes across different patient subgroups. Fairness in healthcare requires transparent, impartial, and inclusive treatment for all.

In our continuous endeavor to detect and measure bias, we are incorporating a new tool into the open-source CyclOps toolkit, specifically designed to effectively measure and assess the fairness of ML model predictions across patient subgroups.

Methods

This study leverages the General Medicine Inpatient Initiative (GEMINI) data to identify delirium instances during hospitalization. A variety of patient and encounter-level features, such as administrative data, diagnosis codes, lab results, medications, and interventions, are fed to a tree-based classifier model. These features, provide a holistic overview of patient health and facilitate the accurate prediction of delirium likelihood.

Subsequently, we use our tool to evaluate the fairness of the model, focusing particularly on patient demographics. These demographics are key to our fairness assessment due to their known propensity to introduce bias in predictive models. Our main focus lies on essential fairness measures such as disparate impact and equalized odds, allowing us to conduct a comprehensive fairness evaluation of our model.

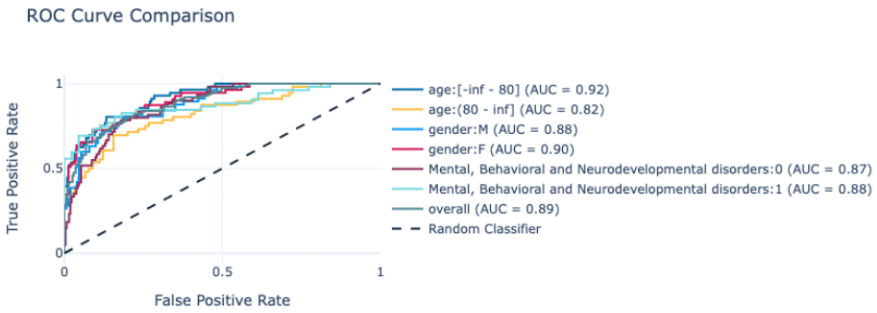
Results

We evaluate the model using the tool on a validation dataset with 420 patients. We show results in figure 1.

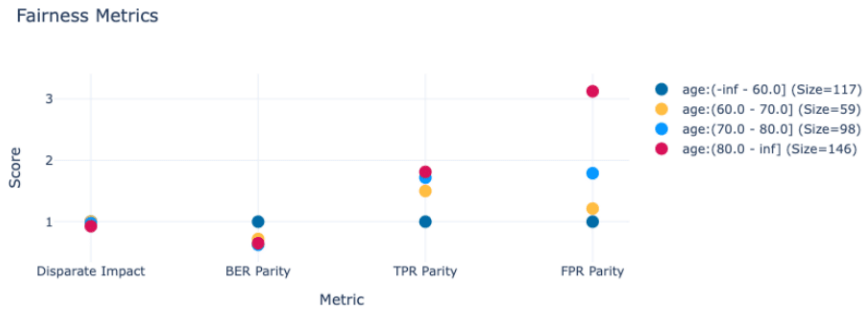
Discussion/Conclusion

The study verifies the effectiveness of our tool in identifying and quantifying bias in ML models. The analysis revealed a disparity in the predictive performance for patients aged above 80, indicating a need for future model improvements. The result findings highlight the relevance of the tool for ensuring equitable healthcare decision-making. More interdisciplinary research is needed for broader adaptability and improved functionality.

Supporting information



(a) We look at Area Under the Curve (AUC) across patient subgroups which include (i) patients aged less than or equal to 80 and above 80, (ii) gender, (iii) patients with and without associated ICD-10 diagnosis codes for Mental, Behavioral and Neurodevelopmental disorders. The classifier had a significant drop in performance for patients aged above 80, while the other subgroups had nominal performance close to the performance over the entire dataset.



(b) We further compute parity metrics for Disparate Impact, Balanced Error Rate (BER), True Positive Rate (TPR) and False Positive Rate (FPR) for patients grouped by age, where the base (favoured) group includes those under and equal to the age of 60. We note that the TPR and FPR ratios for the subgroups are not equal to one, hence they fail the equalized odds tests.

Figure 1: Fairness results

[P34] Artificial intelligence-enabled cough detection for monitoring pulmonary tuberculosis treatment response in Madagascar: a preliminary report

Alexandra Zimmer , McGill University
Mihaja Raberahona, Infectious Diseases Department, University Hospital Joseph Raseta Befelatanana
Etienne Rakotomijoro, Infectious Diseases Department, University Hospital Joseph Raseta Befelatanana
Patrick Andrianiaina Andrianarisoa, Infectious Diseases Department, University Hospital Joseph Raseta Befelatanana
Christophe Elody Andry, Infectious Diseases Department, University Hospital Joseph Raseta Befelatanana
Garcia Rambeloson, Infectious Diseases Department, University Hospital Tambohobe Fianarantsoa
Dera Andriantahiana, Centre d'Infectiologie Charles Mérieux, University of Antananarivo
Mandranto Rasamoelina, Centre d'Infectiologie Charles Mérieux, University of Antananarivo
Lola Jover, Hyfe

Introduction

Cough has traditionally been considered a non-specific indicator of disease. Recent advances in artificial intelligence (AI) are transforming this symptom into an objective biomarker to guide clinical decision-making. For pulmonary TB, this includes examining longitudinal trends in cough patterns to monitor disease evolution and treatment response. Accordingly, the objectives of this study are (1) to assess the feasibility of AI-powered longitudinal cough monitoring in a low-resource setting and (2) to describe the temporal trajectory of cough patterns in PTB patients during the first 14 days of treatment.

Methods

A prospective cohort study was conducted in primary health care centres in Madagascar. Participants aged ≥18 with a new onset of cough and who were suspected to have PTB were recruited and underwent confirmatory PTB testing using GeneXpert MTB/RIF Ultra and culture. All participants (PTB positive and negative) longitudinally monitored their spontaneous coughs for 14 days using a study-provided Android smartphone with the Hye Research application, which uses an AI model detecting and recording coughs. Recording adherence was descriptively assessed and non-parametric Wilcoxon test analysed changes in cough frequency from Day 1 to Day 14.

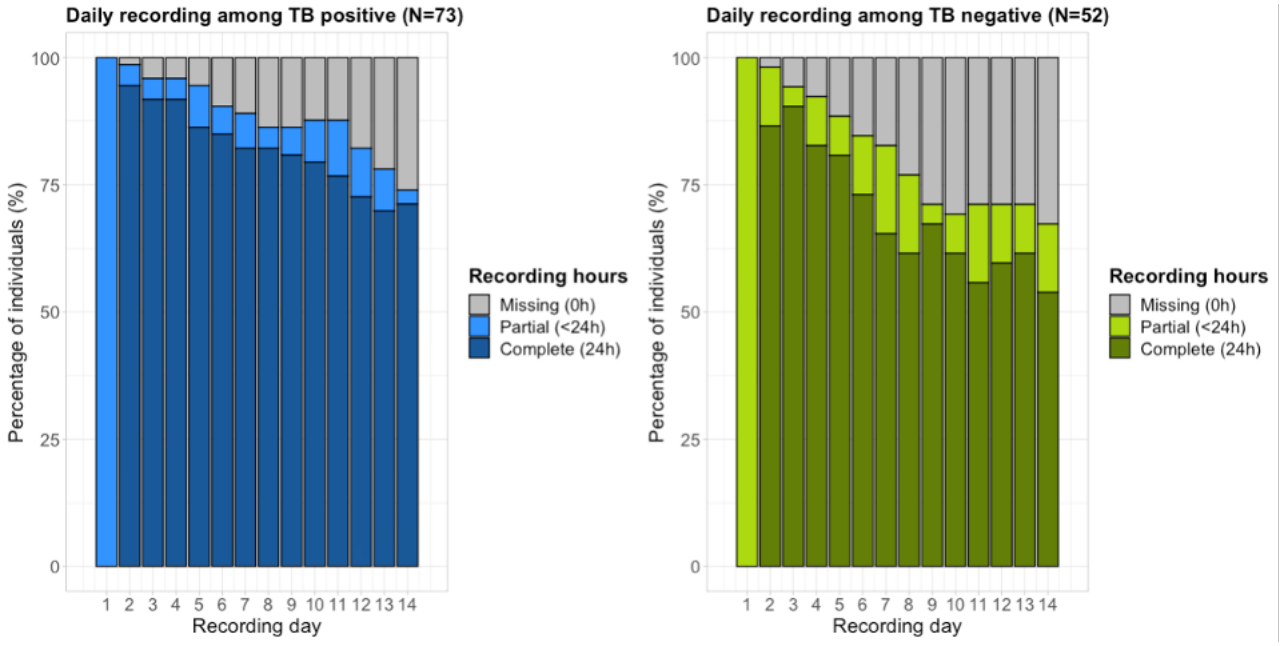
Results

129 individuals were enrolled. The median age was 32 years (IQR: 24-46), 41.1% were female, and 56.6% were PTB positive. Four individuals had indeterminate PTB results. Recording adherence steadily declined over 14 days (Figure). Overall, adherence was higher among PTB positive patients. Day 1 median coughs per hour (medCPH) for PTB positive individuals was 17 (IQR: 3-33), which significantly decreased by Day 14 to 5.75 (IQR 2-13.5) ($P < 0.001$) .

