2025 T-CAIREM Conference

The Evolution of Generative A.I.

November 14, 2025 • Marriott Downtown at CF Toronto Eaton Centre

ABSTRACT BOOKLET

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

PRESENTATION ABSTRACTS

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

[OR-10-01] Evaluation of a Safe and Collaborative AI Agent for Clinical Decision Support in Liver Transplant Candidate Assessment

Ghazal Azarfar, University Health Network

"Introduction: Transplantation is one of the few areas in medicine where the definitive treatment is rationed. Subjective decision-making pose challenges towards the transplant selection process. Large language models (LLMs) through autonomous artificial intelligent (AI) agents have exciting implications in objective decision-making to solve complex problems, but there are gaps in literature regarding its clinical applications. Thus, we examined the performance of a multidisciplinary selection committee of AI agents (AI-SC) in the liver transplant (LT) selection process as a proof-of-concept towards objective decision-making.

Methods: This was a hybrid cohort study of adult (≥18-years-old) LT candidates between 2004-2023 from the Scientific Registry of Transplant Recipients (SRTR) database. Patients receiving LT were retrospectively analyzed and a hypothetical cohort of patients with standard absolute contraindications to LT were generated. The AI-SC's performance to 1) waitlist candidates if LT would offer a 6-month or 1-year survival benefit or 2) decline candidates if contraindications to LT were present or if LT would not offer a survival benefit were analyzed. The AI-SC consisted of four LLMs (transplant hepatologist, transplant surgeon, cardiologist, and social worker).

Results: Of 8,412 patients, 83·6% were waitlisted and 16·4% had contraindications to LT. The AI-SC was able to accurately identify contraindications to LT (accuracy: 98·2%, 95%CI 97·9%-98·4%), predict 6-month (94·9%, 95%CI 94·4%-95·3%), and 1-year (92·0%, 95%CI 91·4%-92·6%) survival. HCC burden beyond Milan criteria was the most common reason for accepted patients who were declined (False Negative). Malignancy was the most common cause of death prior to 6-month or 1-year end points (False Positive).

Discussion/Conclusion: LLMs can be leveraged through a Multi-AI agent system to simulate the LT-SC meetings and provide accurate, objective insights on patients who may or may not benefit from LT. Lessons learned from this study are a provocative step towards making the LT selection process more equitable and objective."

[OR-10-02] Building Trust: LLM-as-a-Judge for the evaluation of chatbots for adolescent and young adult patients living with Type 1 Diabetes

Xinyi Liu, University of Toronto
Pedro Velmovitsky, University of Toronto

Target Audience:
Clinicians, Clinician-Investigators, Researchers, Computer and Data Scientists, Statisticians, Engineers, and Health System Administrators interested in the design, deployment, and real-time evaluation of Large Language Models (LLMs) in clinical practice.

Introduction:
Large Language Models (LLMs) hold great transformative potential for healthcare, but their clinical utility is challenged by potential low-quality responses which affect safety. This study addresses the need for real-time, continuous evaluation of LLMs deployed in patient-facing applications, specifically a chatbot designed to support adolescents and young adults with Type 1 Diabetes (T1D) transitioning to adult care.

Methods:
An LLM-based T1D support chatbot, built with a Retrieval-Augmented Generation (RAG) approach, was developed. An LLM-as-a-Judge (LLM-J) evaluator was co-designed with two clinicians and two patient partners. The LLM-J was tested on synthetic and real patient T1D questions, evaluating Accuracy, Safety, and Global Quality Score (GQS). Agreement between human raters and the LLM-J was analyzed using Gwet's AC2, and scoring consistency was tested with the Wilcoxon signed-rank test. A scalability analysis measured time and cost on 2,856 synthetic questions.

Results:
Agreement between human rater groups and the LLM-J was reliable for all metrics ($p<0.05$). The Safety metric in particular showed strong agreement between the LLM-J, clinicians (AC2 = 0.74), and patient partners (AC2 = 0.78). The Wilcoxon signed-rank test confirmed no significant difference in Safety scores between the human groups and the LLM-J, suggesting the LLM produces similar scores to human experts. The T1D chatbot consistently performed well, with a Safety score of 1 (no major concerns) given by the majority of human raters. Scalability analysis demonstrated that the system is fast and cost-effective, replying and evaluating 2,856 questions in 4 hours for $89.88 CAD.

Discussion/Conclusion:
The chatbot shows promise in safely supporting T1D self-management. Crucially, the LLM-J provides a reliable, efficient, and cost-effective method for establishing continuous, real-time safety guardrails to mitigate the risks associated with deploying LLMs in high-stakes healthcare environments."

[OR-10-03] Feasibility of an LLM-based decision aid for collaborative decision-making in juvenile idiopathic arthritis: a multimethod evaluation

Antonia Barbaric, University of Toronto
Susa Benseler, University of Calgary
Chitra Lalloo, The Hospital for Sick Children
Celine Liu, University of Toronto, Toronto
Alex Mariakakis, University of Toronto
Quynh Pham, University of Toronto
Joost Swart, University Medical Center Utrecht
Bas Vastert, Utrecht University
Rae S M Yeung, University of Toronto
Joseph A. Cafazzo, University of Toronto

"Introduction
Juvenile idiopathic arthritis (JIA) is the most common pediatric chronic inflammatory rheumatic condition, with substantial quality-of-life impacts. Treatment selection is complex: limited guideline evidence, unpredictable response, and difficult trade-offs, so choices rely on physician expertise and should integrate patient preferences. We evaluated the feasibility of an LLM-based decision aid (LLM-DA) to support collaborative decision-making in JIA.

Methods
We conducted a multi-method feasibility study to identify patient-perceived value of the LLM-DA and quality of DA based on evaluations from rheumatologists and an LLM-as-a-Judge. Patients (n=34) participated in a qualitative descriptive study, through focus groups and interviews, and were analyzed thematically. Three rheumatologists and LLM-as-a-Judge independently reviewed 10 de-identified cases, generated treatment recommendations, and rated LLM-DA outputs on safety, factual consistency, treatment appropriateness, and acknowledgment of patient preferences (1-3; acceptable=2-3). We computed Gwet's AC1 for agreement and compared treatment selections among real-world care, clinicians, and the LLM-DA.

Results
Five patient themes were identified: the tool complements clinician expertise, prepares patients through reflection and actionable information, builds confidence for self-advocacy, value depends on the patient-provider relationship, and potential for broader JIA self-management. Clinicians rated 97% of outputs acceptable for safety and preference acknowledgment, 60% for factual consistency, 70% for appropriateness. Treatment recommendation agreement was low (Table 1). Inter-rater agreement among rheumatologists was almost perfect for safety and preference acknowledgement, moderate for treatment appropriateness, and poor for factual consistency, and remained the same or improved with LLM-as-a-Judge.

Discussion
The majority of clinician ratings across all metrics indicates the LLM-DA was acceptable, and patients reported better preparation, confidence, and dialogue. The observed wide clinician variability in scoring and treatment recommendations underscores the existence of multiple acceptable treatment approaches and highlights the critical importance of integrating patient preferences. Notably, the inter-rater agreement among rheumatologists was not consistently high across all metrics and LLM-as-a-Judge demonstrated comparable performance."

[OR-10-04] AI4TO: Towards Empathic Agentic Large Language Models for Supporting Immigrant Breast Cancer Survivors in Toronto

Divy Wadhwani, University of Toronto Scarborough
Elena Ospiyan, University of Toronto Scarborough

"Immigrant women in Toronto diagnosed with breast cancer experience persistent barriers to culturally and linguistically appropriate survivorship support. Available digital health resources are largely restricted to English and fail to address distinct emotional and practical needs related to migration, language, and community belonging. This lack of tailored guidance contributes to reduced trust, delayed care navigation, and discontinuities in follow-up engagement.

This study aims to evaluate AI4TO, an artificial intelligence–based conversational agent designed to deliver multilingual, empathetic, and culturally responsive support for immigrant breast cancer survivors in Toronto. The system integrates Chain-of-Thought and Reality Therapy Chain-of-Empathy (RT-CoE) prompting to enhance reasoning, emotional sensitivity, and cultural adaptability.

The AI4TO framework employs Qwen-3 235B as its core large language model, embedded within a modular architecture that includes dynamic web retrieval, long-term memory, and multilingual processing. The agent's system prompt enforces strict ethical and clinical boundaries while guiding responses through RT-CoE stages: emotion recognition, trigger inference, unmet-need identification, and agency-aligned response. Emotional intelligence was quantitatively assessed
using EQ-Bench, a standardized benchmark for evaluating empathic reasoning in large language models. The system's performance was compared with 46 peer models across commercial and open-weight benchmarks.

Integration of RT-CoE prompting substantially improved empathic reasoning and response authenticity. AI4TO achieved an Elo score of 1315.4 (95% CI 1301–1329; $\sigma = 3.9$), ranking eighth among 47 models and within the top quintile of the EQ-Bench leaderboard. This performance surpassed multiple commercial systems, including Claude Opus 4, GPT o4-mini, and Gemini 2.5-pro. The improvement relative to the baseline Qwen-3 model (Elo 1274) was attributed to structured empathic reasoning rather than scale. AI4TO's multilingual components effectively maintained emotional tone across languages, though minor limitations persisted in cultural nuance and retrieval precision.

AI4TO demonstrates that combining Chain-of-Thought reasoning with structured empathic prompting enables large language models to deliver linguistically accessible and emotionally resonant support for immigrant breast cancer survivors. Its results highlight the feasibility of safe, multilingual AI counseling systems that integrate live information retrieval and ethical guardrails. Ongoing work focuses on clinical pilot testing, participatory co-design with survivor groups, and expanded benchmarking for inclusivity and cultural relevance."

[OR-5-05] Machine Learning for the Prediction of Massive Hemorrhage in Trauma: A Systematic Review, Meta-Analysis, and APPRAISE-AI Study

Anglin Dent, University of Toronto
Gemma Postill, University of Toronto
Richard Cheng, Queens University
Anton Nikouline, University of Alberta Hospital
Teruko Kishibe, Unity Health Toronto
Melissa McGowan, St. Michael's Hospital, Unity Health, Toronto
Jethro Kwong, University of Toronto
Brodie Nolan, St. Michael's Hospital, Unity Health, Toronto

"Background: Timely provision of appropriate blood supplies is an essential aspect of care for patients with traumatic injuries. In the literature, machine learning (ML) algorithms have demonstrated potential for automated prediction of Massive Hemorrhage Protocols (MHP). However, the model quality and performance of such algorithms has yet to be comprehensively evaluated.

Methods: We conducted a systematic review and meta-analysis of studies published between January 1, 2004, to March 24, 2025, reporting the performance of ML models prediction of MHP need among patients with traumatic injuries. In duplicate, title and abstracts and then full texts were screened. Data was extracted from included texts. A meta-analysis using a random-effects model was completed to compare performance metrics. A bivariate mixed-effects model and summary receiver operating characteristics curve was completed to assess the sensitivity and specificity for each ML model type. Quality was assessed using APPRAISE-AI.

Results: Following review of 4542 abstracts, 21 studies (capturing 50 ML models) met inclusion criteria. Meta-analysis revealed ML models to outperform the clinical reference model (Assessment of Blood Consumption [ABC] score) in area under the curve for the receiver-operating characteristic (AUCROC), sensitivity, and specificity. ML model types with the highest pooled AUROC were random forest (0.89, 95% CI: 0.85-0.93), neural networks (0.88, 95% 0.83-0.93), and XGBoost (0.86, 95% CI: 0.80-0.93). XGBoost (sensitivity: 0.86, 95% CI: 0.82-0.90; specificity: 0.82, 95% CI: 0.81-0.83) and neural networks (sensitivity: 0.86, 95% CI: 0.82-0.90; specificity: 0.80, 95% CI: 0.72-0.87) demonstrated the most promising pooled sensitivity and specificity. Models generally scored moderately by APPRAISE-AI assessment.

Conclusions: ML models outperformed traditional clinical prediction tools for MHP prediction. Overall, the reporting quality and reproducibility of ML models was low. Future ML algorithm types should improve reporting, specifically of model training data, data preprocessing, and subgroup performance, to facilitate comparison of model development procedures and increase confidence in newly developed models for MHP prediction."

[OR-5-06] Leveraging large language model personas for synthetic instrument validation of trust assessment instruments in pediatric emergency and surgical care

Ella Boone, McGill University
Katya Loban, McGill University
Elena Guadagno, The Montreal Children's Hospital, Montreal
Dan Poenaru, McGill University

"Background: Trust is foundational to patient-physician relationships and is associated with improved care-seeking and adherence in primary care. However, validated trust instruments for pediatric emergency and surgical contexts are lacking, and traditional instrument development is slow and resource-intensive. Large language models (LLMs) could streamline the validation process by serving as scalable, systematic expert panel surrogates.

Methods: We developed four trust assessment instruments: Experience of Trust in Pediatric Surgical and Emergency Care (ETP), Building Blocks for Trust (BBTP) for patient families, Physician-Perceived Trust in Pediatric Emergency Scale (PPTS) and Physician-Identified Building Blocks for Trust (PIBBT) for physicians. We developed expert personas in three LLMs (Claude Sonnet 4-Thinking, GPT-5-Thinking, Grok4), then completed 10 evaluation runs per instrument using RAND/UCLA Appropriateness methodology, with 7 human experts as ground-truth comparators. Both Scale-Content Validity Index (S-CVI) and Fleiss' kappa for inter-rater reliability targets were set at ≥0.80.

Outcomes: Inter-rater reliability across the three LLMs and human experts revealed high agreement for patient family and physician instruments in dimensional validation (Fleiss' κ = 0.84, 95%CI [0.72, 0.96]; 0.87, 95%CI [0.72,1.00]). For contextual validation, Claude Sonnet 4-Thinking and human experts achieved a high level of agreement across instruments (Fleiss' κ = 0.83, 95%CI [0.73,0.93]; 0.88, 95%CI [0.80,0.96]). Both validation rounds achieved strong S-CVI scores across all instruments (dimensional, contextual: ETP:0.98;0.94, BBTP:0.91;0.97, PPTS:1.00;0.98, PIBBT:0.90,0.96). This supports the validity of LLM-assisted instrument validation methodology.

Conclusions:
Persona-prompted LLMs can efficiently supplement human expertise for instrument validation, maintaining scientific rigor while shortening timelines. This synthetic instrument validation can streamline trust research and ultimately enhance patient-centered care."

[OR-5-07] Digital Twins of Ex Vivo Human Lungs Enable Accurate and Personalized Evaluation of Therapeutic Efficacy

Xuanzi Zhou, University Health Network, Toronto
Bo Wang, University of Toronto
Yiyang Wei, University Health Network, Toronto
Serena Hacker, University Health Network, Toronto
Sumin Kim, University Health Network, Toronto
Thomas Borrillo, University Health Network, Toronto
Abby McCaig, University Health Network, Toronto
Haaniya Ahmed, University Health Network, Toronto
Youxue Ren, University Health Network, Toronto
Olivia Hough, University Health Network, Toronto
Luca Orsini, University Health Network, Toronto
Bonnie T. Chao, University Health Network, Toronto
Micheal McInnis, University of Toronto
Marcelo Cypel, University Health Network, Toronto
Mingyao Liu, University Health Network, Toronto
Jonathan C. Yeung, University Health Network, Toronto
Lorenzo Del Sorbo, University Health Network, Toronto
Shaf Keshavjee, University Health Network, Toronto
Andrew T. Sage, University Health Network, Toronto

"Background: Digital twins show great potential in transplant medicine for evaluating new therapies to repair injured donor organs. Ex vivo lung perfusion (EVLP) sustains donor lungs before transplantation and generates real-time, multi-modal data, offering a unique opportunity to train machine learning (ML) models to forecast lung function and create digital twins of human lungs. Herein, we developed a ML-based approach that leverages >300 ML algorithms that work simultaneously to accurately simulate future lung function. As a proof of concept, we demonstrate the clinical utility of digital twins of ex vivo lungs for the preclinical evaluation of the safety and efficacy of an EVLP-targeted therapy.

Methods: Lung physiology, biochemistry, proteomic and metabolomic biomarkers, transcriptomics, and imaging features were derived from n=1000 EVLP cases performed at our centre (2008-2024) (Fig. 1). For each parameter, a multi-modal time-series forecasting model (XGBoost, gated recurrent unit) was trained to predict future lung function using baseline EVLP data, with mean absolute percentage error as the primary model evaluation metric. Therapeutic efficacy (pulmonary arterial pressure, PAP) and safety (edema) of a thrombolytic treatment were evaluated using the digital twins of n=16 EVLP cases with suspected pulmonary emboli.