Discussion/Conclusion

Longitudinal AI-based cough monitoring proved feasible in remote Madagascar, with high adherence. Early data suggests that anti-TB treatment reduces cough, indicating that cough monitoring could serve as a non-invasive, low-cost, and person-centric approach to monitor PTB treatment response.

Supporting information



[P35] Assessing Hand Function in Spinal Cord Injury Patients: A Personalized Egocentric Video-Based Hand Analysis Approach

Mehdy Dousty, UofT/BME/Vector

David Fleet, UofT/Vector

José Zariffa, KITE - Toronto Rehabilitation Institute - University Health Network; Institute of Biomedical Engineering, University of Toronto; Rehabilitation Sciences Institute, University of Toronto; Edward S. Rogers Sr. Department of Electrical and Comp

Introduction

The evaluation of hand function after spinal cord injury (SCI) is conducted in clinical settings, which may not accurately reflect hand function in the real world, thereby limiting the efficacy assessment of new treatments. Wearable cameras, also known as egocentric video, are a novel method to evaluate hand function in non-clinical environments. Nonetheless, manual processing of vast quantities of complex video data is difficult, highlighting the need for automated data analysis. The objective of this study was to automatically identify distinct hand postures in egocentric video using unsupervised machine learning.

Methods

Seventeen participants with cervical SCI recorded activities of daily living in a home simulation laboratory. A hand pose estimation algorithm was applied on detected hands to determine 2D joint locations, which were lifted to 3D coordinates. The resulting hand posture information was subjected to a number of clustering techniques. Hand grasps were manually labelled into four categories for evaluation purposes: power, precision, intermediate, and non-prehensile

Results

K-Means clustering consistently exhibited the highest Silhouette score, which reflects the presence of discrete clusters in the data. When comparing with manual annotations, Spectral Clustering applied to a feature space consisting of 2D pose estimation with confidence scores yield the best performance as quantified by maximum match (0.48), Fowlkes-Mallows score (0.46), and normalized mutual information (0.22).

Discussion/Conclusion

In conclusion, this study marks a great milestone in the field of hand function analysis for individuals with SCI using egocentric video. By introducing an innovative unsupervised data-driven hand taxonomy, we offer clinicians a valuable tool to assess hand grasp in an unbiased manner, ensuring quick and effective analysis.

[P36] Bridge2AI-Voice as a Biomarker of Health: Building an ethically sourced, bio-acoustic database to understand diseases like never before (Bridge2AI-Voice Consortium)

Frank Rudzicz, Faculty of Computer Science, Dalhousie University

Jordan Lerner-Ellis, Advanced Molecular Diagnostics, Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Mount Sinai Hospital

Alistair E W Johnson, The Hospital for Sick Children

Lochana Jayachandran, Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Mount Sinai Hospital

Introduction

Background & Aim: The Bridge2AI-Voice as a Biomarker of Health, funded by NIH, aims to integrate voice as a biomarker of health into clinical care by creating a substantial multi-institutional voice database. This resource will fuel voice AI research, predictive modeling, disease screening, diagnosis, and treatment. The study seeks to build an ethically sourced, diverse database of human voices, speech, and respiratory sounds linked to health biomarkers using a secure mobile app and IT infrastructure with federated learning for data privacy.

Methods

Study Approach: Multi-institutional collaborations between voice experts, AI engineers, bioethicists, and social scientists promote ethical data sourcing. The app will integrate voice data with imaging, demographics, and genomics and federated learning will enable collaborative research without sharing raw or private data. Research focuses on voice changes associated with five disease categories (respiratory disorders, voice disorders, neurological disorders, mood disorders, and pediatric voice and speech disorders) in patients at USF specialty clinics and participating institutions. Collaborations with High Volume Expert Clinics and Community Clinics ensure database diversity. The four-year project spans 11 academic sites in the US and Canada, with potential expansions. De-identified voice and clinical data will be accessible on a cloud-based infrastructure for future Voice AI research.

Results

N/A

Discussion/Conclusion

Conclusion: Bridge2AI-Voice as a Biomarker of Health is a ground-breaking endeavour leveraging voice as a biomarker for disease understanding on an unprecedented scale. Federated learning strengthens data privacy and enables valuable insights from voice changes associated with known diseases across five categories. This project sets a new trajectory in voice integration for healthcare, revolutionizing disease diagnosis, patient management, and enabling personalized data-driven healthcare.

[P37] Capacity of Language Learning Models to Generate Medical Residency and Undergraduate Medicine Progress Test Questions

Ryan S. Huang , Temerty Faculty of Medicine, University of Toronto

Hanu Chaudhari, University of Toronto Department of Family and Community Medicine

Alexandra Athanaselos, Department of Family and Community Medicine, University of Toronto

Katina Tzanetos, Department of Family and Community Medicine, University of Toronto

Fok-Han Leung, Department of Family and Community Medicine, University of Toronto

Introduction

Advancements in artificial intelligence, particularly in natural language processing, have led to the development of models such as Chat Generative Pre-Trained Transformer (ChatGPT) by OpenAI, which is a language learning model (LLM) with the ability to generate human-like responses in an ongoing dialogue. Despite concerns about its implications in education, there are potential benefits including faster access to feedback and information. Given its successful application in taking examinations in various areas of medical education, we explored its utility in generating quality multiple-choice questions (MCQs) for the postgraduate family medicine residency and undergraduate medicine progress tests at the University of Toronto.

Methods

The study consists of a multi-phase analysis of ChatGPT's ability to generate examination questions. A series of categorical variables will be considered: Subject matter expert (SME) generated questions/scenarios, GPT4 generated questions using cloning from existing question banks, GPT4 generated questions using novel prompts/inputs, and GPT4 generated questions with human appraisal. Data collection will be managed using REDCap, with scoring of questions done by experienced physicians.

Results

To be determined after the completion of the study phases. The study will compare mean scores relative to each categorical variable using linear mixed models. Non-inferiority will be assessed by investigating conditional expectation of quality scores in each intervention arm (SME, GPT4, GPT4 with human appraisal) and associated 95% confidence intervals.

Discussion/Conclusion

The utilization of AI models like ChatGPT may offer a more cost-effective and time-efficient solution to the laborious process of creating high-quality MCQs. Our work will directly generate usable study tools to support the growth of exam banks for both assessment and student practice as well as create a new framework for medical students and teachers. As the field of AI continues to evolve rapidly, this study could provide key insights into the possible intersections of AI and medical education.

[P38] Clinician and health system leader perspectives on the use of AI for deriving social determinants of health data in primary care settings

Stephanie Garies , Unity Health Toronto

Simon Liang, Unity Health Toronto

Steve Durant, Unity Health Toronto

Karen Weyman, Unity Health Toronto (St. Michael's Hospital)

Noor Ramji, Unity Health Toronto (St. Michael's Hospital); University of Toronto

Mo Alhaj, Unity Health Toronto

Andrew Pinto, Unity Health Toronto; University of Toronto

Introduction

Artificial intelligence (AI) is increasingly used in healthcare settings, but it needs to be thoughtfully developed and co-designed with end-users to solve real-world challenges, build trust, and ensure health inequities are not exacerbated. The aim of this study is to understand the perspectives of clinicians and decision-makers on the use of AI to derive and present patient social data.

Methods

Semi-structured interviews were conducted between June 2022-January 2023 with clinicians (e.g. family physicians, allied health professionals) and decision-makers (e.g. practice managers, government). A thematic content analysis was conducted to understand how participants currently access social determinants of health (SDoH) for patients, perspectives on using AI to derive SDoH from electronic medical records (EMR), ethical considerations, and how best to present and use AI-derived social data.

Results

Twelve interviews were conducted in total. Just over half of participants (n=7) said they were minimally knowledgeable about AI, with the remaining moderately or very knowledgeable. Most participants described limited or inconsistent access to patient SDoH outside of sex and gender. A number of potential benefits of using AI to derive social data were reported, such as reducing clinician time and effort to capture SDoH, improving patient care, and avoiding uncomfortable conversations about social needs. Participants also had many concerns related to biases in the data/AI, protecting patient confidentiality, further stigmatizing patients, and the potential to lead to inappropriate care or patient harm. A wide range of suggestions for future use of AI-derived social data were supplied, both for direct patient care and at a practice- or population-level.

Discussion/Conclusion

Patient social data are necessary for the provision of high-quality, equitable health care. These findings will inform future work to co-design a useful, actionable AI-based tool for use in primary care settings to better identify social needs of patients.

[P39] Delineate cell-cell communication (CCC) in anti-cancer drug resistance by deep learning based multi-modal single-cell methods

Fatema Zohora , Princess Margaret Cancer Center

Gregory Schwartz , Princess Margaret Cancer Center

Introduction

Drug resistance is responsible for up to 90% of cancer related deaths. Recent studies show CCC as a potential non-genetic factor that greatly influences anticancer drug resistance of the cancer cells. Single-cell data can better explain CCC networks and such data come from multiple molecular layers, e.g., transcriptomic, epigenomic, and spatial molecular layers. Existing computational methods for predicting CCC in tumor samples often involve experiments that collect a single molecular layer from individual cells and do not take the spatial location of the cells into account. Moreover, most tools predict communication between cell populations instead of detecting CCC at single-cell resolution. As a direct result of these limitations, existing tools have low true positive rates.

Methods

I am developing deep learning based methods to pinpoint CCC at single-cell resolution by integrating multiple molecular layers of data along with cell location. Technically, I am using Graph Neural Network based models and performing unsupervised training of the model through Deep Graph Infomax - a contrastive learning approach.