Results: A digital twin of an ex vivo human lung was successfully developed, accurately predicting lung function across more than 75 distinct functional parameters measured during EVLP. The digital twin approach achieved >90% accuracy in both short- and long-term forecasting of lung function. In a cohort of n=16 human lungs treated with tPA, the digital twin paired analyses approach revealed a significant reduction in PAP within two hours of treatment (p-value = 0.031). In contrast, using a conventional two-arm study design, no significant differences in PAP were observed at 1- or 2-hours post-treatment. Thus, only a digital twin approach was able to identify preclinical treatment-induced reductions in PAP.

Conclusion: Digital twins enable direct comparisons between treated organs and their personalized virtual models, advancing precision medicine at the organ level. By creating digital twins of ex vivo lungs, researchers can conduct more advanced preclinical evaluations of therapies, leading to more efficient clinical trials in transplant medicine."

[OR-5-08] MEDfl: A Collaborative Framework for Federated Learning in Medicine

Ouael Nedjem Eddine Sahbi, Université de Sherbrooke
Martin Vallières, McGill University
Cynthia Gagnon, Université de Sherbrooke
Bessam Abdulrazak, Université de Sherbrooke
Haithem Lamri

"Medical institutions possess vast amounts of valuable clinical data essential for training artificial intelligence (AI) models. However, strict data protection regulations such as GDPR and HIPAA restrict data sharing across hospitals, limiting dataset diversity and scale. In addition to these regulatory barriers, there remains a pressing need for tools that facilitate collaboration between computer scientists and medical researchers. Despite advances in AI, interdisciplinary cooperation is still hindered by the absence of integrated, user-friendly platforms bridging the gap between technical and clinical domains.

To address these challenges, we propose MEDfl, a federated learning (FL) framework enabling collaborative research AI modeling across multiple hospitals without moving sensitive data outside institutional boundaries. Each site retains full control over its datasets, while only encrypted model updates are exchanged with a central coordinating server. MEDfl integrates several complementary technologies: Transfer Learning (TL) to improve performance in data-scarce environments, Differential Privacy (DP) to prevent potential re-identification of patients, and Secure Aggregation. Beyond these privacy and performance aspects, MEDfl also tackles the collaboration challenge between computer scientists and medical researchers. It combines a user-friendly interface that allows researchers to design and launch federated learning experiments without deep programming knowledge, automated code generation that translates configurations defined in the interface into executable pipelines, and an integrated code package.

The first version of the MEDfl package is already available on PyPI (https://pypi.org/project/MEDfl/), supporting both simulation and real-world deployment. We provide YouTube tutorials (https://www.youtube.com/playlist?list=PLEPy2VhC4-D7Y4lkGMRpHG8ydVZQonkMJ) and open-source repositories for the Python package (https://github.com/MEDomicsLab/MEDfl) and the desktop application (https://github.com/MEDomicsLab/MEDomics/tree/dev_medfl_sqllite). Using FedAvg, models with transfer learning achieved faster and stronger performance (AUC 0.85–0.9) compared to models without it (AUC 0.75–0.78), confirming the benefit of combining FL and TL for faster convergence and improved adaptation across heterogeneous clients. These experiments were conducted on a publicly available binary classification dataset on diabetes from Kaggle, containing 100,000 instances.

Expected outcomes include preserving data privacy without hindering multi-institutional research, empowering computer scientists to easily configure and compare FL pipelines, and simplifying collaboration between medical and technical researchers through accessible, no-code tools."

[OR-5-09] AI-enabled OSCE: A Modular Platform for Simulated Clinical Assessment and Rubric-based Feedback Using Large Language Models

Shlok Panchal, McMaster University
Alice Kam, University of Toronto

"Introduction
Objective Structured Clinical Examinations (OSCEs) are the established standard for evaluating clinical reasoning, communication, and professional behaviour through simulated patient interactions. Despite their educational value, OSCEs require extensive resources, including standardized patients, trained examiners, and logistical support. This limits their scalability and frequency in medical training. Recent advances in large language models (LLMs) have the potential to automate aspects of simulation and feedback generation, creating a consistent and accessible training environment. This study aims to develop an AI-enabled OSCE platform that uses LLMs to simulate patient encounters and generate feedback aligned with standard clinical communication and reasoning rubrics.

Methods
A modular system was developed with three core components: (1) a learner interface that simulates patient dialogue through a constrained LLM and supports voice input via speech-to-text; (2) a facilitator dashboard for case creation, rubric selection or upload, and session control; and (3) an analysis module that uses retrieval-augmented generation to link learner responses to rubric criteria and reference materials such as marking guides or clinical frameworks. A vector database supports targeted retrieval, and prompt constraints maintain consistent role behaviour and structured feedback. All interactions, transcripts, and generated outputs are logged for review and performance analysis.

Results
The platform was deployed in a development environment and tested across multiple clinical scenarios. Automated case generation produced consistent patient presentations, including histories, symptoms, and vital signs. The analysis module generated rubric-based feedback summarizing strengths and areas for improvement, while prompt constraints minimized off-role responses and maintained scenario coherence. A pilot study protocol has been prepared to evaluate usability and assess agreement between AI-generated and human scores through correlation and inter-rater reliability analyses.

Discussion/Conclusion
This platform demonstrates a feasible approach to extending clinical assessment using LLMs beyond traditional OSCE formats. It delivers rubric-based feedback that supports learner development while easing the demands of examiners. Future work will evaluate usability, alignment with faculty scoring, and the impact on training accessibility."

[OR-5-10] Applying Machine Learning to Improve Causal Inference: Impact of Non-Pharmaceutical Interventions on SARS-CoV-2 Infection Risk in Children

Mary Aglipay, University of Toronto
Jeff C. Kwong, University of Toronto
Ashleigh Tuite, University of Toronto
Muhammad Mamdani, University of Toronto,
Charles Keown-Stoneman, Unity Health Toronto
Catherine Birken, University of Toronto
Jonathon Maguire, University of Toronto

"Introduction
Children in Ontario experienced the longest COVID-19 restrictions in North America. Measures such as masking, distancing, and staying home reduced community transmission but also affected children's social, educational, and emotional well-being. While population-level effects of these interventions are well studied, little is known about the impact of individual adherence on a child's risk of infection. To address this evidence gap, we emulated a target trial using real-world pediatric data and applied modern causal inference and machine learning methods to estimate the effect of sustained adherence to non-pharmaceutical interventions (NPIs) on SARS-CoV-2 infection risk.

Methods
Data came from the TARGet Kids! COVID-19 Study, a longitudinal cohort of Toronto children (April 2020–March 2023). Eligible participants were aged 0–10 years whose parents completed at least one COVID-19 questionnaire. The intervention compared sustained regular adherence versus minimal adherence to masking, physical distancing, and staying home. We estimated the per-protocol effect of sustained adherence on time to parent-reported infection using a Cox marginal structural model with stabilized inverse probability weights for adherence and censoring. Confounders included calendar time, community incidence, household infection, vaccination status, socioeconomic indicators, and child demographics. To address complex time-varying confounding and enhance model stability, we are applying gradient boosting to optimize the estimation of these weights. This machine learning approach dynamically learns the best weighting functions across time, improving balance diagnostics and reducing model dependence.

Results (In Progress)
Preliminary findings indicate that children with regular adherence to NPIs had lower rates of parent-reported infection, though the effect attenuated when accounting for changing community incidence and vaccination. Ongoing analyses with gradient-boosted weights are expected to yield more robust hazard estimates and improved precision.

Discussion/Conclusion
This study demonstrates the use of machine learning to strengthen causal inference in pediatric epidemiology. Integrating gradient boosting with marginal structural models may help uncover nuanced intervention effects in complex longitudinal data. These results will inform future infection-prevention strategies and exemplify how AI-driven analytic tools can enhance the evidence base for child and family health policy."

[OR-5-11] Improving the Explainability and Performance of Pediatric Brain Tumor Molecular Subtype Classification through Multimodal Learning

Sara Ketabi, University of Toronto
Matthias W. Wagner
Cynthia Hawkins
Uri Tabori
Birgit Betina Ertl-Wagner
Farzad Khalvati

“Introduction: Detecting the genetic markers of pediatric low-grade glioma (pLGG), the most common brain tumor in children, is crucial for targeted treatment planning. Despite the significant performance of convolutional neural networks (CNNs) in diagnosing these genetic markers from brain magnetic resonance imaging (MRI), these models are not widely used in clinical settings mainly due to the lack of explainability. In other words, the features influencing a model's prediction are not typically understandable to radiologists. Radiology reports can be integrated with MR images to improve CNN explainability. In this study, we propose a Contrastive Learning (CL) framework to minimize the distance between the MRI and the corresponding report representations and maximize the distance between the mismatched MRI and reports. The learnt MRI representations can then be applied to improve the explainability and performance of pLGG genetic marker classification.

Materials and Methods: The dataset for this REB-approved retrospective study contains 341 FLAIR MR images for pLGG cases aged between 5 and 18, and associated radiology reports. To encode the MRI and reports, we apply 3D Residual Network (ResNet), initialized with a set of pretrained weights named “MedicalNet”, and a transformer called Clinical Longformer, respectively. The encoded representations are then compared based on the cosine similarity to adjust their distance within a CL framework. Consequently, similar MRI and report representations would be close to each other, while dissimilar pairs would be far apart. Next, we initialize ResNet with the weights extracted from this framework and fine-tune it on the genetic marker classification task using a subset of 204 MRIs. This subset relates to the two most important pLGG genetic markers, i.e., BRAF fusion and BRAF V600E mutation.

Results: Using a 5-fold cross validation, we calculated the area under the ROC curve (AUC) and Dice score between the model’s attention maps and manual tumor segmentation masks. The ResNet model initialized with the proposed weights outperforms a baseline classification model trained from scratch, achieving an AUC of 0.877 +/- 0.072 versus 0.748 +/- 0.083. Furthermore, this initialization increases the attention maps Dice score from 2% to 15.8%.”

[OR-5-12] AI evaluating AI: A study comparing AI-based evaluations of AI-generated lay summaries from radiology reports against human evaluations

Nanziba Tasneem, McMaster University
Christian van der Pol, McMaster University
Ambreen Zahoor, McMaster University
Nitin Juggath, McMaster University
Kyle McGowan, McMaster University
Cynthia Lokker, McMaster University
Ashirbani Saha, McMaster University

"Audience: Healthcare AI researchers and clinicians
Introduction: Human experts' review of AI-generated clinical content is essential to ensure patient safety. However, these reviews can be subjective, with most evaluations requiring multiple raters, leading to increased time commitment of experts and prolonged research timelines. Large Reasoning Models (LRMs), AI systems trained for human-like reasoning, can be used to supplement human evaluations or provide additional assessments. The objective of this study was to evaluate Large Language Model (LLM)-generated lay summaries of radiology reports by simulating an LRM (Gemini 2.5–Pro) both as an expert radiologist and lay participants; and compare the LRM evaluations with experts and matched lay participants, respectively.

Methods: Five LLMs (GPT-4, GPT-4o mini, Gemini 1.5–Pro, Gemini 1.5–Flash, and Llama 3.1) generated lay summaries for 100 radiology reports from the publicly available "BioNLP 2023 Report Summarization" dataset. The LRM (Gemini 2.5-Pro) was prompted to play the roles of (a) an expert radiologist (10+ years' experience) and (b) demographically matched lay participants. Two blinded experts (radiology fellows) evaluated all summaries for quality and accuracy, while 15 lay participants assessed a subset of 15 summaries (3 from each LLM) for understanding and confidence. The percentage agreement between the human and AI opinions was calculated, and Mann-Whitney U tests were performed for pairwise comparison between the lay participants' and LRM's evaluations.

Results: The LRM-as-expert strongly aligned with the experts on quality; with <1% complete polarizing disagreements (agree vs. disagree with both experts) and 25% partial (agree vs. disagree with one expert) polarizing disagreements. The LRM largely agreed with experts on accuracy (<11% disagreement per LLM, except Llama 3.1: 39%). Mostly, the discrepancies occurred as experts judged summaries as accurate while the LRM did not. The LRM-as-layperson consistently overestimated layperson understanding (100% for all LLMs) and had significantly different responses vs. laypersons for confidence ($p<4.65 \times 10^{-5}$).
Conclusion: This study introduces a novel LRM-based proxy evaluation technique, showing strong alignment with expert radiologists and variability for laypersons. It highlights the potential of LRMs in clinical evaluation and the need for refinement in layperson simulation. "

[OR-5-13] Evaluation of Collaborative Human–LLM Workflow for Automated Discharge Summaries for Hospitalists

Ghadir Ali, University of Toronto
Ervin Sejdic, University of Toronto
Nihal Haque, North York General Hospital, Toronto

"Target Audience:
Physicians, AI Researchers

Introduction:
Nearly 50% of physicians' time is spent on documentation-related tasks, reducing time for patient care and contributing to burnout. The rapid success of large language models (LLMs), steered attention to their potential in automating clinical documentation. Prior work shows promise but largely benchmarks LLMs against physician performance. In practice, however, legal and ethical requirements mandate that physicians remain in charge. A more realistic approach is for LLMs to assist in a supervised setting. This creates a trade-off, as clinicians must review and correct the generated text. In this work, we empirically evaluate LLM-assisted documentation in a real-world collaborative workflow.

Methods:
We conducted an exploratory study deploying an LLM-based system to generate discharge summaries from hospitalists' progress notes. Prompting strategies were iteratively refined based on feedback from two physicians. Finally, outputs were evaluated through a structured survey that covered eleven dimensions, spanning accuracy, risk, added value, and open-ended feedback. Semi-structured interviews with three additional physicians provided further qualitative insights on usability and design requirements for clinical integration.

Results:
Physicians rated the system as effective in reducing cognitive workload and time by 40-59%, with low risks to the patient (0.5/10). Incompleteness was identified as the most frequent weakness (5.5/10). Correctness was rated as 7.5/10, and clinical sense was rated as 8.5/10. Qualitative feedback noted that the LLM-generated summary should better reflect the narrative of the patient's presentation. Another insight from the interviews is underspecified follow-up plans in the generated summaries risking critical actions being unaddressed or delayed.

Conclusion:
A supervised, physician–LLM workflow shows strong potential to reduce documentation burden. A highlighted finding is that omissions (incompleteness) are more problematic than hallucinations (inaccuracy), a gap that may stem from LLMs' general-purpose training versus physicians' specialist judgment of clinical relevance. Addressing this mismatch requires further investigation into adapting LLMs to align with physician reasoning and documentation practices through iterative collaboration."

[OR-5-14] Machine learning models for estimating counterfactuals in a single-arm inflammatory bowel disease study

Dan Liu, CHEO Research Institute, Ottawa
Fida Dankar, CHEO Research Institute, Ottawa
Jennifer C. deBruyn, University of Calgary
Amanda Ricciuto, University of Toronto
Khaled El Emam, University of Ottawa

“Target audience: clinicians, researchers

Introduction:
Single-arm clinical trials can accelerate results by reducing the number of patients that must be recruited. Virtual control arms can then use the available external control data to simulate the counterfactual outcomes of the treatment arm. In this study, we build a machine learning (ML) outcome model that can estimate the control (Infliximab - IFX) counterfactual outcomes for the treatment (Adalimumab ADA) arm patients from an existing pediatric Crohn’s disease study.

Methods:
Gradient boosted decision trees and TabPFN models, both with augmentation, were used mto train counterfactual models on historical IFX data. The two IFX models were used to predict the counterfactual outcomes for the ADA arm patients. Odds ratio (OR) with 95% confidence interval (CI) was used to assess the effect of IFX compared to ADA. These results were compared to results obtained using propensity score matching (ground truth), which is the default method for these types of comparative studies.

Results:
Data augmentation using synthetic data generative models improved the performance of both ML models significantly. Light gradient boosting machine yields the best augmented OR closest to the ground truth, and all 95% CI results align with the conclusion from the ground truth study that no statistical difference in the clinical outcomes has been observed between the ADA-treated and IFX-treated patients.

Conclusions:
Our study illustrates the usefulness of virtual controls as an alternative to costly, challenging or unethical patient recruitment in an inflammatory bowel disease trial. The gradient boosted prediction model we developed can be used as a pre-trained model in future studies to predict counterfactual outcomes for the IFX control."