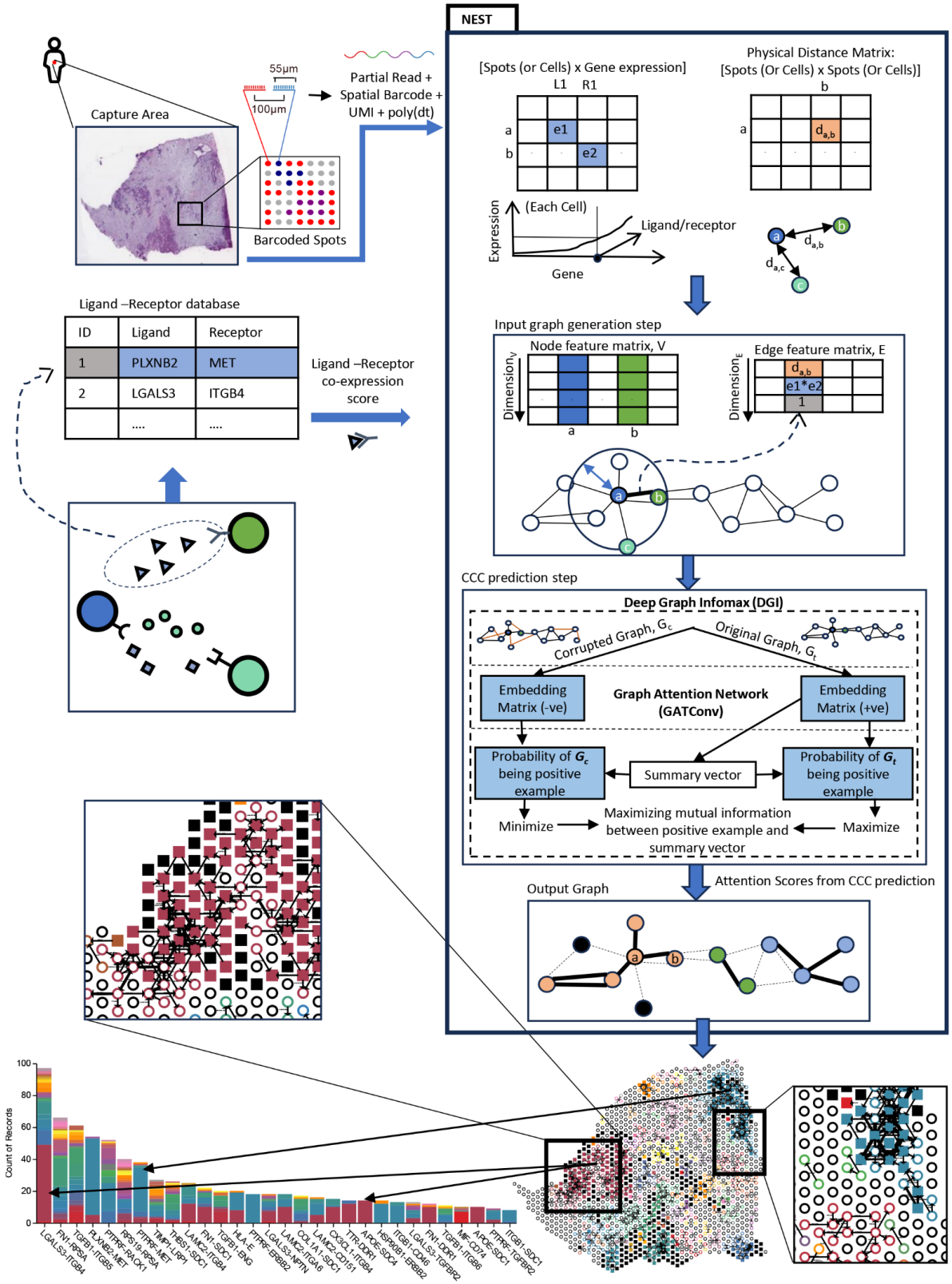
Results

We validated the ligand-receptor based CCC detected by our model through well-studied published dataset, e.g., Visium data of human lymph node, and MERFISH data of mouse hypothalamus preoptic region. Besides that, we validated the model through synthetic data. Then we show that our model can find potential CCC from Lung adenocarcinoma data (published) that supports original findings and also provides additional biomarker CCC. Finally, we apply our model on in-house pancreatic ductal adenocarcinoma (PDAC) samples to answer new biological questions.

Discussion/Conclusion

My model should guide in identifying CCC-related therapeutic targets to help improve patient outcomes. This model is transferable to any other related fields where a graph representation exists. Therefore, interested researchers from cancer, single-cell, and other related research communities can use the model as well as change/adapt it to serve other projects ideas.

Supporting information



[P40] Exploring Canadian Sentiments on COVID-19 Vaccination: A Twitter-based Analysis

Hassan Maleki Golandouz, University of Manitoba

Wendy Xie, National Collaborating Center for Infectious Diseases, Winnipeg, Canada

Lisa M. Lix, University of Manitoba

Introduction

Vaccination rollout was vital for protecting Canadians against severe outcomes from COVID-19 infection. Social media platforms, such as Twitter, can be valuable sources of data for investigating public opinions about COVID-19 vaccination. Our objective was to identify key topics from COVID-19 vaccine-related English tweets posted by Canadian users and examine trends in public opinions and sentiments.

Methods

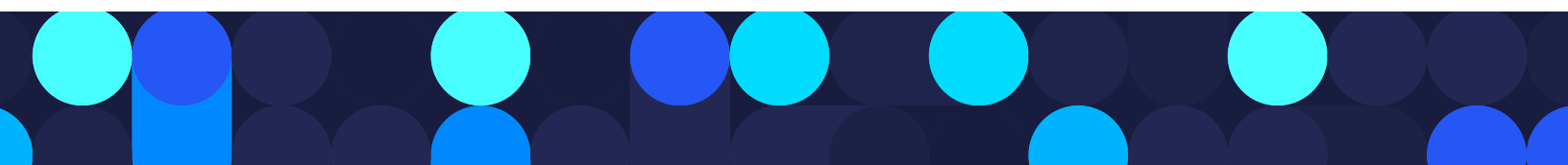
Our study analyzed 8,909 COVID-19 vaccine-related tweets (excluding retweets) from 4,587 users, posted between November and December 2021. During this period, Health Canada's child vaccination approval and the federal vaccination mandate for employees significantly impacted public sentiment. We focused on Canada-relevant tweets, as ascertained using Twitter user location. Using Correlation Explanation (CorEx) topic modeling, we analyzed the tweets, setting variable topics per iteration, to determine the optimal numbers of topics based on correlation scores. The model extracted keywords for topic inference and a random tweet set was assessed per topic for consistency. We also employed VADER, a sentiment analysis tool particularly sensitive to social media sentiments, to calculate sentiment scores and examine trends, considering both the polarity and intensity of expressed emotions in tweets.

Results

Our analysis revealed 17 topics in COVID-19 vaccine-related tweets. The most common topics were "scheduling vaccination appointments"(21.0%), "variants and immunity" (10.5%), and "vaccination legal actions, controversies"(9.1%). The sentiment analysis revealed that topics such as scheduling vaccination appointments, vaccine perspectives and decisions, vaccine dosing, pediatric vaccination, vaccine policies and public response, and variants and immunity witnessed more days of positive sentiment than negative sentiment. However, topics like vaccine adverse effects, vaccination legal actions, controversies, and vaccine impact on healthcare workers attracted more days with negative sentiment.

Discussion/Conclusion

Analyzing tweet trends during specific time periods like major policy changes can aid healthcare policymakers in strategically developing adaptable programs, leveraging diversified viewpoints and public sentiment insights for enhanced efficiency in COVID-19 vaccination.



[P41] Exploring patient perspectives on how they can and should be engaged in the development of artificial intelligence (AI) applications in health care

Samira Adus, University of Toronto

Andrew Pinto, Unity Health Toronto; University of Toronto

Jillian Macklin, University of Toronto

Introduction

Artificial intelligence (AI) is a rapidly evolving field which will have implications on both individual patient care and the health care system. There are many benefits to the integration of AI into health care, such as predicting acute conditions and enhancing diagnostic capabilities. Despite these benefits potential harms include algorithmic bias, inadequate consent processes, and implications on the patient-provider relationship. One tool to address patients' needs and prevent the negative implications of AI is through patient engagement. As it currently stands, patients have infrequently been involved in AI application development for patient care delivery.

Methods

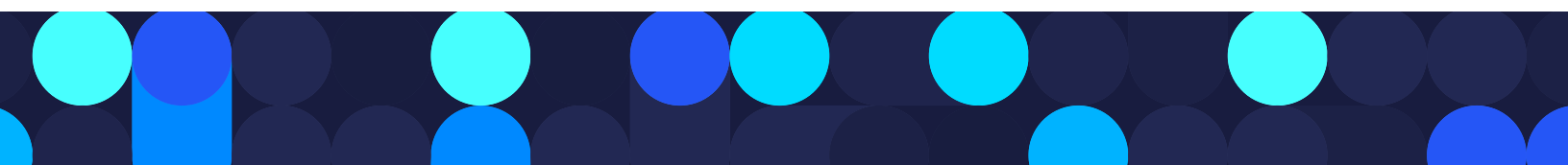
We conducted four virtual focus groups with thirty patient participants to understand of how patients can and should be meaningfully engaged within the field of AI development in health care. Participants completed an educational module on the fundamentals of AI prior to participating in this study. Focus groups were analyzed using qualitative content analysis.

Results

We found that patients want to be engaged at the problem-identification stages using multiple methods such as surveys and interviews. Participants preferred that recruitment methodologies for patient engagement included both in-person and social media-based approaches with an emphasis on varying language modalities of recruitment to reflect diverse demographics. Patients prioritized the inclusion of underrepresented participant populations, longitudinal relationship building, accessibility, and interdisciplinary involvement of other stakeholders in AI development. We found that AI education is a critical step to enable meaningful engagement.

Discussion/Conclusion

Given the novelty and speed at which AI innovation is progressing in health care, patient engagement should be the gold standard for application development. Our proposed recommendations seek to enable patient-centered AI application development in health care. Future research must be conducted to evaluate the effectiveness of patient engagement in AI application development to ensure that both AI application development and patient engagement are done rigorously, efficiently, and meaningfully.



[P42] Identifying Individualized Neurophysiological Causal Features for Working Memory Performance: Implications for Non-Invasive Brain Stimulation

Mina Mirjalili, Centre for Addiction and Mental Health

Reza Zomorrodi, Centre for Addiction and Mental Health

Zafiris J. Daskalakis, Department of Psychiatry, School of Medicine, University of California, San Diego

Daniel M. Blumberger, Centre for Addiction and Mental Health

Sean L. Hill, University of Toronto, Vector Institute for Artificial Intelligence

Tarek K. Rajji, Centre for Addiction and Mental Health

Introduction

Non-invasive brain stimulation (NIBS) has been extensively used to target neural oscillations for improving working memory performance. However, current NIBS paradigms have two main limitations: first, lack of personalization, and second, a need for a randomized controlled trial (RCT) to determine causality between the targets of NIBS and performance. RCTs are expensive, time-consuming, and sometimes simply not possible. Computational personalized causal modeling is a tool to discover the optimum targets for NIBS to improve working memory performance. Therefore, our aim was to use this method and observational data to discover causal relations between neural oscillations and working memory performance, and to identify potential targets and stimulation parameters for personalized NIBS.

Methods

We used electroencephalography (EEG) data from 66 young healthy participants collected while performing a 3-back working memory task. Using graphical causal modeling, we discovered causal neural oscillations of working memory performance and compared the causal features between two groups: high and low performers.

Results

Total number of causal features in high performers was higher than low performers. Among the causal features, right temporal gamma oscillation was ~5 times (z-score = 3.87, p = 0.0001) more frequently a causal feature among high performers than low performers. However, the power of causal temporal gamma oscillation was not different between the two groups.

Discussion/Conclusion

Our findings suggest that a potential approach to improve working memory performance using NIBS is to induce more causal gamma oscillations by, for example, generating more local gamma entrainment over the right temporal cortex and not necessarily by increasing gamma power.

[P43] Identifying, Characterizing and Tracking Suspicious Skin Moles

Mahla Abdolahnejad , Skinopathy Inc.

Rim Mhedbi , Skinopathy Inc.

Hannah Chan , Skinopathy Inc.

Rakesh joshi , Skinopathy Inc.

Joshua N. Wong , University of Alberta Hospital

Collin Hong , Scarborough Health Network and Skinopathy Inc.

Introduction

Most adults have 10-40 moles on their body and one can potentially be melanoma, the cause of 70% of skin cancer deaths globally. The definitive assessment for melanoma is biopsy, but it is invasive and not practical to test all moles. A proof-of-concept solution uses computer vision to map all moles on a colour 2D image, segments them, and uses a digitized clinical method, the ABCD (asymmetry, border, colour, diameter) protocol, to flag all suspicious moles.

Methods

We created an algorithmic pipeline whose first component maps all individual moles in wide-angle 2D-colour images. Tested images had a wide diversity in skin tone, mole numbers, and mole size. Individual mole images are then pre-processed and loaded into a novel ABCD algorithm. The algorithm leverages computer vision and neural network components to ascertain ABCD scores for each mole. Moles that meet the “suspicious” criteria are then automatically flagged.

Results

The mole mapping algorithm detects most (>90%) moles in a wide-angled 2D colour image. The individual moles are segmented by a novel saliency Boundary-Attention Mapper (BAM), built from a neural network that achieved 87% accuracy in classifying skin lesions. Benchmarking BAM on the ISIC2017 dataset gave a 90.45% pixel-wise accuracy, 86.06% pixel-wise sensitivity, and 94.35% pixel-wise specificity. Asymmetry, colour, and diameter (in pixels) scores for suspicious moles are within acceptable error rates, however border irregularity scores have low confidence.

Discussion/Conclusion

The digitized mole mapping and ABCD system has clinical relevance. However, the major challenges that still exist are the need for high resolution individual images for accurate border irregularity scores, and the feasibility of using fiducial markers for lesion measurements to facilitate at-home monitoring.

[P44] Investigating model failures by patient (profiles) for safer clinical deployment

Olivier Lefebvre , Université de Sherbrooke
Martin Vallières, Université de Sherbrooke
Jean-François Ethier, Université de Sherbrooke
Félix Camirand Lemyre, Université de Sherbrooke

Introduction

Machine learning models are powerful tools for clinical applications, like mortality prediction. However, they may exhibit different levels of performance, including some very poor one, for different patient groups with different profiles of attributes, despite strong global performance in external test sets. Secondly, ensuring model reliability and consistency over time is of utmost importance. Hence, our study seeks to develop a method to assess predictive uncertainty of classification models for individual patients and profiles at implementation and over time.

Methods

We first develop an error variable indicating model mispredictions, then train a second-layer model called conditional accuracy (CA) to characterize mispredictions for individual patients and groups. This enables us to identify patient profiles with higher prediction errors (lower CA) from the first model, as well as profiles undergoing behavior change.

Results

Our approach helps determine if uncertainties compromise model usability for patients. Using a Random Forest model trained on 2011-2016 admissions (n=122,860) for 1-year mortality prediction from hospitalization data, we observed significant changes in model uncertainty for different patient profiles between 2017-2018 (n=22,034) and 2020-2021 (n=27,664) cohorts. In particular, younger patients (< 57yo) not admitted to medicine/surgery and with peripheral oedema had an accuracy drop from 86% to 38%, conditional accuracy from 0.69 to 0.39, and mortality rates surged from 10.8% to 61.8%. This could represent higher mortality than previously for younger hospitalized COVID patients during that time period.

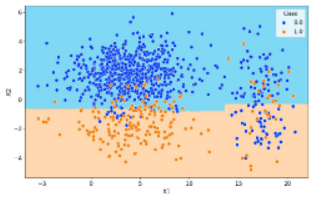
Discussion/Conclusion

This study emphasizes exploring model failures and uncertainties in clinical deployments. The two-layer approach identifies patients with reduced model reliability and evaluates variability across patient profiles over time, considering profile behavior changes. Automatically identifying these uncertainties empowers responsible utilization of machine learning models in healthcare applications, providing improved support to users. Continued research in this direction is crucial for ensuring the safe and effective integration of machine learning models in clinical practice.

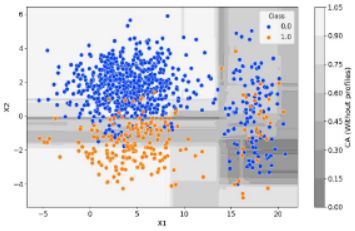
Supporting information

Rationale: Global metrics don't tell the whole story

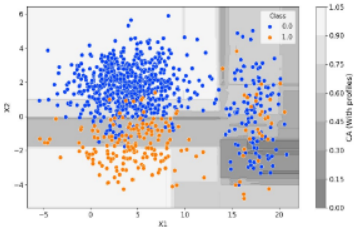
Method (simulated Data)



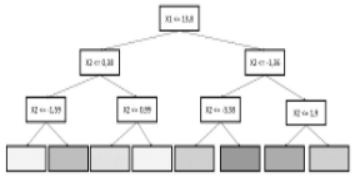
Original classification model



Conditional accuracy (CA) for individual patients



Conditional accuracy (CA) with profiles



Assess mispredictions and identify patient profiles with higher prediction errors

Results (example of use on One Year Mortality model)

Global metrics of the One Year Mortality model

OYM model (2011-2016, n=122.860)	2017-2018 Cohort	2020-2021 Cohort
n	22.034	27.664
AUC	0.873	0.87
Accuracy	87%	86%
Mortality rate	15.1%	17.5%

Profiles for One Year Mortality model

Automatic identification of profiles with change in behavior

2017-2018 Cohort	
CA	0.69
AUC	0.771
Accuracy	86%
Mortality rate	10.8%

2020-2021 Cohort	
CA	0.39
AUC	0.784
Accuracy	38%
Mortality rate	61.8%

Identification of patients with reduced model reliability and evaluation of variability across patient profiles over time

[P45] Investigating the relationships between social media discourse and ICU bed demand to inform healthcare supply-chain decisions: COVID-19, Twitter and Causal Analysis

Mahakprit Kaur, York University
Jude Kong, York University
Taylor Cargill, York University
Kevin Hui, York University
Minh Vu, York University
Nicola Bragazzi, York University

Introduction

The COVID-19 pandemic has highlighted gaps in current handling of medical resource demand surges and the need for prioritizing scarce medical resources. In this study, Twitter sentiment analysis was utilized to determine if an increase in negative sentiment COVID-19 related tweets in Brazil, India, and the US correspond to a greater use of hospital Intensive Care Unit (ICU) beds in these countries.