[OR-5-15] Mapping the Adoption of Artificial Intelligence in Canadian Healthcare

Tania Tajirian, University of Toronto
Marissa Binstock, Centre for Addiction and Mental Health, Toronto
Abhishek Chopra, University of Toronto
Francesca D'Angelo, Digital Health Canada, Toronto
Shangjucta Das Pooja, University of Toronto
Mannie Chhabra, University of Toronto
Shveta Bhasker, University of Toronto
Shaaf Farooq, University of Toronto
Stephanie Rintoul, Hamilton Health Sciences, Hamilton
Marwa Naimi, University of Toronto
Victorine Maikem, University of Toronto
Annika Tork, University of Toronto
Shelagh Maloney, Digital Health Canada, Toronto

"Target Audience
Health system leaders, policymakers, digital health researchers, clinicians, and technology developers interested in the implementation and impact of artificial intelligence (AI) in healthcare

Introduction
Artificial intelligence (AI) has transitioned from conceptual promise to applied clinical tools, shaping diagnosis, monitoring, and treatment worldwide. In Canada, however, AI adoption has been fragmented and unevenly distributed, with visibility highest in acute care and diagnostic imaging. No comprehensive national catalogue existed prior to this study, limiting coordinated investment, governance, and policy planning. The objective of this project was to conduct the first environmental scan of AI-driven clinical initiatives across Canadian healthcare, aligning with the theme of FHLIP 2026: 'From Silos to Synergy.'

Methodology
A volunteer working group conducted a structured environmental scan of publicly available AI initiatives across Canadian provinces and territories. Using a standardized taxonomy, 152 initiatives were coded for clinical function, venue, technology type, and deployment stage. Data was validated independently and analyzed via pivot tables to summarize patterns and signals.

Results
Of 152 initiatives identified, 43 were hospital-based and 30 in acute care. Machine learning was the most common technology type (67 initiatives). The majority (89) were pilots, with only 18 scaled initiatives. Key signals included workflow integration driving adoption, pronounced equity gaps with underrepresentation in primary care, long-term care, and Indigenous remote settings, thin peer-reviewed evidence, and emerging use of large language models and robotics.

Discussion & Conclusion
The study establishes a national baseline illustrating widespread fragmentation in Canadian AI adoption, underscoring the need for intentional policy and governance to promote equitable, evidence-based system-wide implementation [1,2]. Addressing equity gaps and enhancing outcome transparency are critical for advancing responsible AI integration in healthcare [3,4,5,6]"

[OR-5-16] Patient Perspectives on Use of Artificial Intelligence in Clinical Practice: A Narrative Review

Dimitri Oreopoulos, McMaster University
Sebastian, Mafeld, University of Toronto

"Patient perspectives on Artificial Intelligence (AI) use in healthcare have not been well studied. Yet, this understanding is crucial for ensuring these perspective are addressed in AI development and deployment in clinical settings.

This review synthesizes existing literature and identifies key themes regarding patient perspectives on AI. The electronic search strategy, which was designed in consultation with Western University's Allyn and Betty Taylor library, sourced 351 studies from five databases: Medline Proquest, Ovid Embase, the Cumulative Index to Nursing and Allied Health Literature (CINAHL), PubMed, and Google Scholar. Ultimately, 20 studies were included in this review, along with four additional sources from grey literature.

Key themes that emerged across these studies included: 1. the inability of AI to replace physicians, 2. elevated status of physician authority in high-stakes clinical situations, 3. lack of education or information relating to AI, 4. mistrust of AI, 5. the need for more transparent AI regulations. Importantly, patient concerns within these categories exhibited remarkable heterogeneity, which reinforces the need for flexible AI tools that address the diverse needs of their patients. Despite these concerns, patient consensus overwhelmingly favoured the inclusion of AI in healthcare as a tool for physician support. Consensus stemmed from patients' hope for the AI-supported clinician of the future to be the 'ideal physician'. This paper is intended to serve as a practical guide to aid healthcare policymakers' and practitioners' understanding of patient perspectives regarding AI in healthcare. Only once this understanding has been achieved can AI technologies truly reach their full potential."

[OR-3-17] Empowering Researchers: Generative AI Tools for Medical Data Analysis and Sharing

Daniel Hekman, Upside Lab

"Artificial intelligence is increasingly leveraged in healthcare, with substantial efforts focused on recognizing patterns in data. In recent years, however, the emergence of generative AI and, more recently, agentic AI has expanded the potential applications of AI into new fields. While these approaches may not be directly suited for patient-facing predictions or interventions, they offer significant opportunities to advance medical research.

In this presentation, we describe our work on using generative AI to enable researchers to more intuitively and efficiently explore complex medical datasets. We will introduce a novel tool that we are developing that facilitates data exploration, analysis, and preparation for sharing with collaborators and for publication. Using an example medical dataset, we will demonstrate how this tool can streamline workflows, promote collaboration, and enhance the reproducibility of research."

[OR-3-18] Assessing Diagnostic Accuracy of ChatGPT-4o and Gemini Pro in Classifying Skin Lesions

Rick Jaggi, University of Saskatchewan
Abdullah Qureshi, University of Saskatchewan
Yara Eita, University of Saskatchewan

"Background: General practitioners (GPs) often face challenges diagnosing dermatologic conditions due to limited specialist access. With growing interest in artificial intelligence (AI), large language models (LLMs) like ChatGPT and Gemini may offer support in clinical decision-making. This study evaluates their accuracy in classifying skin lesions as benign or malignant.

Methods: We tested ChatGPT-4o and Gemini Pro using 200 ISIC images (100 benign, 100 malignant). Each model assessed the same images using two prompt types: (1) open-ended ("What is the top differential diagnosis?") and (2) closed-ended ("Is this skin lesion (A) malignant or (B) benign?"). Sensitivity, specificity, and accuracy were calculated, and McNemar's test was used to assess intra- and inter-model performance.

Results: Gemini outperformed ChatGPT across both prompt types. With the open-ended prompt, Gemini showed 81.0% sensitivity and 67.7% accuracy, while ChatGPT showed 33.0% sensitivity and 49.3% accuracy. Closed-ended prompts increased Gemini's sensitivity to 92.0% but reduced specificity. ChatGPT's performance remained unchanged regardless of prompt type.

Conclusions: Gemini demonstrated higher diagnostic sensitivity but with a malignancy-seeking bias. ChatGPT showed more conservative predictions. While LLMs show potential, their current diagnostic reliability is limited, and will need rigorous evaluation protocols in future clinical applications."

[OR-3-19] The Write stuff: Ambient Listening and beyond: building AI capacity

Maxim Ben-Yakov, Humber River Health, Toronto

"The Write Stuff: Ambient Listening and beyond: building AI capacity,"" delves into the transformative potential of AI Scribe in healthcare documentation. This innovative tool leverages artificial intelligence to streamline patient interactions by automatically generating drafts of clinical notes through ambient listening technology. The reported benefits are substantial, including increased clinician efficiency, a reduction in documentation burden, and more accurate and complete medical records, ultimately leading to quicker and more precise medical coding.

A pilot program was conducted at Humber River Hospital, involving 26 users (24 physicians and 2 nurse practitioners) across nine departments. The focus of this pilot was to evaluate clinician satisfaction, documentation time, and patient satisfaction, while also exploring secondary benefits. The findings revealed promising outcomes in high-volume, structured environments like Cardiology and Emergency, where AI Scribe significantly reduced documentation time and improved focus. However, the tool showed limited value in more complex specialties such as Psychiatry and Oncology, where notes often required extensive editing, sometimes increasing the workload.

A major barrier identified was the lack of Electronic Medical Record (EMR) integration, which necessitated manual copy-pasting of notes and diminished the potential for time savings. Despite these challenges, patients were highly receptive to the use of AI Scribe, with a vast majority expressing comfort and reporting improved provider engagement. The presentation concludes with a recommendation for a ""Bring Your Own Scribe"" (BYOS) strategy, emphasizing the need for robust governance and risk management to ensure privacy, security, and clinical compliance as AI adoption accelerates within clinical practice. This strategy aims to meet immediate clinician demand while a long-term plan for enterprise-grade documentation tools is developed."

[OR-3-20] Integrating Equity, Diversity, and Inclusion in Generative AI Applications for Healthcare Education: A Scoping Review

Danielle Shin, Centre for Addiction and Mental Health, Toronto
Jisan Lee, Gangneung-Wonju National University, South Korea
Zerina Lokmic-Tomkins, Monash University, Australia
Sang Bin You, University of Pennsylvania, United States
Mollie Hobensack, Vanderbilt University, United States
Lisa Reid, Flinders University of South Australia, Australia
Charlene E. Ronquillo, University of British Columbia
Max Topaz, Columbia University, United States
Jiyoun Song, University of Pennsylvania, United States

"Introduction: Capabilities of generative AI (GenAI) tools continue to advance, positioning them as promising resources for supporting healthcare education. However, their implementation presents unknown implications related to Diversity, Equity, and Inclusion (DEI). In this review, DEI is used to represent any intentional effort to achieve equal opportunities and outcomes for all, including those from traditionally marginalized or underrepresented groups. To address the knowledge gap, we aim to synthesize the literature on how DEI considerations are discussed in GenAI applications.

Methods: Following the JBI scoping review guidelines, we searched CINAHL, Embase, and PubMed on October 20, 2024. We included all types of empirical studies published in English in any country. Specifically, this review examined how DEI is recognized, the strategies used for its integration, and the reported outcomes of these efforts. It also identified author-reported limitations and recommendations for integrating DEI in GenAI within healthcare education.

Results: Sixteen studies were included. Content generation for educational materials was the dominant use case of GenAI (n=12), while fewer studies explored interactive virtual patient simulations (n=3) and simulation-based, scenario-driven training (n=2). At the research design level, nine studies reported DEI integration strategies, such as inclusive recruitment, and reporting of socioeconomic status, and tailored data collection and interpretation for low-resource settings. At the GenAI lifecycle level, nine studies reported strategies such as bias mitigation, multilingual content generation, and randomized patient characteristics to prevent stereotyping. Notable findings included a lack of consistent conceptual clarity on how to integrate DEI, limited engagement with diverse partners in the development of GenAI, and a lack of evaluations examining how DEI is incorporated or addressed. Most studies instead evaluated aspects of GenAI performance.

Discussion and Conclusions: Although early efforts to incorporate DEI into GenAI for health education and training are evident, this integration remains limited and insufficiently evaluated. Findings from our study can enhance understanding of directions for future empirical inquiry, including improving transparency, encouraging the inclusion of diverse perspectives, and establishing frameworks that systematically integrate DEI in healthcare educational contexts."

[OR-3-21] Convolutional Neural Networks Outperform Automated Six-Point Morphometry for Vertebral Fracture Recognition in DXA Lateral Spine Images

Barret Monchka, University of Manitoba
John Schousboe, University of Manitoba
Douglas Kimelman, University of Manitoba
Lisa Lix, University of Manitoba
William Leslie, University of Manitoba

"Introduction: Prevalent vertebral fractures (VFs) strongly predict future osteoporotic fractures, making their timely detection critical for managing osteoporosis. Automated VF screening could enable earlier treatment of high-risk patients and may be achieved through (1) six-point morphometry, a rule-based software approach, or (2) machine learning (ML). We compared the clinical utility of these two approaches.

Methods: Dual-energy X-ray absorptiometry (DXA) images were obtained from the Manitoba Bone Mineral Density Registry (Feb. 2010 - Dec. 2017) and linked with administrative health records through Mar. 2018. Automated morphometry identified VFs based on vertebral body height reduction. Convolutional neural network (CNN) probabilities ≥ 0.5 were classified as VFs. Agreement with human experts was measured using prevalence- and bias-adjusted kappa (κ). Associations between baseline risk factors and identified VFs were estimated as odds ratios (ORs). Hazard ratios (HRs) for incident major osteoporotic fractures (MOFs) were estimated using Cox regression. 95% confidence intervals (CIs) were reported.

Results: Among 12,395 patients, 16.6% were diagnosed by experts with a VF at baseline. Predicted VF prevalence was 17.3% for CNNs and 70.2% for automated morphometry (51.0% when excluding grade 1 VFs). Agreement with experts was nearly perfect for the CNNs (κ=0.86), but weak for automated morphometry (κ=0.17 overall, κ=-0.14 when excluding grade 1 VFs). Baseline osteoporosis risk factors were strongly associated with CNN predictions but weakly associated with automated morphometry (Table). CNN-predicted VFs were strongly predictive of incident MOFs (HR=2.33, 95% CI: 2.03-2.68), similar to associations with expert diagnoses (HR=2.54, 95% CI: 2.21-2.92), while automated morphometry had significant but weaker associations (HR=1.49, 95% CI: 1.28-1.73 overall; HR=1.62, 95% CI: 1.42-1.85 when excluding grade 1 VFs).

Conclusion: ML is superior to automated morphometry at identifying VFs in DXA images and is more predictive of clinically relevant outcomes. These findings support recommendations discouraging the use of quantitative morphometry for VF ascertainment."

[OR-3-22] Practical Applications of Artificial Intelligence in Geriatric Medicine: A Scoping Review

Asghar Khan, Western University
Sai-Amrit Maharaj, University of Toronto
Kelly Kay, Provincial Geriatrics Leadership Ontario
Justin H.S. Kim, North York General Hospital, Toronto
Daniel Yacoub, North York General Hospital, Toronto
Nihal Haque, North York General Hospital, Toronto

"Target Audience: Clinicians, researchers, and policy-makers

Introduction: Advancements in artificial intelligence (AI) offer promise for addressing geriatric concerns such as falls, frailty, dementia, polypharmacy, and homecare provision, which significantly impact older adults (55+). This scoping review examined the current state of AI technology implementation in the care of older adults.

Methodology: Our search identified 1816 studies published between 2020 and 2023. Two reviewers screened titles and abstracts, followed by full-text review of 442 articles, with conflicts resolved by a third reviewer. Among the 169 articles meeting inclusion criteria, 97 corresponded to our research question of current and near-term clinical utility of AI across five key geriatric themes.

Results: AI technologies applied to geriatric concerns were categorized as: Prediction (51.5%), Classification (27.8%), Assistive Devices (11.3%), or Robot (9.3%). Predictive AI was seen in studies of falls (21.6%), dementia (15.5%), frailty (5.2%), homecare (4.1%), mild cognitive impairment (MCI) (3.1%) and delirium (2.1%). Classification using AI included two activities, detection of a condition or differentiation between individuals with or without a condition, and was used in dementia (17,5%), frailty (4.1%), polypharmacy (3.1%), and falls, delirium and MCI (1% respectively). Across all studies, only 9% reported an area-under-the-curve (AUC) value greater than 0.8, suggesting limited studies demonstrating clinical utility (Figure 1). Few studies were randomized-controlled trials (RCTs) (9.3%), Most studies focused on dementia (26.8%), or specifically Alzheimer's disease (17,5%), falls (24.7%) and frailty (10.3%). No studies compared AI performance with non-AI technologies and 75.2% of studies did not report the ethnicity of participants. One study reported on Black participants and no studies included Indigenous participants.

Conclusion: Most AI studies focused on older adults remain theoretical, lacking near-term clinical utility. Future research should emphasize RCTs, AUC reporting, comparison against established non-AI clinical tools and AI use among participants of different ethnicities to support real-world application."

[OR-3-23] Is Generative AI the new NVivo? A Systematic Review on the Validity and Reliability of Large Language Models for Qualitative Data Analysis

Theodore C.K. Cheung, The Hospital for Sick Children, Toronto
Amanda Seyler, University of Calgary
Nhu Huynh, The Hospital for Sick Children, Toronto
Roxy Helliker O'Rourke, The Hospital for Sick Children, Toronto
Christie Burton, The Hospital for Sick Children, Toronto
Russell Schachar, The Hospital for Sick Children, Toronto
Paul Arnold, University of Calgary
Jennifer Crosbie, The Hospital for Sick Children, Toronto

"Introduction: Qualitative research methods, including interviews and focus groups, provide rich, contextualized insights into human experience and complex social phenomena. Engaging closely with qualitative data is integral to interpretation; however, the growing scale of textual datasets presents new challenges and opportunities for analysis. Advancements in generative artificial intelligence, particularly large language models (LLMs), offer tools that may assist researchers in identifying patterns and organizing qualitative data while preserving interpretive depth. As interest in these applications expands, it is important to assess the validity and practical suitability of LLMs for qualitative analysis. This systematic review evaluated the extent to which LLMs, specifically generative transformer models such as ChatGPT, can identify themes and patterns consistent with human analysts.

Methods: PubMed, PsycINFO, Web of Science, Google Scholar, and the reference lists of included studies were searched up to June 2025 using pre-defined keywords. Eligible studies directly compared LLM- and human-generated outputs derived from thematic, narrative, and discourse analyses.