Methods

Tweets were collected from a publicly available dataset that contained COVID-19 tweets with sentiment labels and geolocation between February 1, 2020 and March 31, 2021. Negative sentiment tweets were analyzed using the Granger Causality test and Convergent Cross Mapping (CCM) to determine if the time series for negative sentiment tweets can be useful for forecasting ICU bed shortages in the United States, Brazil, and India.

Results

For the United States, the Granger test was significant for 14 of the 50 regions which passed the Augmented-Dickey Fuller test at lag 2 ($p < 0.05$) and, according to the CCM analysis, there was a significant relationship between ICU bed demand and negative COVID-19 tweet sentiment for 46 of 50 states ($p< 0.05$). For Brazil, the Granger test was significant for 6 of the 27 subregions ($p < 0.05$) and there was a significant relationship for 26 of 27 Brazilian subregions ($p< 0.05$) according to the CCM analysis. For India, the results of the Granger test were significant for 6 of the 26 states ($p< 0.05$) and all 26 states exhibited a significant relationship from CCM analysis.

Discussion/Conclusion

This study provides a novel approach for identifying regions of high hospital bed demand by analyzing Twitter sentiment data supporting Twitter as a useful tool for organizing relief efforts in a timely manner.

[P46] Latent Variable Energy Based Model with Self-Supervised Approaches for Cancer Grading Problem

Kayvan Tirdad, Toronto Metropolitan University
Alex Dela Cruz, Toronto Metropolitan University
Kenneth Wenger, Toronto Metropolitan University
Alireza Sadeghian, Toronto Metropolitan University

Introduction

Labeling medical image datasets are time-consuming and prohibitively expensive, requiring hundreds of hours of effort from expert diagnosticians. Recent advances in semi-supervised learning algorithms (SSL) have made great strides in reducing the training dependency on labeled datasets and requiring that only a subset of the data be labeled. In addition, using Energy Based Model (EBM) in medical domain problems, such as image reconstruction, Image segmentation, and Image to Image translation, has recently led to much stronger systems. Adding a Latent variable to EBM has been presented to better capture the hidden information that such systems deal with. This work presents a platform for using Latent variable Energy Based Model (LVEBM) with a self-supervised learning approach to solve the cancer grading problem.

Methods

We applied a joint embedding predictive architecture with the implementation of a variation of variance-invariance-covariance regularization (VICReg) for the self-supervised approach to the problem of cancer grading. The model was trained to grade Whole Slide Image (WSI) of bladder cancer. The system tried to match the vectorized version of both inputs (original and distorted version) while learning whether it is cancerous or not.

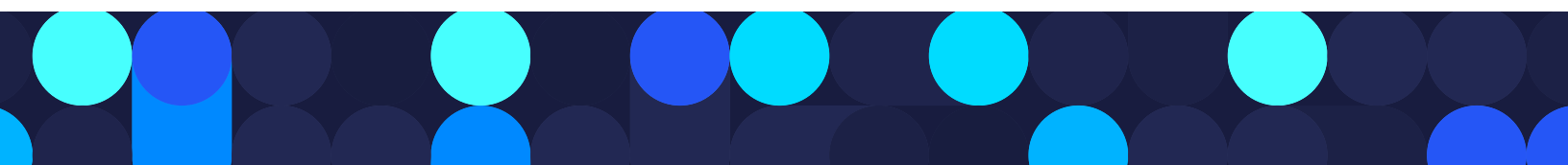
The research involved researchers from the Computer Science Department of Toronto Metropolitan University.

Results

The performance of the final system was on par with the state-of-the-art approaches that applied to the same problem. At the same time, we show that an increase in data produced better performance faster than any other approaches. In addition, the final system shows the importance of the criteria applied to the system as a latent variable to the performance of the systems.

Discussion/Conclusion

This project shows that Latent variable Energy Based Model (LVEBM) with Self-supervised approaches can be considered a very effective method for solving cancer grading problems while it can perform with very little labeled data.



[P47] Off-Label Drug Use during the COVID-19 Pandemic in Africa: Topic modeling and sentiment analysis of Ivermectin in South Africa and Nigeria as a case study

Zahra Movahedi Nia, York University
Nicola Bragazzi, York University
Ali Asgary, York University
Bruce Mellado, Witwatersrand University
James Orbinski, York University
Jianhong Wu, York University
Jude Kong, York University

Introduction

Although rejected by the World Health Organization (WHO), the human and even veterinary formulation of Ivermectin has widely been used for prevention and treatment of COVID-19, around the globe. In this work we leverage Twitter to understand the reasons for the drug use from Ivermectin-supporters, their source of information, their emotions, and their gender demographics, in Nigeria and South Africa.

Methods

Topic modeling is performed on a Twitter dataset gathered using keywords “ivermectin” and “ivm”. A model is fine-tuned on RoBERTa to find the stance of the tweets in two countries, namely, Nigeria and South Africa. Statistical analysis is performed to compare the stance and emotions of the tweets.

Results

Most Ivermectin supporters either redistribute conspiracy theories posted by influencers, or refer to flawed studies confirming Ivermectin efficacy in vitro. Three emotions have the highest intensity, optimism, joy, and disgust. The stance distribution on topics is very different for the two countries (p=0.0086). South Africa and Nigeria have a higher disgust and optimism/joy intensities, respectively. The number of anti-Ivermectin tweets has a significant positive correlation with vaccination rate.

Discussion/Conclusion

This work makes the effort to understand public discussions regarding Ivermectin during the COVID-19 pandemic. This work helps policy makers inform more targeted policies to discourage self-administration of Ivermectin. Moreover, it is a lesson to future outbreaks.

Supporting information

Table 1: Correlation between number of cases and total number of tweets on Ivermectin, and correlation between different stances and number of administered vaccines for South Africa and Nigeria

	Nigeria			**South Africa**	
	Correlation with Number of cases	**p-value**		**Correlation with Number of cases**	**p-value**
Total tweets	0.6094	0.00094	Total tweets	0.62784	0.00011
	Correlation with vaccination rate	**p-value**		**Correlation with vaccination rate**	**p-value**
Anti-Ivermectin	0.4760	0.01396	Anti-Ivermectin	0.66016	3.8710^{-5}
Pro-Ivermectin	-0.2912	0.14882	Pro-Ivermectin	-0.34718	0.05154
Neutral	0.08968	0.66304	Neutral	-0.20192	0.26774

[P48] On the factuality of Large language model-generated summaries of clinical abstracts

Wael Abdelkader, McMaster University
Cynthia Lokker, McMaster University
Alfonso Iorio, McMaster University
Sophia Ananiadou, National Centre for text mining, University of Manchester
Zheheng Luo, National Centre for text mining, University of Manchester
Qianqian Xie, National Centre for text mining, University of Manchester

Introduction

Large Language Models (LLMs) have emerged as powerful tools for handling complex text-processing tasks. Utilizing LLMs to summarize lengthy scientific documents may save time for researchers and non-experts, thereby expediting the dissemination of scientific information. However, factual inconsistency is a critical problem in generated summaries that may include information that is absent or not logically deduced from source documents.

Methods

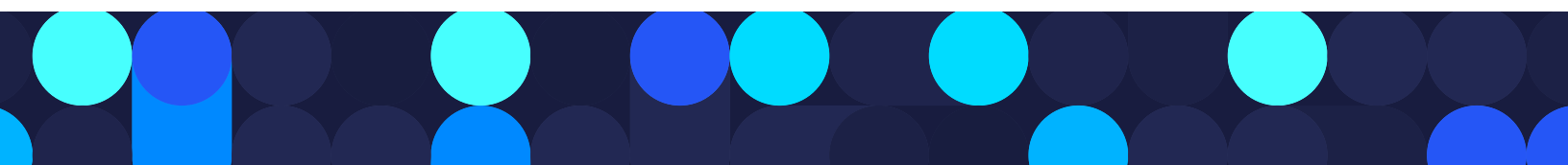
This study assessed the degree of factual consistency in summaries generated by two widely used LLMs, ChatGPT and Vicuna-13B, from titles and abstracts of 85 clinical treatment articles, resulting in 170 abstract-summary pairs. Factual consistency was evaluated using a detailed protocol by four experts in Evidence-Based Medicine. The experts assessed each summary for fine-grained aspects, including PICO (Population, Intervention, Comparator, Outcome), direction of conclusion, and strength of claim, and assigned an overall factuality score ranging from 0 to 3. To assess inter-assessor agreement, 78 summaries were evaluated by two experts.

Results

Our analysis revealed that 50.6% of ChatGPT and 37.6% of Vicuna-13B generated summaries achieved full factual scores. A closer examination of the consistency ratio showed the LLMs' struggle in accurately capturing the Comparator, Intervention, and Direction of Conclusion, all of which scored below 40% in consistency. Further inspection of the Inter-Assessor Agreement demonstrated that agreements were mostly below 90% in most aspects, except for Intervention. Particularly, with its relatively larger range of choice, annotators only agreed on 41% of cases for the overall score.

Discussion/Conclusion

This pilot study represents an initial effort to evaluate summaries generated by LLMs. The relatively small percentage of factually consistent summaries highlights the potential risks associated with the direct application of LLMs in processing clinical text. Moreover, the agreement among annotators underscores the challenges involved in evaluating the factuality of clinical texts, emphasizing the need for further refinement of the annotation protocol.