Results: A total of 39 studies met inclusion criteria. Most used versions of ChatGPT (70.8%), while others utilized Gemini (8.3%), Claude (4.2%), among others. Comparative analyses spanned diverse disciplines (e.g., psychology, public health, education, and health care, etc.), though standardized protocols for conducting LLM-assisted qualitative analysis were lacking. Overall, LLM performance relative to human analysts was rated as moderate to high. Namely, among those studies reporting quantitative indexes, the median % agreement score was 0.80 (range: 0.19-1.00), mean Cohen's Kappa was 0.74 (range 0.38-0.82), and median Cosine coefficient was 0.58 (range: 0.47-0.80), with stronger alignment for semantic or descriptive themes and lower accuracy for interpretive, context-dependent themes (see supporting information).

Discussion: Findings provide support for the use of LLMs as complementary tools in qualitative data analysis. Rather than replacing human interpretation, LLMs may serve as efficient aids for data organization and inter-rater comparison, supporting, but not substituting, the depth of qualitative engagement."

[OR-3-24] Machine-learning-assisted preoperative prediction of pediatric appendicitis severity

Aylin Erman, University of Ottawa
Julia Ferreira, McGill University Health Centre
Waseem Abu Ashour, McGill University Health Centre
Elena Guadagno, McGill University Health Centre
Etienne St-Louis, McGill University
Sherif Emil, McGill University
Jackie Cheung, McGill University
Dan Poenaru, McGill University

"Introduction
Acute appendicitis is the most common surgical emergency in children, and its severity determines therapeutic options and outcomes. This study investigates the use of machine learning (ML) algorithms for the accurate preoperative prediction of the severity of appendicitis.

Methods
Our dataset consisted of anonymized medical records of children undergoing emergency appendectomy between 2014-2021 at the Montreal Children's Hospital. The built ML pipeline imputed missing values, upsampled infrequent classes and predicted 5 appendicitis grades (1 - non-perforated, 2 - localized perforation, 3 - abscess, 4 - generalized peritonitis, and 5 - generalized peritonitis with abscess). The best combination of imputation strategy, class balancing technique and classification model was chosen based on validation performance. Model explainability was verified by a pediatric surgeon.

Results
The study included 1980 patients (60.6% males, average age 9.3 years). Grade of appendicitis in the cohort was as follows: grade 1 - 70%; grade 2 - 8%; grade 3 - 7%; grade 4 - 7%; grade 5 - 8%. Every combination of 6 imputation strategies, 7 class-balancing techniques, and 5 classification models was tested. The best-performing combined ML pipeline distinguished non-perforated from perforated appendicitis with 82.8 ± 0.2 % NPV and 56.4 ± 0.4 % PPV, and differentiated between severity grades with 70.1 ± 0.2 % accuracy and 0.77 ± 0.00 AUROC.

Discussion/ Conclusion
The key variables driving our model's predictions are consistent with established clinical knowledge and existing literature, indicating that the ML model successfully identified meaningful patterns within the data. Additionally, our model outperforms existing pediatric appendicitis prediction tools. Following external validation and silent clinical testing, this ML model has the potential to enable personalized severity-based treatment of pediatric appendicitis and optimize resource allocation for its management."

[OR-3-25] ECG-FM: A self-supervised foundation model for 12-lead ECG interpretation and rapid LVEF risk stratification in acute care

Sameer Masood, University Health Network, Toronto
Kaden McKeen, University Health Network, Toronto
Bo, Wang, University Health Network, Toronto
Barry Rubin, University Health Network, Toronto

"Target Audience
Emergency physicians, cardiologists, hospitalists, clinical informaticians, and ML researchers.

Introduction
Timely estimation of left ventricular ejection fraction (LVEF) can inform early decisions when echocardiography is delayed. We developed ECG-FM, a transformer-based model for automated ECG interpretation and LVEF prediction using routinely acquired 12-lead ECGs from tertiary-care hospitals. We aim to enable bedside risk stratification within seconds in emergency/inpatient settings and to prepare for clinical deployment at UHN.

Methods
Data comprised 12-lead ECGs from UHN-ECG (≈622k ECGs/211k patients, 2010–2018) plus public PhysioNet 2021 and MIMIC-IV-ECG for pretraining. ECGs were resampled to 500 Hz, z-score normalized, and segmented into non-overlapping 5-s windows. A wav2vec 2.0-style encoder (CNN feature extractor + transformer) was pretrained with masking, Contrastive Multi-Segment Coding (CMSC), and Random Lead Masking (RLM). For LVEF, labels were regex-extracted from echo reports; each ECG was paired to the closest echo within ±7 days. Splits were patient-temporal (≈80/10/10) to avoid leakage.

Results
On ECG interpretation, ECG-FM achieved high performance across clinically salient rhythms/diagnoses (e.g., atrial fibrillation AUROC 0.996, accuracy 0.978). For reduced LVEF, performance was strongest for severe dysfunction (≤30%): AUROC 0.918, AUPRC 0.674, accuracy 0.876, NPV 0.938; NPV exceeded 92% across LVEF thresholds (≤50%, ≤40%, ≤35%, ≤30%). These results were obtained on a large, representative in-hospital cohort.

Discussion/Conclusion
ECG-FM demonstrates strong ECG interpretation and clinically meaningful LVEF prediction, supporting early decision-making when echo is unavailable or delayed. We are extending the model by integrating additional clinical variables and diagnostic data and advancing toward pragmatic clinical deployment at UHN to deliver actionable, bedside estimates rather than delayed results. Future work will evaluate workflow integration, prospective performance, and impact on downstream care."

[OR-3-26] A Deep Learning Screening Model for Accurate Traumatic Brain Injury Triage

Julia Wiercigroch, University of Toronto
Christopher Smith, St. Michael's Hospital, Toronto
Christopher Witiw, St. Michael's Hospital, Toronto

"Background: The Automated Surgical Intervention Support Tool for Traumatic Brain Injury (ASIST-TBI) is an AI-based clinical decision-support tool designed to identify traumatic brain injury (TBI) requiring neurosurgical intervention from head CT imaging alone. During silent deployment in an emergency department (ED), ASIST-TBI encountered pathologies beyond isolated TBI, including non-traumatic and post-operative scans. These heterogeneous cases reduced reliability and revealed the need to automatically distinguish between traumatic and non-traumatic images. Our goal was to incorporate additional clinical context into ASIST TBI's workflow to improve its interpretability and reliability in real-world practice.

Methods: We curated 1,995 non-contrast head CTs from ED patients. Each scan was labeled for hemorrhagic findings, non-hemorrhagic pathology, post-operative changes, and artifacts. Three specialized models were trained: (1) a 2.5D Convolutional Neural Network (CNN) for artifact detection, (2) a 2D CNN using maximal-intensity projections for post-operative changes, and (3) a pretrained CNN for hemorrhage segmentation with an added XGBoost classification head to distinguish hemorrhagic from non-hemorrhagic pathology. Individual model performance was evaluated, and the integrated pipeline was applied to 148 low-confidence scans (prediction probability 0.2–0.8) from the ASIST TBI deployment.

Results: The pipeline improved the interpretability of ASIST TBI's predictions. Individual models achieved strong performance (e.g., AUROC = 0.97 for artifact detection). In all true positive cases, the pipeline correctly identified hemorrhages consistent with ASIST TBI's neurosurgical decisions. For true negatives, 75.2% were non-hemorrhagic, artifacts, or post-operative changes. Among false positives, a non-hemorrhagic cause was identified in 95.5% of cases, suggesting these scans were flagged incorrectly. For false negatives, where ASIST-TBI missed a hemorrhage, the non-hemorrhagic pathology model detected the bleed in 80.0% of cases. These results demonstrate that added clinical context makes model predictions more interpretable and trustworthy.

Conclusion: By analyzing ASIST TBI's lower confidence predictions, our multi-model pipeline addresses the challenges revealed during ED deployment. It has the potential to explain model failures, flag false positives, and recover missed findings. Integrating this pipeline into existing clinical workflows demonstrates a practical approach to improving real-world performance and safety of AI-based decision support tools in complex care environments."

[OR-3-27] A Process-Driven Governance Model for Artificial Intelligence Software Not Classified as Medical Device in Canada

Remziye Zaim, University of Toronto
Victoria Chui, University of Toronto
Shion Guha, University of Toronto
Ibukun-Oluwa Omolade Abejirinde, University of Toronto
Ijeoma Uchenna Itanyi, University of Toronto
Lorraine Lipscombe, University of Toronto
Jennifer L. Gibson, University of Toronto
Laura C. Rosella, University of Toronto
James A. Shaw, University of Toronto

"The governance of Artificial Intelligence (AI) software not classified as a medical device remains an under-addressed challenge in health systems worldwide. AI systems that fall outside formal Software as Medical Device (SaMD) regulatory pathways often evade comprehensive oversight, raising concerns around data privacy, safety, equity, and trust, and introducing new public risks.

This study presents a process-driven governance model designed to guide the responsible implementation of AI for population-level disease risk prediction and prevention. Drawing on the Quintuple Aim —enhancing patient and provider experiences, improving population health outcomes, promoting equity, and ensuring system sustainability— the model addresses the fragmented and ad hoc oversight currently characterizing non-SaMD AI tools.

We developed the Responsible AI for Health Systems (RAIHS) framework, which integrates participatory governance approaches to engage policymakers, healthcare providers, patients, local agencies, and community members in decision-making. The framework embeds ethical and social considerations throughout the AI lifecycle—from problem identification and procurement to deployment, monitoring, and evaluation. Using a case study in Peel Region, Ontario, we demonstrate how the RAIHS framework can be applied to the deployment and implementation of AI for Type 2 diabetes (T2D) prediction and prevention. The case illustrates how community engagement and attention to sociodemographic complexity can strengthen trust and ensure equitable outcomes.

The RAIHS framework offers a scalable, values-based approach to governing AI systems responsibly in public health. By embedding ethics, transparency, and community input at every stage, it promotes trust in AI, enhances health outcomes for populations at risk, and builds the adaptive capacity of local health systems. This work underscores the urgent need for standardized, context-sensitive governance models that can be applied across jurisdictions to mitigate unintended risks and maximize the societal benefits of AI in health.

This study was funded by the Canadian Institute for Advanced Research (CIFAR), as part of the Artificial Intelligence for Health Solution Networks Project, 2023-2026."

[OR-3-28] Assessing reconstruction fidelity in conditional models for 12-lead electrocardiogram synthesis from limited leads

Christopher Cheung, Sunnybrook Health Sciences Centre, Toronto
Alex Mariakakis, University of Toronto
Mithun Manivannan, Carleton University

"Introduction:
Reconstructing a diagnostically complete 12-lead electrocardiogram from a limited subset could broaden access to cardiac assessment in settings without full multilead acquisition. However, previous ECG reconstruction methods often suffer from regression to the mean, where generic models produce average morphologies that fail to capture patient-specific abnormalities, particularly in pathological conditions like atrial fibrillation or myocardial infarction. We implemented and evaluated a conditional hierarchical variational autoencoder that synthesizes eight missing leads from four inputs (I, II, V3, V5), with remaining leads derived algebraically, using pathology-aware conditioning to overcome this limitation.

Methods: The model reproduces the reference architecture with a multi-level latent Gaussian hierarchy and a channel-conditioned discretized logistic mixture likelihood. Conditioning used synthetic pathology embeddings obtained from lead pattern heuristics. Training used 17,418 10 sec 500 Hz 12 lead recordings from PTB XL with per lead min max scaling, double precision arithmetic, KL annealing with balance coefficients, and numerically stable objective evaluation.
Results: The implementation achieved stable training with finite losses across all experiments, demonstrating architectural feasibility, as evidenced by the monotonic reduction of training and validation loss in (Fig. 1A). However, reconstruction quality remained low, with a mean per-lead MSE of 0.15 ± 0.05 and spectral correlation of 0.01 ± 0.02, consistent with the lead-wise error bars and near-zero correlations (Fig. 1B) and the overall validation MSE of 0.189 (Fig. 1C), indicating that synthetic label conditioning provides insufficient signal for effective reconstruction of unobserved channels. Qualitatively, observed inputs retained plausible baselines while unobserved channels lacked patient-specific morphology and frequency content.

Discussion: Preliminary findings show numerical stability but cross-lead fidelity below clinical acceptability. Synthetic pathology embeddings conveyed insufficient mutual information, promoting regression toward average morphology. We therefore hypothesize that a generic model with cohort-aware conditioning can outperform patient-specific fitting, although this was not evaluated. Future work will test this in electrophysiology ablation cohorts by incorporating intra-procedural signals, catheter coordinates, and pacing epochs. This cohort-specific conditioning may enable clinically reliable, limited lead reconstruction."

[OR-3-29] Artificial Intelligence-Powered Wound Assessment at the Point of Care: Enhancing Equity Across Skin Tones

Tahirih Nasseri, University of Toronto
Samantha Bestavros, University of Toronto
Rachel Tyli, University of Toronto
Robert D Fraser, University of Western Ontario
Sheila C Wang, Women's College Hospital, Toronto

"Target Audience:
Clinicians, researchers, and healthcare innovators interested in practical, efficient, and equitable wound assessment and AI-supported decision-making at the bedside.

Introduction:
Wound assessment in clinical practice often relies on subjective visual inspection and measurement, which is less reliable in darker skin tones and contributes to inequities in diagnosis and management. Many artificial intelligence (AI)-based wound evaluation tools are limited by non-representative datasets dominated by lighter skin tones. We present a multimodal, AI-powered wound assessment platform designed for real-time use at the point of care to provide accurate and equitable evaluation across all skin tones.

Methods:
The AI algorithm was trained on 465,187 scientifically calibrated wound images spanning the Fitzpatrick skin tone spectrum. The system integrates visible light, fluorescence, and infrared thermography (IRT) to automatically segment and classify wound tissue (granulation, slough, eschar, epithelialization), detect infection, and assess inflammation. Validation included clinician comparisons (n=59) and sub-studies examining (1) wound tissue classification, (2) infection detection using bacterial fluorescence and IRT, and (3) inflammation assessment in hidradenitis suppurativa through IRT temperature symmetry and clinical severity scoring.

Results:
AI predictions showed strong concordance with clinician ratings, with 91% of cases rated "very good" to "fair."1 Infection detection achieved 74% overall accuracy and 100% sensitivity, accurately distinguishing infected, inflamed, and non-inflamed wounds.2 IRT-based inflammation assessment correlated strongly with clinical severity (Spearman's $\rho = 0.65–0.84$) and achieved near-perfect interrater reliability (0.98), identifying subclinical inflammation invisible on darker skin tones (Fitzpatrick V–VI).3

Discussion/Conclusion:
This AI-driven multimodal imaging platform provides clinicians with a reliable, objective, and equitable point-of-care wound assessment tool. By reducing inter-rater variability and visual bias, it enhances accuracy in tissue classification, infection detection, and inflammation monitoring. Its integration into bedside workflows could support earlier recognition of complications, more consistent treatment decisions, and improved wound care outcomes for patients of all skin tones."

[OR-3-30] Bias to Better: A Stepwise Checklist for Equitable Child & Youth MH AI

Adan Amer, University of Toronto
Nicholas Mitsakaki, Children's Hospital of Eastern Ontario (CHEO) Research Institute, Ottawa
Kathleen Pajer, Children's Hospital of Eastern Ontario (CHEO) Research Institute, Ottawa
Christina Honeywell, Children's Hospital of Eastern Ontario (CHEO) Research Institute, Ottawa

"Introduction: Artificial intelligence (AI) can streamline paediatric mental-health care but may also introduce or magnify inequities when data, design choices, or deployment contexts do not reflect real clinical populations. Existing "fairness" guidance is often high level and leaves clinicians and researchers without concrete steps to detect and mitigate bias across the project life cycle. Our work develops an actionable, stage-specific toolkit to operationalize equity considerations for AI/ML applied to child and youth mental health.

Methods: We conducted a targeted narrative review on bias types, fairness metrics, and mitigation strategies that span all distinguishable stages of the AI/ML development lifecycle: conception, data collection, data pre-processing, model evaluation, clinical deployment, and monitoring. Findings were iteratively synthesized into a stage-specific checklist. Co-design input came from patient and family stakeholders at the Children's Hospital of Eastern Ontario (CHEO) Research Institute and from clinician-researchers and AI developers within the Precision Child and Youth Mental Health (PCYMH) Collaboratory. Development was aligned with a concurrent institutional roadmap for integrating AI/ML in clinical research to ensure fit with workflows and governance.

Results: The resulting checklist translates fairness principles into concrete prompts and decision points for each lifecycle stage (e.g., inclusion criteria, label quality, subgroup performance, distributional shift, harm modelling, and post-deployment monitoring). It pairs each prompt with suggested actions (e.g., stratified error analysis, data augmentation options, alternative thresholds, stakeholder validation steps) and links decisions to potential equity impacts. Co-design sessions surfaced usability requirements (plain-language items, minimal overhead, case examples) that were incorporated into subsequent versions. The checklist is prepared to be piloted as an interactive Microsoft Copilot agent that guides teams, records decisions, and surfaces resources.

Discussion: This work operationalizes bias mitigation for paediatric mental health AI by embedding concrete checks into routine research and implementation steps. By making equity considerations explicit and auditable, the tool supports safer and more trustworthy models and fosters shared accountability among clinicians, data scientists, and family partners. Next steps include a pilot to evaluate usability, decision completeness, and effects on subgroup performance monitoring, followed by iterative refinements and open dissemination."

[OR-3-31] Prompt-engineering in Large Language Models (LLMs) to detect emotions from statements made by patient/caregivers in online networks

Muhammad Haseeb Aslam, McMaster University
Ashirbani, Saha, McMaster University

"Target Audience: Healthcare AI researchers, clinicians, and patients

Introduction: To determine the effective combination of pre-trained Large Language Model (LLM) architecture with prompting, using different hyperparameters and conditions (input curation), for emotion recognition in statements made by cancer patients/caregivers in Cancer Survivors Network (CSN).

Methods: We used a publicly available dataset (CANCEREMO, Sosea and Caragea., EMNLP, 2020) containing 19,917 sentences compiled from CSN, annotated for Plutchik's eight basic emotions (Trust, Surprise, Fear, Anger, Anticipation, Sadness, Joy, and Disgust), and categorized for training, validation, and testing. After removing duplicates (9,120) and sentences with irregular characters (1,193), 5,704 sentences expressing only single emotions were considered.

Our study involved five widely studied LLMs, GPT-4o-mini, Gemma-2-2B-IT, LLaMA-3.2-3B-Instruct, Falcon-7B-Instruct, and Mistral-7B-Instruct-v0.2 and two recent ones: GPT-5-mini and GPT-oss-20b, under both zero-shot and few-shot conditions. The experiment followed three steps: (a) selection of 10 random sentences (from the training subset) per emotion as learning examples (few-shot conditions), (b) evaluation using 10 random sentences (from the validation subset) to select top-performing models under relevant hyperparameter sets: first-five LLMs {15 sets from temperature (0.1-1.5) and top-p (0.90-1.0)}, GPT-5-mini {9 sets from verbosity and reasoning (low, medium, high)}, and GPT-oss-20b {7 sets from temperature (0.1-1.5), top-p (0.90-1.0), repetition penalty (1.0-1.2) and max_new_tokens (2048, 3072, 4096)} and (c) final testing on 150 random samples (from the test subset). Multi-class accuracy was used as the evaluation metric.

Results: The top-three models in the final test were Falcon-7B-Instruct (temperature: 0.1, top-p: 0.95), GPT-5-mini (reasoning: medium, verbosity: low), and GPT-oss-20b (temperature: 0.3, top-p: 0.9, repetition_penalty: 1.0, max_new_tokens: 3072) with accuracies of 78.67% (95%CI: 72%-85.3%), 74% (95%CI: 66.7%-80.7%) and 69.33% (95%CI: 62%-77.3%), respectively, all under few-shot conditions.

Conclusions: Prompt engineering improved emotion detection accuracy beyond the reported 73% in the literature, which used fine-tuned models (i.e., weight-altered models). Notably, the best results came from an older LLM (Falcon-7B-Instruct). Future work will focus on robust benchmarks to identify optimal conditions for detecting single/multiple emotions in cancer patient's statements."

[OR-3-32] Medevi-RoB: A Cross-Domain Benchmark of Large Language Models for Automated Risk of Bias 2 (RoB 2) Assessment in Randomized Trials

David Nguyen, McMaster University
Andrew Ganea, Queen's University
Dylan Kwan, McMaster University
Nathan Lai, Western University
Minh Hoang Ton, Queen's University
Henry Liu, McMaster University
Jessica Song, McMaster University
Sarah Pimenta, McMaster University
Michael Nguyen, George Mason University, USA
Dena Zeraatkar, McMaster University

"Background: Manual RoB 2 appraisal is essential for trustworthy evidence synthesis but remains slow and expertise-intensive. Large language models (LLMs) could help automate structured critical appraisal, yet their performance on RoB 2, across specialties and outcomes, has not been comprehensively evaluated.

Objective: To benchmark state-of-the-art LLMs on RoB 2 assessments of randomized controlled trials (RCTs), comparing model outputs with consensus ratings from high-quality systematic reviews, and to explore factors influencing accuracy, consistency, and potential biases.

Methods: We conducted a cross-sectional diagnostic accuracy study using Medevi, an open, reproducible platform for AI-assisted evidence synthesis. Randomized trials were sampled from Cochrane reviews containing validated RoB 2 assessments. Multiple LLM families (reasoning-optimized, general-purpose, and resource-efficient) were tested across structured prompting strategies ranging from zero-shot to per-sub-domain logic. Outputs were compared with expert consensus ratings. Primary endpoints were agreement metrics between model and reviewer assessments, with secondary analyses exploring model confidence, systematic biases, and cost-effectiveness.

Results: Early analyses suggest promising agreement between LLM-generated and human RoB 2 assessments across domains and prompting approaches. Structured, rule-based aggregation methods yielded the most stable and interpretable outputs. Quantitative findings will be finalized and presented at the conference.

Conclusions: Preliminary findings indicate that LLMs, when guided by structured prompts and transparent algorithms, can feasibly support semi-automated RoB 2 appraisal at scale. The Medevi-RoB benchmark provides a reproducible framework to evaluate and refine AI integration into critical appraisal workflows, with planned extensions to ROBINS-I, QUADAS-C, and AMSTAR."

[OR-3-33] Guidelines for Understanding and Implementing Documentation Education with AI (GUIDE-AI) – A Multimodal Study Assessing Canadian Medical Trainee Perspectives on AI Medical Scribes

Nima Toussi, University of Saskatchewan
Rajesh Girdhari, University of Toronto
Chris Gilchrist, University of Toronto
Sharon Domb, University of Toronto
Cody Jackson, University of Toronto
Azadeh Moaveni, University of Toronto
Noah Crampton, University of Toronto

"INTRODUCTION
Artificial Intelligence (AI) scribes, software that utilizes natural language processing to transcribe and summarize patient-physician interactions, have experienced rapid uptake among Canadian physicians. While the practicality of these tools has been demonstrated in practice and their uptake encouraged through public subsidies, there is a significant gap regarding how medical trainees assess these tools. This study serves as an exploratory analysis of medical trainee perspectives on AI scribes.

METHODS
This study employed a mixed-methods needs assessment consisting of an online survey and optional post-survey interviews. All medical students and residents (PGY1-3) at English-speaking Faculties of Medicine were eligible. The five-part survey gathered data on demographics, exposure to AI scribes, perceived benefits and concerns, and educational priorities related to AI Scribes. Post-survey interviews were semi-structured and explored themes of professional identity, perceived risks and benefits, and educational needs in the context of emerging AI tools in medicine. Research Electronic Data Capture (REDCap) was utilized to collect survey responses. REB approval was received (UToronto #00048659) prior to recruitment.

RESULTS
Complete responses were obtained from 110 medical trainees (27 pre-clerkship, 27 clerkship, 56 residents), with 10 follow-up interviews conducted to date. 74.5% were at least moderately familiar with the concept of AI scribes, only 30.0% had used them directly. 62.7% had observed their use, most commonly in Family Medicine (73.5%). Upon Analysis of Covariance, higher stage of training was found to be significantly correlated with greater concern with AI Scribes inducing errors in medical documentation ($p = 0.18$).

Preliminary thematic analysis of interview data revealed four overarching themes: promise of efficiency versus skill atrophy, privacy and accuracy as foundational concerns, urgency of formal AI education, and AI for collaboration over replacement. Trainees expressed cautious optimism, emphasizing enthusiasm for administrative relief and apprehension about privacy and liability.

CONCLUSIONS
Canadian medical trainees endorse commonly held benefits of AI Scribes. More senior medical trainees were more likely to hold reservations — underscoring the risks of competency erosion and privacy. These findings highlight an urgent need for experiential, ethics-informed AI curricula that prepare trainees to integrate documentation technologies safely and effectively."

[OR-3-34] Methodological Quality and Reporting Standards of AI Models for Predicting Clinical Deterioration: A Systematic Review Using the APPRAISE-AI Framework

Eric Zhang, University of Toronto
Daniel Kwan, University of Toronto
Jill Shah, University of Toronto
Sara Pollonen, University of Toronto
Kalia Kamini, Trillium Health Partners, Toronto

"Background: Artificial intelligence (AI) models are increasingly used to predict clinical deterioration in hospitalized adults, aiming to prompt earlier interventions and reduce preventable morbidity and mortality. Despite promising discriminative performance, concerns persist regarding methodological rigor, risk of bias, and inconsistent reporting standards. The APPRAISE-AI framework provides a structured method for evaluating the quality, transparency, and reproducibility of clinical AI studies.

Objectives: To synthesize the performance metrics of AI-and machine learning-based models predicting clinical deterioration among adult inpatients and to assess their methodological quality and risk of bias using the APPRAISE-AI framework.

Methods: This systematic review was prospectively registered in PROSPERO (CRD420251091119). Comprehensive searches of MEDLINE, Embase, Scopus, and IEEE Xplore identified studies evaluating AI models for outcomes such as ICU transfer, cardiac arrest, and in-hospital mortality. After screening 1,388 records and assessing 125 full-texts, 37 studies were included. Data extraction has been completed, and structured evidence tables are being finalized. APPRAISE-AI quality assessment is ongoing, with 10 of 37 studies appraised to date, focusing on domains of data quality, methodological conduct, robustness, and reproducibility.

Results: The 37 eligible studies (2013–2025) represent over 3.2 million inpatient encounters across North America, Europe, and East Asia. Most were retrospective external validations or pragmatic cohort evaluations of deployed AI tools on medical–surgical wards. Common algorithms included gradient-boosted trees and recurrent neural networks. Models such as eCART, CHARTwatch, and Epic Deterioration Index were frequently evaluated. Reported AUROC values ranged from 0.72–0.93, with top performance in temporal learning models (eCARTv5 AUROC 0.83–0.90; HAVEN 0.90). Preliminary APPRAISE-AI assessments indicate strong scores for clinical relevance, data quality, and reporting quality, but weaker methodological conduct, robustness of results, and reproducibility.

Conclusion: AI-based early warning systems demonstrate strong discriminatory performance but variable methodological rigor. Applying APPRAISE-AI highlights persistent gaps in reproducibility and transparency, emphasizing the need for standardized evaluation to ensure trustworthy clinical integration."

POSTER ABSTRACTS

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

[P01] For full abstract please see [OR-10-01]

[P02] Responsible Adoption of Multimodal AI in Healthcare: Promises and Challenges

Ghazal Azarfar, University Health Network

"Introduction: Clinicians rely on various data modalities—such as patient history, clinical signs, imaging, and lab results—to enhance decision-making. Multimodal artificial intelligence (AI) systems are emerging as powerful allies to process these diverse data types, but their clinical adoption faces challenges due to data heterogeneity and integration complexities.

Methods: The 2024 Temerty Centre for AI Research and Education in Medicine (T-CAIREM) symposium explored the potential and challenges of implementing multimodal AI in healthcare. This paper summarizes insights from the symposium.

Results: This study discusses current applications like early sepsis diagnosis and cardiology and identifies key barriers, including fusion techniques, model selection, generalization, fairness, safety, security, and international consideration. We provide practical strategies to overcome these obstacles, emphasizing technologies like federated learning to reduce biases and promote equitable healthcare.

Conclusion: By addressing these challenges, multimodal AI can transform clinical practice and improve patient outcomes worldwide."

[P03] For full abstract please see [OR-10-04]

[P04] For full abstract please see [OR-5-05]

[P05] For full abstract please see [OR-5-06]

[P06] For full abstract please see [OR-5-07]

[P07] For full abstract please see [OR-5-08]

[P08] For full abstract please see [OR-5-09]

[P09] For full abstract please see [OR-5-11]

[P10] For full abstract please see [OR-5-13]

[P11] For full abstract please see [OR-5-14]

[P12] For full abstract please see [OR-5-15]

[P13] For full abstract please see [OR-3-16]

[P14] For full abstract please see [OR-3-17]

[P15] For full abstract please see [OR-3-18]

[P16] For full abstract please see [OR-3-19]

[P17] For full abstract please see [OR-3-20]

[P18] For full abstract please see [OR-3-21]

[P19] For full abstract please see [OR-3-22]

[P20] For full abstract please see [OR-3-23]

[P21] For full abstract please see [OR-3-24]

[P22] For full abstract please see [OR-3-25]

[P23] For full abstract please see [OR-3-26]

[P24] For full abstract please see [OR-3-28]

[P25] For full abstract please see [OR-3-29]

[P26] For full abstract please see [OR-3-30]

[P27] For full abstract please see [OR-3-32]

[P28] Large Language Model-Assisted Critical Appraisal for Medical Education Research

David Nguyen, Michael G. DeGroote School of Medicine, McMaster University

Andrew Ganea, Faculty of Health Sciences, Queen's University, Kingston, Ontario, Canada

Minh Hoang Ton, School of Computing, Queen's University, Kingston, Ontario, Canada

Matthew Sibbald, Michael G. DeGroote School of Medicine, McMaster University

"Background: Systematic reviews are critical to evidence-based medical education but remain resource-intensive. Existing automation tools are limited in scope, flexibility, or transparency. Medevi, an open-source platform leveraging large language models (LLMs), aims to provide transparent, customizable, and cost-effective machine-assisted data extraction. This pilot study assessed Medevi's performance in medical education research, a domain with few validated benchmarks.

Methods: We evaluated Medevi using 52 peer-reviewed medical education studies with existing Medical Education Research Study Quality Instrument (MERSQI) and Newcastle-Ottawa Scale–Education (NOS-E) scores from a prior systematic review. Medevi outputs were compared against human-rater ground truths. Agreement was quantified using intraclass correlation coefficients (ICC) and mean absolute error (MAE). Efficiency was benchmarked by comparing time and cost per study against conventional human data extraction.

Results: Medevi achieved substantial agreement with human raters for MERSQI (ICC = 0.617, MAE = 0.904) and fair agreement for NOS-E (ICC = 0.365, MAE = 0.962). Domain-level agreement ranged from 46.2%-98.1% (MERSQI) and 32.7%-86.5% (NOS-E), with lower performance in more subjective areas. Compared to estimates for manual extraction (65-130 minutes, $21-43 per study), Medevi reduced appraisal time to <1 minute and cost to $0.10-0.57, yielding over 97% in cost savings.

Discussion: Medevi demonstrates promise as a transparent, open-source tool for accelerating systematic reviews in medical education. While best suited to objective criteria, it highlights the value of human oversight for subjective judgments. By reducing cost and time barriers, Medevi may help democratize high-quality evidence synthesis in medical education."

[P29] For full abstract please see [OR-3-33]

[P30] For full abstract please see [OR-3-34]

[P31] Responsible Adoption of Multimodal AI in Healthcare: Promises and Challenges

Ghazal Azarfar, University Health Network

"Introduction: Clinicians rely on various data modalities—such as patient history, clinical signs, imaging, and lab results—to enhance decision-making. Multimodal artificial intelligence (AI) systems are emerging as powerful allies to process these diverse data types, but their clinical adoption faces challenges due to data heterogeneity and integration complexities.

Methods: The 2024 Temerty Centre for AI Research and Education in Medicine (T-CAIREM) symposium explored the potential and challenges of implementing multimodal AI in healthcare. This paper summarizes insights from the symposium.

Results: This study discusses current applications like early sepsis diagnosis and cardiology and identifies key barriers, including fusion techniques, model selection, generalization, fairness, safety, security, and international consideration. We provide practical strategies to overcome these obstacles, emphasizing technologies like federated learning to reduce biases and promote equitable healthcare.

Conclusion: By addressing these challenges, multimodal AI can transform clinical practice and improve patient outcomes worldwide."