Supporting information

AI in medicine: Abstract supplementary information

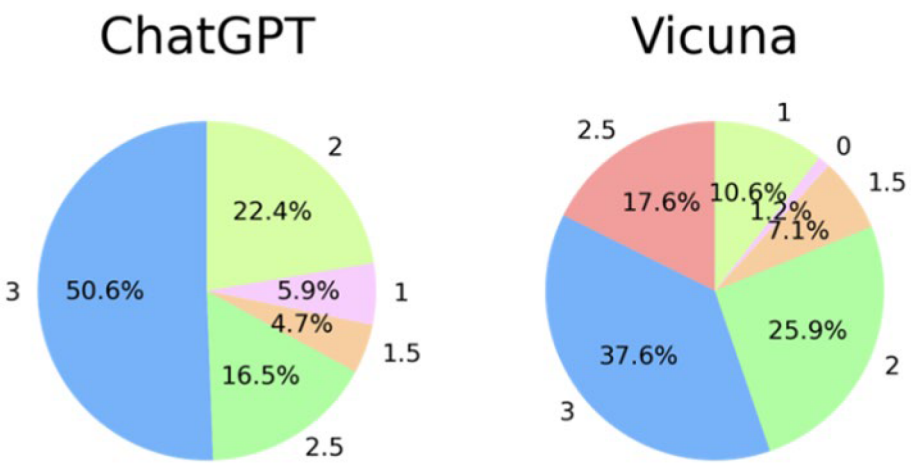


Figure 1: Pie chart for the models' overall scores percentage

[P49] Predicting Pre-eclampsia in Pregnant Women: An ML-based Approach using the Lavndr App

Shveta Bhasker, Institute of Health Policy, Management and Evaluation, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada
Kadriye Candas, Institute of Health Policy, Management and Evaluation, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada
Ashley Girgis, Institute of Health Policy, Management and Evaluation, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada
Natasha Rozario, Institute of Health Policy, Management and Evaluation, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada
Praveena Santhakumaran, Institute of Health Policy, Management and Evaluation, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada

Introduction

Pre-eclampsia (PE) is a high-risk pregnancy complication with at least 60% of pre-eclampsia deaths being preventable. One major challenge in managing pre-eclampsia is the lack of a reliable method for predicting its occurrence. In addition, pre-eclampsia disproportionately affects vulnerable populations including Black and Indigenous women who are at greater risk of developing PE.

Methods

We propose an idea for a mobile health (mHealth) application (app) named Lavndr that leverages a machine learning (ML) algorithm to provide early identification of high-risk individuals and promote proactive maternal health management. A nationwide health insurance dataset of the BPJS Kesehatan in Indonesia was examined. A cost-benefit analysis will be performed to inform health policymaking since the mHealth application will reduce health expenditure.

Results

This mHealth app can assist in transitioning to a more proactive approach to maternal health. The ML algorithm to predict PE is projected to result in a model with accuracy above 90%. Similarly, based on publicly available information from the Canadian Institute for Health Information, a cost-benefit analysis is projected to show >$1 million CAD in cost savings over one year since the mHealth app can prevent unnecessary surgeries related to PE like cesarean sections.

Discussion/Conclusion

There is a lack of high-quality maternal health data available for ML algorithm training. This is especially concerning as pre-eclampsia disproportionately affects Black women and Indigenous women. By leveraging the rich demographic diversity of Canada, we can develop a more comprehensive and representative dataset that would enable the development of more accurate and effective ML algorithms for predicting pre-eclampsia. This would also result in millions of dollars in cost savings for the Canadian health system.

Supporting information

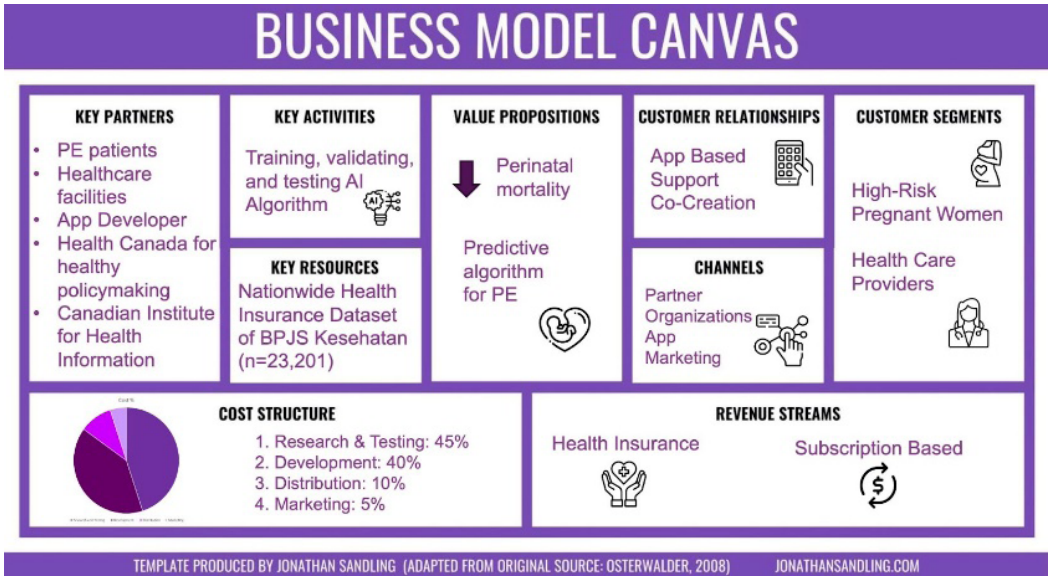


Figure 1. This is the business model canvas for the mHealth app. By partnering with the stakeholders listed, we can achieve our goals of reducing perinatal deaths and developing an accurate predictive algorithm.

[P51] Solving Healthcare's Last Mile Problem

Karim Keshavjee, IHPME/University of Toronto
Anson Li, EY
Mark Dayomi, IHPME/University of Toronto
Pooyeh Graili, Quality HTA
Ali Balouchi, InfoClin
Aziz Guergachi, Ted Rogers School of Management/Toronto Metropolitan University

Introduction

AI could revolutionize healthcare, but not if we implement it in traditional ways--through physicians. Physicians are overwhelmed. They need to work 24 hours per day just to deliver existing evidence-based interventions to their patients, let alone deliver new advancements in care. AI will only increase the number of hours physicians need to work if we plan to deliver it through them. We propose an app marketplace that can deliver AI directly to patients and individuals at risk by making evidence-based care accessible outside physician visits.

Methods

We identified key interestholders who would benefit from or be supportive of such a platform, developed draft requirements for each interestholder and drew a draft reference architecture based on the requirements. We asked interestholders (N&#3f10) to provide feedback on requirements, the architecture diagram, the design and the implementation approach. We elicit attractive features, features they are skeptical of, areas for improvement, assessment of feasibility and ways to improve feasibility.

Results

Patients find the concept appealing, but are skeptical that AI can deliver promised benefits. App publishers like it because it lowers customer acquisition costs. Policy players are concerned about feasibility because of the coordination required by multiple players. Physicians are willing to participate as long as it doesn't increase their work effort and wonder how the system will compensate them for use of their EMRs and data. Health charities see the value for their constituents. Digital health deployment partners see the value and are interested exploring it further.

Discussion/Conclusion

Increasing access to evidence-based prevention and treatment is essential to reducing pressure on our healthcare system. If implemented correctly, this reference architecture could deliver AI directly to patients and significantly decrease population risk and improve outcomes. One challenge could be the older population's unfamiliarity or distrust of technology. Further research is needed to address this potential problem.

[P52] Stability-Based Biomarker Development to Identify Pathological Brain Areas Responsible for Freezing of Gait in Parkinson's Disease

Nooshin Bahador, UHN

Robert Chen, University of Toronto

Milad Lankarany , University of Toronto

Introduction

Freezing of gait (FOG) as one challenging symptom in Parkinson's disease (PD) is linked to cognitive impairment. Various treatments for FOG have been suggested, including medication, surgery, neuropsychiatric methods, and non-invasive brain stimulation. However, none have shown consistent effectiveness. The goal of this research was to address the lack of reliable biomarkers for identifying pathological brain areas responsible for FOG.

Methods

Stability characteristics of neural activities were used as a biomarker for identifying the target area. To develop this biomarker, a linear model was created to represent EEG activity dynamics, and the relationship between different electrodes was extracted. By testing the model's sensitivity to individual electrode perturbations, vulnerable electrodes were identified. Our study had been performed on data collected from nineteen patients diagnosed with PD—14 men and 5 women with an average age of 66.8±7.1 (mean ± SD). The study had been approved by the UHN Research Ethics Board, and written informed consent had been obtained from all participants. Each participant had performed an experimental task in a darkened room, viewing a virtual hallway on a monitor. They depressed right and left foot pedals alternately to move forward through narrow and wide doorways along the hallway.

Results

We found that the left supplementary motor area was associated with FOG. Additionally, we observed that the brain underwent an intermediate phase prior to the onset of FOG, during which instability propagated medially. We have not observed a significant difference in alpha-band power between frozen gait and voluntary stop in this brain region. However, beta-band power markedly decreased during frozen gait compared to voluntary stop in the same area.

Discussion/Conclusion

Our algorithm helped to identify what specific brain regions are engaged. In future works, non-invasive neuromodulation technique can be used to determine whether stimulating the area improves symptom.

[P53] SurvdigitizeR: R Package to Automate the Digitization of Published Kaplan-Meier Curves

Jasper Zhongyuan Zhang, The Hospital for Sick Children
Juan David Rios, The Hospital for Sick Children
Tilemanchos Pechlivanoglou, York University
Alan Yang, The Hospital for Sick Children
Qiyue Zhang, The Hospital for Sick Children
Dimitri Deris, McMaster University
Ian Cromwell, CADTH
Petros Pechlivanoglou, The Hospital for Sick Children

Introduction

Economic evaluations and meta-analyses frequently depend on survival probabilities digitized from published Kaplan-Meier (KM) curves. Manual digitization of KM curves is time-consuming, expensive, and can lead to errors. This study aims to develop an efficient and accurate algorithm for automating the extraction of survival probabilities from KM curves and providing a user-friendly open-source tool.

Methods

We developed an automated digitization algorithm which processes images, converts them in their hue, saturation, and lightness scale and uses optical character recognition to detect axis location and labels. It also uses a k-medoids clustering algorithm to separate multiple overlapping curves on the same figure. To validate performance, we created 36 survival plots from an exponential distribution. We added random censoring onto each plot with 1, 2, and 3 curves in both base R graphics and ggplot. We compared automated digitization and manual digitization performed by well-trained researchers. We calculated the root mean squared error (RMSE) at 100 time points for both methods. The algorithm's performance was also evaluated by Bland-Altman analysis for the agreement between automated and manual digitization on a real-world set of published KM curves.

Results

The automated digitizer accurately identified survival probabilities over time in the simulated KM curves. The mean RMSE for automated digitization was 0.011, while manual digitization had a mean RMSE of 0.013. The algorithm's performance was negatively correlated with the number of curves in a figure and the presence of censoring markers and positively correlated with sample size. In real-world scenarios, automated digitization and manual digitization showed close agreement.

Discussion/Conclusion

The algorithm streamlines the digitization process and requires minimal user input. It effectively digitized KM curves in simulated and real-world scenarios, demonstrating accuracy comparable to conventional manual digitization. The algorithm has been developed as an open-source R package and is publicly available on GitHub: https://github.com/Pechli-Lab/SurvdigitizeR, and a Shiny App: https://pechlilab.shinyapps.io/Digitizer_APP/.

[P54] Time Motion-Study for Artificial Intelligence Automation to Improve Family Medicine Workflow: Protocol for a Mixed Methods Study

Karen Li, Temerty Faculty of Medicine, University of Toronto
Noah Crampton, Toronto Western Hospital; University of Toronto
Andrew Pinto, Unity Health Toronto; University of Toronto
Hanu Chaudhari, University of Toronto Department of Family and Community Medicine
Omri Nachmani, University of Toronto Department of Family and Community Medicine
Stephanie Garies , Unity Health Toronto
Serena Jeblee, University of Toronto
Jane Zhao, University of Toronto's Institute of Health Policy, Management and Evaluation (IHPME)
Christopher Meaney, University of Toronto

Introduction

The performance of administrative tasks using electronic medical records (EMRs) is one of the leading causes for physician burnout. Family medicine has one of the highest frequencies of EMR use amongst all specialties. Artificial intelligence (AI) can help reduce the clerical burden associated with EMRs. However, no previous studies have characterized the time profile of primary care providers (PCPs) using EMR in a Canadian context.

Methods

This study is a mixed-methods study with three components: (1) Digital charts will be reviewed to measure the types and volumes of particular clerical tasks (e.g. prescription, requisition, referral) emerging from patient encounters in primary care. Total EMR work time, if available, will also be recorded and analyzed to determine associations between work time and task events. (2) PCPs will be directly observed at their clinical sites. Observers will document EMR activities on the WorkStudy+ software using a predetermined process-task coding scheme. Results from participant demographic surveys will be used to ensure equal representation of each major EMR system. (3) Focus groups will be held with PCPs from various clinical sites to explore staff perspectives regarding EMR use.

Results

The primary outcome is a time utilization profile detailing the proportion of time PCPs spend on each aspect of the EMR system. Outcomes of interest from focus group discussions include EMR usability pain points, opportunities for automation in EMR, and anticipated barriers and facilitators for the implementation of AI tools in primary care.

Discussion/Conclusion

This will be a novel area of study in Canada. Findings will provide a roadmap for the successful design, development, and implementation of AI-based technologies that automate EMR tasks in primary care. Such technologies aim to improve staff work environments and to reduce provider burnout.

[P55] Understanding Patterns in Variants of Uncertain Significance to Facilitate Reclassification Using Machine-learning Based Variant Effect Predictors

Cindy Zhang, University of Toronto

Daniel Zimmerman, University of Toronto

Frederick Roth, University of Toronto

Robert C Grant , Princess Margaret Cancer Centre - UHN

Introduction

Genomic sequencing is primed to advance the diagnosis and risk assessment of a broad range of diseases. When rare missense variants are clinically interpreted, most are classified as variants of uncertain significance (VUS), which means it offers essentially no guidance to clinicians or patients. Reclassifying a VUS, e.g., to pathogenic or benign, usually relies on functional evidence, such as with multiplexed assays of variant effects (MAVEs). However, such results are generally not available. To fill this gap, machine-learning based variant effect predictors such as VARITY are used to predict the pathogenicity of a variant. In our study, we aimed to relate variant interpretations to predictor scores, which will facilitate VUS reclassification.

Methods

We used ClinVar to extract clinically interpreted variants and mapped them to pathogenicity scores obtained from VARITY. Then, we used Python to visualize and fit the data. We propose a method of prioritizing genes based on the likelihood of each VUS being pathogenic or benign.

Results

We found that VARITY scores showed a greater correlation to ClinVar variant interpretations than other predictors such as REVEL. Using a likelihood threshold of 0.95, we found genes with the greatest number of movable variants. For example, we found MSH6 and CHEK2 to be among the top five for having pathogenic-leaning VUS, whereas BRCA1 and PALB2 are among the top five for having benign-leaning VUS.

Discussion/Conclusion

Our study is not only useful for testing the efficacy of computational predictors but also to see which genes to prioritize for VUS reclassification. Our results could be used to guide variant effect predictor development and systematic functional testing, ultimately resulting in greater translation to clinical settings.

[P56] Unveiling Potential Hidden Bias in Automated Lateral Spine Image Interpretation: Predicting Demographic and Anthropometric Characteristics using Convolutional Neural Networks

Barret A. Monchka, George & Fay Yee Centre for Healthcare Innovation, University of Manitoba
Douglas Kimelman, University of Manitoba
Parminder Raina, McMaster University
William D. Leslie, University of Manitoba

Introduction

Convolutional neural networks (CNNs) can automate vertebral fracture (VF) recognition in lateral spine images with accuracy comparable to human experts, but generalizability has not been well established. CNNs trained to detect VFs may be unknowingly incorporating demographic and anthropometric characteristics into prediction algorithms—potentially leading to biased models. This study aims to evaluate the ability of CNNs to identify demographic and anthropometric characteristics from vertebral fracture assessment (VFA) images.

Methods

VFA images were acquired from the Manitoba Bone Mineral Density Registry between 2010 and 2017 (n=12,742) and the Canadian Longitudinal Study on Aging (CLSA) at baseline (n=28,651). Manitoba VFAs were randomly split into training (80%) and validation (20%) sets, while CLSA VFAs were randomly assigned for training (60%), validation (10%), and future use (30%). Performance evaluation was conducted on the validation sets. Self-identified ethnicity as White (n=9,083) and Asian (n=303) was available for Manitoba VFAs acquired during 2010-2016. Age and body mass index (BMI) were predicted through regression and evaluated using mean absolute error (MAE) and mean absolute percentage error (MAPE).

Results

Ethnicity (White vs. Asian) was identified with high precision: positive predictive value=81.2%, F1-score=55.9%, AUC=0.97 (Table). Sex was recognized with high balanced accuracy in both CLSA (balanced accuracy=99.0, F1-score=99.0, AUC=1.00) and Manitoba (balanced accuracy=95.5, F1-score=79.9, AUC=0.99) images. Age and BMI were accurately predicted as continuous variables for CLSA (Age: MAE=3.84 years, MAPE=6.02%; BMI: MAE=1.69 kg/m2, MAPE=6.45%) and Manitoba (Age: MAE=4.31 years, MAPE=5.82%; BMI: MAE=3.56 kg/m2, MAPE=12.69%).

Discussion/Conclusion

CNNs recognize demographic and anthropometric characteristics in VFA images with high balanced accuracy. Understanding the extent to which CNNs incorporate demographic and anthropometric traits into VF prediction is crucial for deploying fair and unbiased models in clinical settings with diverse patient populations.

Supporting information

Table. Convolutional neural network performance for demographic characteristic recognition in lateral spine images

Characteristic	Data source	Balanced accuracy	Sensitivity	Specificity	PPV	NPV	F1-score	AUC
Ethnicity								
White (ref) vs. Asian	Manitoba	71.1	42.6	99.7	81.2	98.1	55.9	0.97
Sex								
Male (ref) vs. female	CLSA	99.0	98.9	99.1	99.1	99.0	99.0	1.00
Male (ref) vs. female	Manitoba	95.5	97.2	93.9	99.6	69.5	98.4	0.99
Age (years)								
<65 (ref) vs. 65+	CLSA	84.5	76.8	92.2	88.7	83.3	82.3	0.94
<75 (ref) vs. 75+	Manitoba	70.9	73.0	68.8	74.2	67.4	73.6	0.78

AUC = area under the receiver operating characteristic curve, balanced accuracy = mean of sensitivity and specificity,
CLSA = The Canadian Longitudinal Study on Aging, Manitoba = Manitoba Bone Mineral Density Registry,
NPV = negative predictive value, PPV = positive predictive value, ref = reference category. Classification threshold = 0.5.