[P32] Domain-specific Pretrained Encoder Transformers for the Identification of Methodologically Rigorous Systematic Reviews: A Retrospective Modeling Study

Fangwen Zhou, Faculty of Health Sciences, McMaster University
Muhammad Afzal, Birmingham City University, Birmingham, United Kingdom
Rick Parrish, Faculty of Health Sciences, McMaster University
Ashirbani Saha, Department of Oncology, Faculty of Health Sciences, McMaster University
Wael Abdelkader, Faculty of Health Sciences, McMaster University
R. Brian Haynes, Faculty of Health Sciences, McMaster University
Alfonso Iorio, Department of Medicine, Faculty of Health Sciences, McMaster University
Cynthia Lokker, Faculty of Health Sciences, McMaster University

Introduction
Systematic reviews are considered one of the strongest levels of evidence, providing essential information for clinical guidelines and bedside practice. However, the broad spectrum of review articles complicates the identification of high-quality systematic reviews. This study aims to fine-tune and evaluate pretrained encoder transformers to identify high-quality systematic reviews from review articles using a large, reputable dataset.

Methods
Review articles from McMaster's Premium LiteratUre Service (PLUS) were retrieved. Articles were considered rigorous if they were systematic reviews that 1) stated the clinical topic, 2) described methods (databases and inclusion criteria), 3) searched ≥1 major database, 4) reported numbers retrieved/reviewed/included, and 5) did not exclude RCTs for treatment, primary prevention, quality improvement, or economics reviews; included "inception cohort" for prognosis reviews. Articles from 2003 to 2023 were randomly split 80:10:10 into training, validation, and testing sets. Articles in 2024 were used for external testing. A grid search of 630 configurations was conducted. Titles and abstracts were used as inputs. The model that achieved the lowest log loss on the validation set was further evaluated. A threshold of ≥0.50 was used for classification. Bootstrapping of 1,000 iterations was used to estimate 95% confidence intervals.

Results
BioELECTRA with no class weight adjustments, learning rate 5E-5, batch size 64, and seed 2 had the lowest validation log loss of 0.1840. Table 1 details the characteristics of the datasets and the model's performance, achieving >94% area under the receiver operating characteristic curve and >95% sensitivity. In articles published in 2024, there was a notable drop in specificity and mild degradation in other metrics.

Conclusion
Pretrained encoder transformer models demonstrate robust performance in identifying rigorous systematic reviews based on PLUS's criteria. These models can streamline the identification of high-quality evidence, reducing manual effort and enhancing the efficiency of evidence curation. Performance degradation in articles in 2024 may be due to changes in article structure and content over time. Future efforts should establish standardized benchmarking datasets for systematic reviews, develop models for tools such as AMSTAR 2, and account for temporal data drift to better support knowledge translation."

[P33] Interpretable methodological rigour appraisal of randomized controlled trials with deep learning transformers and large language models

Fangwen Zhou, Faculty of Health Sciences, McMaster University
Muhammad Afzal, Birmingham City University, Birmingham, United Kingdom
Rick Parrish, Faculty of Health Sciences, McMaster University
Ashirbani Saha, Department of Oncology, Faculty of Health Sciences, McMaster University
Wael Abdelkader, Faculty of Health Sciences, McMaster University
R. Brian Haynes, Faculty of Health Sciences, McMaster University
Alfonso Iorio, Department of Medicine, Faculty of Health Sciences, McMaster University
Cynthia Lokker, Faculty of Health Sciences, McMaster University

Introduction
The automation of clinical literature appraisal has garnered wide attention as an increasing number of articles are being published every year. Deep learning transformer models, including BERT and GPT, for classifying published clinical literature to support critical appraisal have demonstrated potential. However, due to the complexity of language models, the lack of transparency remains an important concern. This work explored the performance of BERT models with SHapley Additive exPlanations (SHAP) for explanation and GPT-4o in automating the methodological appraisal of randomized controlled trials (RCTs) for McMaster's Premium LiteratUre Service (PLUS), a gold-standard knowledge translation database.

Methods
Classifiers were trained with titles and abstracts of original articles using a grid search of 630 configurations based on PLUS's criteria for rigour (randomization, ≥10 participants per group, ≥80% follow-up, clinically important outcomes, and preplanned subgroup analyses if applicable). Models used 53,219 PLUS abstracts (2003–2023), split 80:10:10 into train/validate/test sets, with Clinical Hedges and PLUS 2024 as external tests. The SHAP partition explainer determined important tokens (words/subwords) for articles. The mean SHAP values for the most impactful words with ≥100 occurrences were examined. Full-text of a random sample of 800 articles from the validation and test sets was appraised using GPT-4o. GPT was prompted with the role, the task, PLUS's criteria, and the article instance. The temperature was set to 0 for reproducibility.

Results
The top-performing BioLinkBERT model achieved an AUROC of 89% to 95% on the validation and test sets. The top 15 most impactful unique words for positive (rigorous) and negative (non-rigorous) classes are found in Figure 1 and generally align with the manual appraisal criteria. On the subset of 800 articles, GPT-4o demonstrated comparable performance with a Matthews correlation coefficient of 0.429 compared with BioLinkBERT (0.466). GPT-4o provided transparent criterion-specific justifications.

Conclusions
Encoder transformers with explanations enable interpretable and scalable appraisal of RCTs, while generative LLMs achieve comparable zero-shot performance without explicit fine-tuning. SHAP enhances transparency by revealing influential linguistic features, while GPT-4o can provide plain-text justifications. These approaches advance trustworthy, automated critical appraisal for evidence synthesis and knowledge translation workflows.

[P35] Using Ensemble Deep Learning for Exercise Stress Echocardiography Quality Assessment

Khashayar Namdar, University of Toronto/Dalhousie University
Kenneth D'Souza, New Brunswick Heart Centre, University of New Brunswick
Suroush Bastani, Canadian Rivers Institute, University of New Brunswick
Jean-François Légaré, New Brunswick Heart Centre, IMPART Investigator Team
Leo Anthony Celi, MIT, BIDMC, Harvard T.H. Chan School of Public Health
Akram Mikaeilpour, New Brunswick Heart Centre, IMPART Investigator Team
Keith Brunt, New Brunswick Heart Centre, IMPART Investigator Team

Introduction
Stress echocardiography is a widely used functional ischemia testing, yet its diagnostic yield is limited mainly due to constraints in image acquisition and poor image quality. Building on prior work in automated view identification, we propose image quality indices (QI) to assess echocardiographic datasets systematically.

Methods
We curated a balanced dataset of 762 stress echocardiographic video Digital Imaging and Communications in Medicine (DICOM) files (rest and post-exercise) with a total of 17,072 frames from 63 patients, comprising five standard views. An ensemble of ten deep learning models was trained for view identification, including CapsNet, a custom shallow convolutional neural network (CNN), ResNet18/34/50/101/152 pretrained on ImageNet, and ResNet50, InceptionV3, and DenseNet121 pretrained on RadImageNet. Training was limited to 20 epochs, with unified five-fold patient-level stratification. QI metrics included: (i) mean accuracy across models/frames, (ii) uncertainty (average 1-predicted probability), (iii) variation (normalized Shannon entropy of predictions).

Results
Accuracy, Uncertainty, and Variation were strongly correlated (accuracy-uncertainty: r = -0.81; accuracy-variation: r = -0.87; uncertainty-variation: r = 0.90), yet their subtle distinctions proved useful for differentiating video quality. Among 762 cases, only one had accuracy = 0; variation was also 0, indicating unanimous model predictions across 28 frames. Closer review showed the ground truth (Apical 3-Chamber (AP3) based on DICOM tags) was incorrect, while the model correctly predicted Parasternal Long-Axis (LAX). Videos were manually labeled as no shadow (n = 23), low shadow (n = 466), or high shadow (n = 273) to capture shadows due to lung and rib. Shapiro-Wilk tests confirmed non-normality of all three metrics across groups. Consequently, Mann-Whitney U tests were applied; all pairwise comparisons were significant except Uncertainty between no- and low-shadow groups. A five-class balanced classification of echocardiographic views achieved accuracies of 0.85, 0.88, and 0.93 in the high-, low-, and no-shadow groups, respectively.
Conclusion
The results demonstrate that ensemble deep learning can systematically determine the level of image quality in exercise stress echocardiography.

[P36] Transferring Multi-depth CNN Pooling Features to Parallel LSTMs for Diabetes Risk Prediction from Primary Care EMRs

Karim Keshavjee, University of Toronto
Iqra Naveed, University of Management and Technology, Lahore, Pakistan
Muhammad Kaleem, University of Management and Technology, Lahore, Pakistan
Aziz Guergachi, Toronto Metropolitan University, York University

Introduction
Type 2 diabetes causes substantial morbidity and cost. Many prediction models underuse temporal patterns and higher-order biochemical relationships. We tested a deep architecture that preserves information from early, middle, and late convolutional pooling layers and transfers these representations to parallel LSTMs for incident diabetes prediction. We also examined whether simple engineered metabolic features add value.

Methods
We used de-identified Canadian primary care EMR data (1998–2015). Adults without diabetes at first visit were followed for incident diabetes. Features included demographics, vitals, labs, and diagnoses. Two feature sets were built: original, and original plus engineered ratios and flags (for example TG/HDL, LDL/HDL, total-cholesterol/HDL, VLDL≈TG/2.2, obesity from BMI, prediabetes flag). Data were split 80 percent train and 20 percent test. The model routed feature maps from the first, middle, and last CNN pooling layers into three parallel LSTMs whose outputs were fused for classification. Baselines included KNN, SVM, LSTM, Bi-LSTM, and CNN-LSTM. Metrics were accuracy, sensitivity, specificity, precision, F1, and MCC.

Results
The proposed CNN→parallel-LSTM approach outperformed all baselines. With the merged feature set and a larger training cohort, test results reached accuracy 93.2 percent, sensitivity 75.7 percent, specificity 98.8 percent, precision 95.3 percent, F1 84.4, and MCC 81.0. Using only original features reduced accuracy to 92.2 percent. Fasting glucose, A1c, engineered prediabetes, lipid ratios, and BMI ranked highest in importance.

Discussion/Conclusion
Preserving and transferring multi-depth CNN pooling representations to parallel LSTMs improves incident diabetes prediction from routine EMRs. Adding simple, clinically interpretable engineered features further lifts performance. This design supports scalable identification of high-risk patients for prevention in primary care. External validation across sites and incorporation of social and behavioral factors are next steps.

[P37] Semi-automating risk of bias assessment in randomized controlled trials with deep learning methods

Daniel Xie, Department of Integrated Biomedical Engineering and Health Sciences
Fangwen Zhou, Evidence, and Impact, McMaster University
Ashirbani Saha, *Evidence, and Impact, McMaster University*
Cynthia Lokker, Evidence, and Impact, McMaster University
Target Audience: Systematic Reviewers, Health Research Methodologists

Introduction:
Systematic reviews (SRs) consolidate available clinical evidence, and each included study is critically appraised in duplicate to assess validity. The Cochrane risk of bias (RoB) tool is commonly used to appraise randomized controlled trials (RCTs) and the updated version, ROB2, is increasingly being adopted but is more complicated to apply. Reliably automating this process using machine learning (ML) could reduce the double-appraisal burden and accelerate evidence synthesis. This pilot project trained and evaluated multiclass ML models to predict the RoB 2.0 overall bias domain.

Methods:
759 RCTs with RoB2 assessments were retrieved from 64 Cochrane SRs. The 'overall bias' domain was classified as low (n=91), some concerns (n=327), and high (n=336). The dataset was split into 80% (training and validation) and 20% (testing) with five-fold cross-validation. Hyperparameters were tuned via grid search with an early stopping patience of 3. Titles/abstracts were input to train and select three BERT-based models which were ensembled and the resulting classification probabilities were fed into a random forest. The primary evaluation metric was Matthew's Correlation Coefficient (MCC) to account for class imbalances. Metrics were macro-averaged across the three classes.

Results
24 BERT models, and 36 RF models were fine-tuned. The top BioLinkBERT model with class weights achieved an MCC of 0.58, F1 0.70, 67% sensitivity, and 86% specificity. All models overestimated the 'some concerns' class.

Discussion
Our pilot approach with an ensemble BERT model combined with a random forest allowed us to discern between the three RoB 2.0 classes and highlighted the need for further experimentation to improve model performance. Recent automation of RoB 2 with large language models show better performance than deep learning, however agreement with human reviewers remains poor and resource use is greater. This study has laid the foundation for training and evaluation approaches for a multiclass RoB assessment.

[P38] Mapping Convergence and Divergence Between Health Informatics and Learning Analytics Using Topic Modeling

Mi Song Kim, Western University

Introduction/Literature review: Health informatics (HI) and learning analytics (LA) are increasingly intersecting, offering potential to transform health professions education, clinical training, and professional development. In HI, learning occurs at human (clinicians, trainees), system (AI models), and organizational (policies) levels (Coppersmith et al., 2019; Madathil, et. al., 2025; Maurud, et al., 2023). This expanded view raises pressing questions: How do clinicians and algorithms co-learn from the same data? Can organizational learning in healthcare be aligned with learner engagement and curriculum outcomes? Despite these possibilities, the conceptual and methodological relationship between HI and LA remains fragmented, and thematic convergence and divergence have not been systematically examined (Bojic et al., 2023). This study maps the research landscape, identifying overlaps, divergences, and future directions.

Methods: A systematic Web of Science search was conducted for literature on HI, LA, and their intersections in health professions education. HI terms included "health/medical/clinical informatics". LA terms included "learning analytics" and "educational data mining." Searches were limited to English-language articles published between 2000 and 2025. With only two intersection studies, two separate corpora (152 studies) were analyzed for comparative analysis via NLP topic modeling to identify convergent and divergent themes across the two fields.

Findings: Along the descriptive–predictive–prescriptive continuum, convergence and divergence emerged. Descriptively, both fields emphasize multimodal data, visualization, and dashboards (Topics 0, 1, 5), though HI prioritizes data quality and clinical infrastructure, while LA tracks engagement and learning outcomes. At the predictive level, convergence emerges around AI-enabled modeling and decision support (Topic 3). Both anticipate outcomes: HI forecasting patient trajectories and LA predicting learner performance though with different contexts and stakeholders. At the prescriptive level, HI emphasizes system-level governance, policy, and organizational learning (Topic 2), whereas LA centers on adaptive feedback and personalized wellbeing support (Topic 4). Meta-level discussions (Topic 6) signal efforts to integrate research questions and frameworks across domains.

Discussion: These findings suggest cross-pollination: HI's ethical and governance frameworks could inform learner-data privacy, while LA's pedagogical modeling could enhance clinical training. Integrating both can foster adaptive and ethically grounded systems, bridging learning and clinical practice within a shared data ecosystem.

[P39] Automated Translation of Chronic Disease Diagnosis Codes Across Versions of the International Classification of Diseases Using the GPT-4 Large Language Model

Barret Monchka, University of Manitoba

"Introduction: The International Classification of Diseases (ICD) is a medical coding system used for healthcare system administration, public health surveillance, and research. Crosswalk tables, which map diagnosis codes across ICD versions, are time-consuming and costly to produce but are essential to reduce medical coding errors when healthcare systems periodically adopt ICD updates. Automated crosswalk development could better support clinical staff during transition periods and facilitate research spanning multiple ICD versions.

Methods: We evaluated the accuracy of the fourth-generation OpenAI Generative Pre-trained Transformer (GPT-4) model to translate chronic disease diagnoses across the 9th and 10th revisions of U.S. and Canadian ICD systems. Nine prompting strategies were developed. The three most-accurate prompts were combined to form composite prompts for Canadian and U.S. contexts. Accuracy was evaluated against crosswalks developed by Canadian and U.S. health services organizations. Each prompt was executed 10 times to assess variability, with mean accuracy ± standard deviation (SD) reported across replications.

Results: Across the nine prompting strategies evaluated for translating Canadian ICD codes, accuracy ranged from 32.5% to 47.4%, with the highest accuracy attained when prompts included diagnosis code labels. Prompting also impacted model variability, with SDs ranging from 0.6% to 3.6%. When combining the three best-performing prompts, GPT-4 achieved accuracies of 48.3% ± 0.8% (Canada) and 38.8% ± 0.6% (U.S.). Accuracy varied substantially across disease categories: 39.0%-70.0% (Canada), 30.1%-58.3% (U.S.). When evaluating only the first three characters of predicted codes, GPT-4 achieved 84.7% ± 0.6% (Canada) and 86.2% ± 0.4% (U.S.).

Conclusions: Large language models (LLMs) are highly sensitive to prompting strategy, substantially impacting both accuracy and variability. Testing multiple prompting strategies is recommended when employing LLMs for clinical or research applications. While current performance is insufficient for deployment, pre-trained LLMs show promise for automated ICD code translation and could become reliable with targeted refinement.

[P40] Trial Files: Leveraging large language models to summarize practice-changing clinical trials for clinicians

Katarina Zorcic, Sinai Health System
Emily Bartsch, Temerty Faculty of Medicine, University of Toronto
Bryant Lim, Sinai Health System, Department of Medicine, University of Toronto
Genevieve McCallum, Faculty of Medicine, University of British Columbia, Vancouver, BC
Tamara Van Bakel, Sinai Health System, Department of Medicine, University of Toronto
Alison Hacker, School of Medicine, Queen's University, Faculty of Health Sciences
Mike Fralick, Sinai Health System, Department of Medicine, University of Toronto

Target Audience: Clinicians (general internists, subspecialists, and trainees) seeking practical tools to remain up-to-date with rapidly evolving medical literature.

Introduction: Clinicians face an overwhelming volume of newly published randomized controlled trials (RCTs), making it difficult to stay current with emerging evidence. Large language models (LLMs) offer a scalable approach to rapidly summarize trial findings for medical education and clinical practice.

Methods: We developed and prospectively evaluated Trial Files, a newsletter that leverages an LLM to generate summaries of RCT abstracts from five high-impact journals in internal medicine. We created a software tool, called PaperScrape, which leverages the Medline application programming interface (API) to identify relevant RCTs. Information from each RCT's abstract was extracted, and plain-language summaries were generated using OpenAI's LLM API. Accuracy of model-generated summaries was compared against manual review prior to newsletter launch using an earlier version of the model and prompt. Subscriber survey data and growth metrics were analyzed to assess perceived usefulness and dissemination.

Results: Between June 2023 and October 2025, 61 newsletters featuring three RCTs each were distributed to 1,256 subscribers. The accuracy of 96 randomly selected RCT summaries was evaluated. Accuracy after prompt engineering improved substantially across domains: 97.1% for study phase, 92.2% for blinding, 85.4% for sample size, 97.9% for patient population, 94.7% for comparison groups, and 92.7% for primary outcomes. Survey respondents (n=43) rated the newsletter 4.7/5 on average, with 98% agreeing it made it easier to keep up-to-date with new trials.

Conclusion: LLMs can generate concise and accurate summaries of RCT abstracts, providing a feasible approach to improve knowledge translation and help clinicians stay up-to-date on recently published trials.

[P41] Building and Evaluating Predictive Models to Forecast Urgent Dialysis Needs Across Four Tertiary Hospitals

Katarina Zorcic, Sinai Health System
Kevin Zhu, Sinai Health System
Christopher Chan, Division of Nephrology, University Health Network
Mike Fralick, Sinai Health System, Department of Medicine, University of Toronto

Target Audience: Clinicians, nephrologists, hospital administrators, and health data scientists seeking to use predictive modeling to optimize dialysis staffing, resource allocation, and operational efficiency in acute care settings.

Introduction: Urgent dialysis is a critical and resource-intensive procedure that requires specialized nursing staff. Hospitals typically schedule a fixed number of dialysis nurses per day, despite highly variable daily demand, leading to inefficiencies in staffing and care delivery.

Methods: We developed and evaluated statistical and deep learning models to forecast daily urgent dialysis needs for the next seven days across four tertiary hospitals in Toronto, Canada (Hospital A for one hospital and Hospital B for three hospitals combined). The study included both a retrospective (April 2018 to March 2023) and a prospective evaluation period (November 2023 and June 2024). We compared several forecasting approaches, including autoregressive integrated moving average (ARIMA), Holt-Winters exponential smoothing, and temporal convolutional networks (TCNs), against a benchmark model based on the historical daily average number of dialysis procedures. Model performance was assessed using mean absolute error (MAE).

Results: Across all sites, 37,388 urgent dialysis procedures were analyzed. Retrospectively, the ARIMA and TCN models achieved an MAE of 3.0 and 1.5 procedures per day for Hospital A and Hospital B, respectively, outperforming the benchmark (MAE 4.4 and 1.9 for Hospital A and B, respectively). During the two prospective silent deployments, ARIMA achieved an MAE of 2.2 (Hospital A) and 1.5 (Hospital B), compared to an MAE of 3.0 (Hospital A) and 2.0 (Hospital B) for the benchmark.

Conclusion: This multicentre, six-year study demonstrates that urgent dialysis needs can be accurately forecasted using time-series and deep learning models.

[P42] Machine-learning assisted exacerbation detection and cough classification in COPD patients

Nicholas Li, Schulich School of Medicine & Dentistry, University of Western Ontario

Robert Wu, Department of Medicine, Toronto General Hospital

Alex Mariakakis, Department of Computer Science, University of Toronto

Chronic obstructive pulmonary disease (COPD) is a progressive disease characterized by airflow limitation, excess mucus production and worsening dyspnea. Patients with COPD often experience acute exacerbations of COPD (AECOPD), resulting in excess sputum production and dyspnea requiring hospitalization and medical intervention. At-home monitoring systems using AI to detect and record coughs have been proposed to detect imminent AECOPDs. These systems have yielded vast datasets of cough audio clips which have been underutilized in predictive applications. The aim of this study is to leverage machine learning and AI recorded cough audio to predict patient symptoms and outcomes. Two tasks were undertaken: cough classification and AECOPD prediction.

With regards to the first task of cough classification, we manually annotated 1,500 coughs as productive or non-productive, based on audio review. This dataset was combined with cough audio from the COUGHVID dataset, which contains 2000 clips with a similar labelling scheme. With regards to the second task of AECOPD prediction, cough and symptom data for 20 patients was obtained over 90 days. Hyfe Coughwatches were used to detect and record cough audio, and patients reported daily symptom scores using the London COPD questionnaire. Mel-frequency cepstrum coefficients (MFCCs) from 2-second cough clips were extracted, and average MFCCs over 4-hour intervals were used to generate a robust representation of the average cough during that time period. Days were labelled as either having a symptom exacerbation (London score >= 6) or non-exacerbation. Random forest, XGBoost and ElasticNet models were trained using these features with the task of predicting symptom exacerbation the following day.

Current results showed decent performance in cough classification (Model average F1 score = 0.58) and poor performance in AECOPD prediction (Model average F1 score = 0.25). In addition to further patient recruitment, we are currently exploring the potential for predicted cough classification as an engineered biomarker for AECOPD prediction including associations between predicted ""wet vs dry"" and individual symptom scores.

[P43] Accountability and Innovation in Health AI: Outcomes of the Health AI Systems Thinking for Community (HASTC) Workshop

Phoenix Wilkie, University of Toronto
Khashayar Namdar, Dalhousie Medicine New Brunswick; The Hospital for Sick Children
Konrad Samsel, University of Toronto
Sachin Pasricha, Division of Nephrology, Temerty Faculty of Medicine, University of Toronto
Nikhil Jaiswal, Research Institute of the McGill University Health Centre
Zoryana Salo, University of Toronto
Marianne So, University of Toronto
Shlok Panchal, Faculty of Health Sciences, McMaster University
Armaan Malhotra, University of Toronto; Department of Surgery, University of Toronto
Eptehal Nashnoush, University of Toronto; Trillium Health Partners
Sarah Walker, Department of Computer Science, University of Toronto
Somin Lee, Department of Electrical and Computer Engineering, University of Toronto
Muhammad Mamdani, University of Toronto
Leo Anthony Celi, Department of Biostatistics, Harvard T.H. Chan School of Public Health

Target audience: Students, clinicians, data scientists, policymakers, ethicists, and healthcare leaders engaged in AI governance.

Introduction:
Artificial intelligence (AI) is rapidly transforming healthcare, but its deployment raises pressing concerns about equity, accountability, and transparency. Without safeguards, AI risks reinforcing systemic biases and exacerbating inequities, particularly among marginalized groups. To address these challenges, the Health AI Systems Thinking for Community (HASTC) workshop was convened at the University of Toronto in October 2024 to facilitate interdisciplinary dialogue on managing AI-related harms.

Methods:
The workshop gathered 66 participants from academia, healthcare, and non-profit organizations, including students, researchers, clinicians, and data scientists. Attendees were divided into 10 groups, each mentored by experts in AI, clinical practice, and ethics. Using case studies on algorithmic bias, privacy concerns, and unintended clinical consequences, groups engaged in structured discussions to identify worst-case scenarios and propose safeguards. Discussions were organized around three guiding principles: accountability, transparency, and fairness. Each group summarized its findings in short presentations, followed by a plenary debrief.

Results:
Participants emphasized that AI models trained on biased datasets risk perpetuating inequities. Safeguards proposed included bias mitigation checklists, use of diverse and representative datasets, mandatory bias training for developers, and regular audits. Transparency measures such as model cards, tagging AI-generated content, and embedding hyperlinks to source data were recommended to enhance accountability. Participants also highlighted vulnerabilities to cyberattacks and privacy breaches, recommending penetration testing, standardized outputs, and adherence to secure, approved platforms. Across case studies, participants stressed the importance of context-specific validation, adaptive regulations, and the integration of human oversight in all AI-enabled clinical decision-making.

Discussion/Conclusion:
The HASTC workshop demonstrated the value of cross-disciplinary collaboration in shaping equitable and accountable AI governance in healthcare. Key outcomes underscored the need for dynamic, adaptive regulations and safeguards tailored to local contexts. By integrating community-driven advocacy, interactive learning, and continuous monitoring, stakeholders can better anticipate risks and build public trust. Workshops like HASTC should be repeated regularly to sustain dialogue, refine safeguards, and ensure that AI technologies evolve to promote ethical, equitable, and safe healthcare for all communities.

[P44] Deep learning–based tongue motion analysis from unrestricted video in MND

Mahri Kadyrova, University of Toronto

Ervin Sejdic, Electrical and Computer Engineering Department, University of Toronto

Lorne Zinman, Sunnybrook Research Institute

Agessandro Abrahao, Sunnybrook Research Institute

Yana Yunusova, Sunnybrook Research Institute

Target Audience: Clinicians, ALS/MND researchers, and machine learning specialists.

Introduction. Monitoring motor neuron disease (MND) progression, particularly tongue function, is vital for timely intervention but limited by access to and variability of clinical assessments. Remote, automated video-based evaluation using machine learning–derived tongue motion features offers a scalable alternative. This study aimed to (1) identify the best-performing tongue segmentation model and (2) clinically validate a tongue kinematic feature describing its deformation.

Methods. A total of 133 participants with MND (53 females; median age = 65.0 [59.0–70.0]) completed the Amyotrophic Lateral Sclerosis Bulbar Dysfunction Index–Remote (ALSBDI-R) [1] and performed 3 additional clinical tongue tasks—Relax, lateral-fast (L-fast), and lateral-normal (L-normal)—via video in unrestricted environments. From 711 videos, up to 100 frames per video (range = 46–146; median = 100 [90–107]) were manually annotated with high reliability [2]. Two decoders (U-Net, U-Net++) and five pre-trained encoders (MobileNetV2, VGG19, ResNeXt50_32x4d, EfficientNet-B5, InceptionV4) were fine-tuned using an 80/10/10 participant-level split. The top model (U-Net + EfficientNet-B5) generated tongue masks for L-fast videos. Dense optical flow and Jacobian determinants estimated tongue deformation, following the approach of [3], and were correlated with ALSBDI-R total score via Spearman's rank correlation.
Results. All models achieved >99.4% accuracy and specificity for tongue segmentation. While U-Net++ generally outperformed U-Net across encoders, U-Net with EfficientNet-B5 produced the best results (IoU = 87.74 ± 10.00, DSC = 93.05 ± 8.24, recall = 92.81 ± 8.95) and the lowest Hausdorff distance (10.20). Tongue body deformation significantly correlated with ALSBDI-R total score ($\rho = -0.68$, $p << 0.001$).

Discussion. Remote tongue localization and motion quantification in unrestricted videos proved feasible. Tongue body deformation demonstrated strong potential as a non-invasive biomarker for bulbar dysfunction and disease progression, enabling scalable remote MND monitoring.

[P45] Predictive Models on the Therapeutic Effects of Diverse Music Repertoires on Mental Health

Joey Qiao, Queen's University
Cindy Qiao, University of Toronto, Department of Computer Science

Music therapy is a recognized field of therapeutic intervention used in a diverse range of contexts, including hospital settings, retirement homes, palliative care, and community programs. It leverages the intrinsic qualities of music to address issues and enhance overall mental health. A challenge for music therapists working in hospice is a lack of suitable musical repertoires, called “music unpreparedness”. Machine learning (ML) is a form of artificial intelligence used to generate prediction models using information from datasets. This study aims to use a ma**c**hine learning model to evaluate repertoires and predict their effects on an individual’s mental health given their age and mental health condition, if any.

An open-access dataset from Kaggle containing 736 responses to an online survey about music and mental health was used [3]. In total, 33 features were included in the dataset including the respondent’s age, the number of hours they spend listening to music per day, their favourite genre, et cetera. During training, a decision model learned to partition the dataset based on features like mental health conditions and music genres, aiming to predict the corresponding effects on the respondent’s mental health: improved mental health, worsened mental health, or no effect. The most important features identified were age, hours per day spent listening to music, favourite genres (referring to specific genres and their respective indicator variables), frequency of listening to specific genres of music, and depression levels.

The model’s performance was evaluated by generating accuracy scores on training and validation datasets. The relatively high training accuracy score (0.8107) indicates that the model has effectively learned the patterns and relationships within the training data. In addition, the validation accuracy score (0.7419) being close to the training accuracy score suggests that the model is generalizing to new data instead of merely memorizing the data (which would be considered “overfitting”).

A predictive model was developed with the ability to successfully predict the therapeutic effects of music on an individual’s mental health. In the future, this model can be applied clinically to assist professional music therapists in selecting customized music repertoires tailored to their clients’ needs.

[P46] Old EMR, New Tricks: Leveraging a legacy EMR for AI-based prediction of clinical deterioration in a post-acute hospital

Jordan Pelc, Sinai Health
Alexander Kauffman, Sinai Health
Rahul Joshi, Sinai Health
Natasa Lazarevic, Sinai Health

Target audience: Specialists in human factors, clinical informatics, and/or post-acute care

Introduction: Identification of clinical deterioration in hospital settings is essential for patient safety. However, nearly all of the work in this area has been done in acute-care environments, and post-acute environments often have poor data foundations. Our team sought an approach based on aggregate patient data to predict deterioration to improve patient safety in a post-acute facility, despite poor digital maturity.

Methods: A collaboration of human-factors specialists and physicians identified clinically-meaningful discrete data elements that could be pulled from the local legacy EMR, including 10 clinical variables and 41 medication categories. We conducted a retrospective analysis of 2,876 patient discharges from Hennick Bridgepoint Hospital between April 2024 and July 2025. We defined clinical deterioration as discharge to the emergency department. We trained and evaluated six supervised machine learning models on the above variables to predict discharge to the ED and compared their performance on an independent test set. The best-performing model was examined in detail to identify the most influential predictors.

Results: Using area under the precision-recall curve (AUPRC) as the primary performance metric, CatBoost achieved the highest value at 0.738 (95% CI 0.670-0.798). Partial dependence plots of the most influential predictors revealed patterns of association with risk of deterioration that were consistent with expectations from acute-care. For instance, for pulse rate, risk began to rise at approximately 80 bpm, with a steeper increase between 95 and 110 bpm. The remaining most influential variables likewise behaved similarly to expected.

Conclusion: This AI model for predicting clinical deterioration in post-acute care demonstrates: first, that discrete data from a legacy EMR system are adequate for meaningful predictive model of deterioration; second, a specific model for clinical deterioration which can be implemented in the real-world; and third, that predictors of deterioration in acute-care are predictive in post-acute care, as well."

[P47] Bridging the Data Gap: A Low-Cost Human Activity Recognition and Collection System

Ryan Chang, McMaster
Arvin Heydarian, Western University
Thomas, Doyle, McMaster University
Reza Heydarian, McMaster University

Human Activity Recognition (HAR) is the classification of human movements from sensor data, typically captured with accelerometers or gyroscopes. HAR has significant applications in healthcare, rehabilitation, and wearable monitoring, but progress is constrained by the limited availability of datasets and the high cost of commercial devices. To address these issues, we developed a low-cost HAR system that classifies activities in real time and enables scalable collection of new HAR datasets.