[P57] Using Speech Features in a Random Forest Machine Learning Model to Predict COPD Symptoms

Sashini Kosgoda, University Health Network
Robert Wu, University Health Network
Maryann Calligan, University Health Network
Alex Mariakakis, University of Toronto
Sejal Bhalla, University Health Network
Andrea Gershon, Sunnybrook Health Sciences Centre
Alexandre R Zlotta*, Mount Sinai Hospital

Introduction

Chronic obstructive pulmonary disease (COPD) is a prevalent chronic lung disease in older adults. Older adults with COPD often do not seek medical attention until they experience an exacerbation. To address this issue, we developed a machine-learning model which uses speech features to predict COPD symptoms. We aimed to understand important speech features for predicting different COPD symptoms before individuals experience exacerbations.

Methods

Voice recordings were collected using smartwatches from 7 patients over 6 months. Daily symptoms were recorded on a smartphone app and daily symptom scores were calculated to determine the occurrence of an exacerbation. We extracted speech features using the PulmoListener, an end-to-end speech processing pipeline using the openSMILE framework including loudness, pitch and formants. A Random Forest machine learning model was used to detect the occurrence of any COPD symptom and trained using K-fold cross-validation and feature selection (k=42). Given the imbalance in symptom outcomes, we employed different resampling techniques to improve performance. The average F1 score, accuracy and feature importance for each speech feature by symptom was calculated.

Results

The best-performing model using random oversampling achieved an F1 score of 0.488 and an accuracy of 0.474. The most important feature in that model involved loudness. Testing each symptom as a separate model, we saw that the F1 scores were much lower. Further, only 15 of the 42 selected speech features had some importance in classifying patients into the correct symptom.

Discussion/Conclusion

We suspect the lower F1 scores by individual symptoms may be due to the imbalance in the classes of each symptom. These identified speech features and model show promise in developing further models to explore the utility of speech features as a biomarker to predict symptoms.

[P58] Utilizing image denoising and machine learning segmentation to quantify fluid volume in eyes with vascular retinal diseases: the STATIC study

Mohammad Khan, University of Toronto
Samantha Orr, VRM Toronto
Amin Hatamnejad, University of Toronto
John Golding, VRM Toronto
Simrat Sodhi, University of Cambridge
Austin Pereira, University of Toronto
Anuradha Dhawan, VRM Toronto
Niveditha Pattathil, VRM Toronto
Netan Choudhry, VRM Toronto

Introduction

Purpose: To utilize a combination of a denoising algorithm with an automated OCT segmentation algorithm to quantify intra-retinal fluid (IRF) and subretinal fluid (SRF) volumes in patients with retinal vascular diseases.

Methods

Swept-source OCT was used to acquire volume scans of the macula (Triton, Topcon) in patients with diabetic macular edema. Visual acuity (VA) at these time points was also collected. Volume scans were processed with Topcon's denoising algorithm set at 75%. Original and denoised images were then analyzed using an automated OCT segmentation algorithm which provided IRF and SRF volume measurements, which were compared to assess impact of denoising. Fluid volumes were also correlated with 6-month change in VA using the Pearson Correlation Coefficient (PCC); the difference in correlations was examined using a Wald test.

Results

Forty eyes were included in the final analysis. There was a significant difference between paired IRF volumes measured from denoised and original images (P0.05).

Discussion/Conclusion

Denoising led to a significant difference in the IRF volume measurement. Denoising has potential to improve weak OCT signals, as seen with the significant improvement in correlation between IRF and 6-month change in VA. Denoising does not improve correlations where no signal is present, as seen with the correlation between SRF and 6-month change in VA.

[P59] Evaluating the efficacy of an automated, voice-based swallowing dysfunction screening tool utilizing convolutional neural networks

Rami Saab, Sunnybrook Research Institute

Arjun Balachandar, Department of Medicine, University of Toronto, Toronto, Canada

Hamza Mahdi, Sunnybrook Research Institute

Eptehal Nashnoush, Sunnybrook Research Institute

Houman Khosravani , Division of Neurology, Department of Medicine, Sunnybrook Health Sciences Centre, University of Toronto

Introduction

Swallowing dysfunction, dysphagia, is a serious post-stroke complication and a significant cause of mortality, thus timely screening is critical. Existing tools to screen for dysphagia include the gold standard (a barium swallow test) or screening tools administered by trained healthcare professionals. Each approach presents draw-backs including costs, human-resource requirements and subjectivity in the analysis. The deficiencies of existing tools mean patients are often left waiting for a swallowing assessment and are, as a precaution, prohibited from intaking food orally, negatively impacting their quality of life and health outcomes. In this study, we examine the application of convolutional neural networks (CNNs) to rapidly classify patient swallowing status using voice samples alone.

Methods

Vocal samples from 68 post-stroke patients on the neurovascular ward at Sunnybrook Hospital were studied (average age 68±16). These samples consisted of vowel sounds as well as speech components of the National Institute of Health (NIH) stroke scales. Patients were labeled according to their dysphagia screening status; the Toronto Bedside Swallowing Screening Test (TOR-BSST). Individual vocal samples were then segmented into 1,579 audio clips and converted into 6,655 Mel-spectrograms which were used to train two convolutional neural networks (DenseNet and ConvNext separately and in ensemble).

Results

Clip-level and patient-level swallowing-status predictions were obtained through an unweighted averaging ensemble method. The ensemble network demonstrated an F1-score of 0.81 and area under the receiver operating characteristic curve of 0.912 with a sensitivity and specificity of 0.89 and 0.79 respectively.

Discussion/Conclusion

Our study demonstrates the feasibility and effectiveness of applying state-of-the-art CNNs to classify Mel-spectrogram images of vocalizations for the detection of post-stroke dysphagia. This study is relevant to healthcare professionals caring for stroke patients and may offer an avenue for developing rapid, non-invasive, and more objective dysphagia screening tools.

[P60] Using the Wizard of Oz methodology to build a healthbot to improve medication adherence for smoking cessation

Kamna Mehra, CAMH
Sowsan Hafuth, CAMH
Mackenzie V. Earle , CAMH
Ryan Ting A Kee, Centre for Addiction and Mental Health
Matt Ratto, University of Toronto
Jonathan Rose , University of Toronto
Scott Veldhuizen, CAMH
Laurie Zawertailo, CAMH
Peter Selby, CAMH

Introduction

The Wizard of Oz (WoZ) methodology allows researchers to observe user interactions with a healthbot and gather valuable data and feedback, while manually controlling the system's responses behind the scenes. It is a cost-effective and efficient way to test and refine the user experience before investing resources in building a fully automated solution. We present our specific WoZ methodology to refine a preliminary healthbot whose goal is to increase varenicline adherence.

Methods

Nineteen participants (target n=40) interacted with the healthbot for 10-30 minutes per day for three days/asked 5-10 questions per day through a text message interface. The responses were controlled by a "wizard" (a research staff member), who responded with pre-defined responses. This library was iteratively expanded as participants asked new questions. At the end of their interaction, participants were interviewed over the phone about their experiences. Data from the healthbot interactions and interviews were descriptively analyzed.

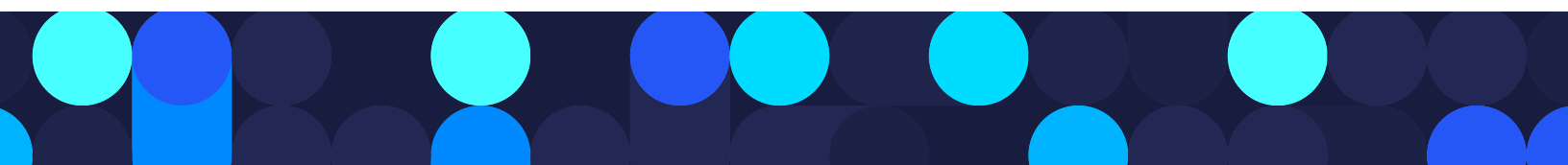
Results

The WoZ methodology helped explore additional information needed by participants, such as details of the effectiveness of the medication, rate of side effects experienced, motivation and strategies to increase medication adherence and quitting smoking. Some novel responses also had to be developed about the healthbot itself, for example, how the healthbot aims to help people, and if the healthbot thinks users can quit smoking, and how it is different from ChatGPT.

The interview data revealed that all participants found the healthbot easy to access, trustworthy, and the language easy to understand. Some participants (22%) mentioned they needed more information or personalized answers to their questions.

Discussion/Conclusion

The WoZ approach proved to be a valuable and pragmatic method in research, facilitating the collection of meaningful data and evaluation of user interactions with the simulated system, thereby offering insights to inform the development of future automated systems.



Thank you for joining

T-CAIREM's free membership is open to students, faculty, researchers, clinicians, and staff at dozens of partner universities, research centres, and hospitals. Join us!
tcairem.utoronto.ca/join-us