The system was built on Raspberry Pi 4B integrated with a 3-axis accelerometer. A custom dataset of six activities (walking, walking upstairs, walking downstairs, sitting, standing and lying) was recorded by five participants with the system attached to the participants' waist. The dataset was segmented into fixed windows and used to train three models: a time-series temporal and contextual contrasting (TS-TCC) model, a feed-forward neural network (FFN) and a support vector machine (SVM). Models were evaluated using k-fold cross-validation, and the best-performing ones were deployed on the Raspberry Pi. Real-time predictions with corresponding windows were uploaded to Amazon Web Services (AWS) IoT Core and stored in Amazon S3.

In k-fold analysis, the TS-TCC, FFN, and SVM models achieved average classification accuracies of 92.4 ± 3.9%, 82.7 ± 9.9%, and 81.6 ± 9.5% (mean ± SD), respectively. The deployed models performed real-time classification on the Raspberry Pi and successfully streamed predictions and signals to the cloud using WiFi.

This work shows that an affordable HAR system can achieve high accuracy while generating high-quality datasets. By lowering barriers to data collection, the system provides a pathway from research prototypes to scalable clinical deployment. Such systems could support rehabilitation monitoring, fall detection, and chronic disease management. Furthermore, by using AWS, the system helps translate AI into everyday healthcare through centralized remote monitoring and notification capabilities."

[P48] Twelve Tips for Integrating Artificial Intelligence into Medical Education: A Practical Framework

Alireza Jalali, uOttawa
Salomon Fotsing, Family medicine department, University of Ottawa
Kadidja Harbi Houssein, Faculty of Medicine, Affaires francophones, University of Ottawa

Introduction:
Artificial Intelligence (AI) is transforming healthcare and medical education. However, educators face challenges in adopting AI effectively and ethically. This work proposes a practical, evidence-based framework of twelve actionable tips to guide AI integration across teaching, research, administration and leadership domains.

Methods:
A narrative synthesis of current literature, institutional practices, and expert recommendations was conducted. Sources included peer-reviewed studies, position statements (e.g., AMA, AAMC), and implementation reports.

Results:
The framework comprises twelve tips organized into five domains. Each tip is supported by research evidence and linked to practical tools.

- Teaching: Personalize learning paths, enhance feedback, integrate AI into case-based learning, and train students in AI literacy.
- Research: Accelerate literature reviews, employ AI in data analysis, and collaborate with AI for writing and publishing.
- Administration & Leadership: Automate routine tasks, use AI for strategic planning, and enhance faculty development.
- Ethics: Uphold transparency, fairness, and accountability.
- Future Readiness: Stay informed and adaptive as AI evolves.

Conclusion:
This framework offers a structured and evidence-based roadmap for integrating AI into medical education while maintaining ethical practices. It addresses immediate applications and anticipates future developments, supporting educators and institutions in preparing learners for AI-enabled healthcare."

[P49] Artificial Intelligence in Health Education: A Qualitative Exploration of Student Perspectives

Alireza Jalali, uOttawa
Joanne Fevry, uOttawa
Salomon Fotsing, Family medicine department, University of Ottawa
Marina Guirguis, Faculty of Medicine, University of Ottawa
Christine Landry, Faculty of Medicine, University of Ottawa
Diane Bouchard-Lamothe, Affaires francophones, University of Ottawa,
Jennifer Lacroix, Affaires francophones, University of Ottawa

Introduction: Artificial intelligence (AI) is increasingly used in education and healthcare. Its growing presence in professional training raises questions about its use by students, particularly in the healthcare field. It is essential to study the impact of this technology on their learning methods. Therefore, this qualitative study aimed to identify the AI tools most used by health students at the University of Ottawa, describe their usage habits, identify the tools that help them acquire new knowledge and develop skills, and explore the best strategies they consider to be the best for raising awareness and training their peers on the appropriate use of AI in learning environments.

Methods: A qualitative study was conducted at the University of Ottawa with students from ten health professions who had used AI in their learning. Data were collected via semi-structured interviews and an online survey with open questions from 51 students. An inductive thematic analysis within an interpretive paradigm was used to identify key themes emerging from the data.

Results: Among the 51 students who participated, AI was described as playing an increasingly important role in their educational experiences. Main themes were adoption patterns, usage habits, and critical assessment of AI tools. AI was mostly seen as complementary, but valued for improving efficiency, knowledge acquisition, and problem-solving. The most widely used tool was ChatGPT, which was adopted out of curiosity, peer influence, and to improve work efficiency.

Discussion / Conclusion: This research highlights students' use of artificial intelligence tools. Although AI is perceived as a valuable complement to medical training, its limitations require further research to support its effective implementation and evaluate long-term educational outcomes. Future studies should examine how AI integration influences professional competencies and patient care outcomes.

[P50] Gaps in Vendor Transparency: An Environmental Scan of AI-CDSS for Primary Care

Angela Coderre-Ball, Centre for Effective Practice
Pippy Scott-Meuser, Centre for Effective Practice
Anjana Ravi, Centre for Effective Practice
Vicky Wu, Centre for Effective Practice

Target Audience: Clinicians, policymakers, and health system leaders assessing artificial intelligence (AI) clinical decision support systems (AI-CDSS) for implementation in primary care.

Introduction: AI is increasingly entering clinical workflows in primary care, supporting administrative tasks as well as diagnostic and therapeutic decisions. While they demonstrate promise in reducing clinician administrative burden, decision makers need transparent information on the AI's evidence base, methodology, and data safeguards to be able to make informed choices before implementing AI-CDSS. This environmental scan (eScan) examined AI-CDSS solution details (as made publicly available by vendors) to assess market transparency and readiness for responsible AI-CDSS implementation in Canadian primary care.

Methods: An eScan identified AI-CDSS solutions with demonstrated or potential implementation in Canadian primary care. Searches included vendor websites, vendors of record, and conference lists. Solution-specific details were extracted using a structured template aligned with TEHAI, FUTURE-AI, and Health Canada's guidance. Data was then synthesized across three domains: AI-knowledge base, AI methods, and privacy.

Results: Of identified vendors (n=53), 55% disclosed some knowledge base details, though disclosure of alignment with evidence-based guidelines varied. Only 15% described AI methodologies beyond "proprietary algorithms." Additionally, whereas 66% published details specific to patient privacy, 15% of vendors referenced compliance with PIPEDA, 26% disclosed Canadian data storage, and 33% explicitly stated that patient data was not used for secondary purposes.

Discussion/Conclusion: The eScan reveals a significant transparency gap between vendor marketing claims and the information required for safe, equitable AI-CDSS implementation. While vendors provided richer details privately through government procurement requests for information, public-facing disclosures remain limited and inconsistent. This disconnect suggests that programs like Evidence2Practice Ontario could play a stewardship role by setting clear expectations for transparency and curating trustworthy information to guide decision makers.

[P51] PAIdoc: Prototyping an AI Billing Assistant for OHIP in Ontario Family Practice

Miranda Chan, U of T Translational Research Program
Linette Penney, U of T Translational Research Program
Vedant Shukla, Translational Research Program
David Huang, Translational Research Program
Auva Zanandi, Translational Research Program
Rebecca Wu, Translational Research Program

Background:
Ontario's family physicians working in fee-for-service models spend significant time converting clinical notes into OHIP billing codes. This administrative work is repetitive, error-prone, and a major contributor to burnout and declining physician retention. Advances in large language models (LLMs) provide an opportunity to automate billing directly from SOAP (subjective objective assessment plan) clinical notes, reducing workload while improving accuracy and efficiency.

Objective:
To develop and test a proof-of-concept AI billing assistant ("PAIdoc") that processes SOAP notes to generate OHIP fee codes for family physicians, and to assess feasibility for integration into clinical workflows.

Methods:
PAIdoc was built using a commercially available LLM configured to analyze SOAP notes from family practice encounters. The system applied natural language processing to identify key clinical details and match them with OHIP codes using a combined AI- and rule-based framework. Due to restricted access to large datasets, testing was performed on anonymized and simulated family medicine cases. Usability and workflow considerations were refined through iterative feedback from family physicians and trainees.

Results (Preliminary):
PAIdoc successfully mapped SOAP note content to clinically appropriate OHIP codes. Early feedback indicated potential to reduce billing time and lower administrative frustration. While large-scale validation is pending, the prototype identified clear next steps: training on larger datasets, integration with electronic medical records (EMRs), and benchmarking against physician-coded claims.

Conclusion:
PAIdoc demonstrates the feasibility of applying AI to a high-burden administrative task in Ontario family practice. By reducing billing workload, this approach may support physician well-being, improve coding accuracy, and protect time for patient care. Future work will focus on dataset access, benchmarking, and pilot testing within EMR environments. Furthermore, PAIdoc could be expanded across provinces and into other specialties to streamline the billing process.

[P52] Exploring Radiologist Burnout Amid the Rise of AI: A Scoping Review of Challenges Impacts and Interventions

Pardaman, Setia, Department of Laboratory Medicine & Pathobiology
Alya Mohmood, Department of Laboratory Medicine and Pathobiology, University of Toronto
Linette Penney, Department of Laboratory Medicine and Pathobiology, University of Toronto

Background: - Burnout among radiologists is a growing concern, with more than one-third of radiologists worldwide experiencing symptoms of emotional fatigue and reduced job satisfaction. The arrival of Artificial Intelligence, particularly generative models, presents both opportunities and challenges in addressing this issue. This paper explores the various factors contributing to burnout such as heavy workloads, administrative tasks, understaffing, technological demands and evaluates how AI especially generative models might either alleviate or amplify the issue.

Methods: - A scoping review was conducted using the PICOS framework. Studies published between 2015 and 2025 were screened using Covidence. Inclusion criteria focused on practicing radiologists in clinical settings. Studies were selected based on their relevance to burnout and the integration of AI technologies, including generative models (e.g., large language models, image synthesis tools). Excluded were studies focusing on trainees, non-radiology specialties, or lacking full-text access.

Results: Out of 2405 screened articles, 87 studies met inclusion criteria. Burnout prevalence among radiologists ranged from 35% to 71%, with contributing factors including high workload, mounting administrative duties and the uncertainty that AI adoption brings. Some papers warned that AI could supplant radiologists while others pointed to generative-AI tools automated report generation, decision-support engines and workflow-optimization modules that appear to lighten burdens and lift efficiency. However, ethical concerns and trust in AI systems are still barriers to adoption.

Conclusion: -Generative AI presents a paradox in radiology: while it can streamline workflows and reduce burnout-related stressors, it also introduces new anxieties around professional roles and ethical boundaries. Generative AI has the promise to transform radiology practice and alleviate burnout, but its implementation must be guided by ethical frameworks, clinician engagement, and robust validation. Addressing burnout through AI requires a balanced approach that prioritizes both radiologist well-being and the delivery of safe, high-quality patient care.

[P53] Mediation CNN (Med-CNN) Model for High-dimensional Mediation Data

Jasper Zhongyuan Zhang, Dalla Lana School of Public Health, University of Toronto

Yao Li, Dalla Lana School of Public Health, University of Toronto

Olli Saarela, Dalla Lana School of Public Health, University of Toronto

Divya Sharma, Dalla Lana School of Public Health, University of Toronto

Wei Xu, Dalla Lana School of Public Health, University of Toronto

Complex biological features such as human microbiome and gene expressions play a crucial role in human health by mediating various biomedical processes that influence disease progression, such as immune responses and metabolic processes. Understanding these mediation roles is essential for gaining insights into disease pathogenesis and improving treatment outcomes. However, analyzing such high-dimensional mediation features presents challenges due to their inherent structural and correlations, such as the hierarchical taxonomic structures in microbial operational taxonomic units (OTUs) and gene-pathway relationships, and the high dimensionality of the datasets, which complicates mediation analysis. We propose the Med-CNN model, an iterative approach using Convolutional Neural Networks (CNNs) to incorporate the complex biological network of the mediation features. The output values from network-specific CNN models are condensed into an integrative mediation metric (IMM), which captures essential biological information for estimating mediation effects. Our approach is designed to handle the high-dimensional data and accommodate their unique structures and non-linear interactive mediation effects. Through comprehensive simulation studies, we evaluated the performance of our algorithm across different scenarios, including various mediation effects, effect sizes, and sample sizes, and compared it to conventional methods. Our simulations demonstrated consistently lower biases in mediation effect estimates, with values ranging from 0.17 to 0.56, which were lower than other established methods ranging from 0.24 to 13.27. In a real data application, our method identified a mediation effect of 0.06 between ethnicity and vaginal pH levels.

[P54] Large language models for electronic health records in pediatric and surgical care: a systematic review

Carmel Daskalo, McGill University Faculty of Medicine and Health Sciences
Waseem Abu-Ashour, The Montreal Children's Hospital, McGill University Health Centre
Jean Marie Tshimula, Université de Sherbrooke, University of Kinshasa
Mohsen Amoei, McGill University Faculty of Medicine and Health Sciences
Elena Guadagno, The Montreal Children's Hospital, McGill University Health Centre
Dan Poenaru, McGill University Faculty of Medicine and Health Sciences

Target Audience: Clinicians (specifically in pediatric and surgical specialties), healthcare AI researchers, health data scientists, healthcare administrators, and policy makers

Introduction:
Large language models (LLMs) are promising tools in healthcare, particularly for accessing unstructured, text-based electronic health record (EHR) data. Studies on the application of LLMs for EHRs in pediatrics and in surgery are limited. This systematic review evaluates the applications of LLMs in the EHR for pediatric and surgical care, model performance compared to traditional methods, and proposed clinical potential to improve healthcare processes, patient outcomes, and overall quality of care.

Methods:
A senior medical librarian, adhering to PRISMA guidelines, searched ten databases from inception until November 19, 2024. The search strategy included large language models or generative artificial intelligence, without language restrictions. Studies were included if they used EHR datasets in pediatric or surgical domains, employed transformer-based LLMs, provided benchmark comparisons or relevant performance metrics, and were primary research articles. Two reviewers independently screened studies for inclusion, with a third reviewer resolving conflicts. Risk of bias was assessed using PROBAST. Data analysis includes descriptive and summative statistics.

Results:
Among 4,326 identified studies, 44 met the inclusion criteria - 30 (68.2%) in all surgical specialties, 3 (6.8%) in pediatric surgical subspecialties, and 11 (25%) in pediatrics. Most studies (59.1%) were published in 2024. LLM types included Bidirectional Encoder Representations from Transformers (BERT) and their variants (52.3%), ChatGPT (29.5%), and other models (18.2%). Most studies (90.9%) relied solely on retrospective unstructured data, and 40.9% focused on classification tasks. LLMs demonstrated performance improvements in 78.1% of studies with a traditional comparator. Clinical documentation assistance (54.5%) and diagnostic and clinical decision support (36.4%) were the most commonly proposed applications for LLMs.

Discussion/Conclusion:
While LLMs offer opportunities for EHR analysis in pediatric and surgical care, most studies remain early-stage, with notable limitations including limited external validation and lack of evaluation in actual clinical workflows. Future research should prioritize rigorous validation and real-world testing to support their safe and effective use in practice.

[P55] Hybrid Spatial–Functional Deep Learning for Tumour Depth Estimation in Fluorescence-Guided Cancer Surgery

Michael Daly, Princess Margaret Cancer Centre
Xinyuan Huang, University of Toronto, Princess Margaret Cancer Centre
Hikaru Kurosawa, Princess Margaret Cancer Centre
Jack Wunder, Princess Margaret Cancer Centre
Sujit Patil, Princess Margaret Cancer Centre
Karthik Kuber, Dept. Computer Science, University of Toronto
Jonathan C. Irish, Dept. Otolaryngology – Head & Neck Surgery, University of Toronto
Michael J. Daly, Princess Margaret Cancer Centre

Introduction: Tumour depth estimation is critical for achieving safe surgical margins in cancer surgery. While fluorescence-guided surgery (FGS) delineates tumour boundaries on the tissue surface, it cannot provide quantitative information on subsurface extension. Spatial frequency domain imaging (SFDI) extends FGS by capturing both fluorescence and tissue optical properties, but converting these 2D optical images into accurate 3D depth maps remains challenging. We present a hybrid deep learning approach that combines spatial feature extraction and functional regression to address this limitation.

Methods: We trained models on 10,000 simulated tumours generated using diffusion theory and Monte Carlo light propagation methods and evaluated transfer to 36 patient-derived phantoms. Inputs included fluorescence images across six spatial frequencies and optical property maps of absorption and scattering. We investigated three approaches: (i) a shape-segmentation U-Net with auxiliary Dice supervision; (ii) a pixelwise Kolmogorov–Arnold Network (KAN) for functional regression; and (iii) a hybrid model coupling a Siamese attention U-Net backbone with a two-layer KAN regression head applied to tumour pixels identified by segmentation. Transfer learning was applied to adapt from simulated to phantom data.

Results: On simulation data, KAN achieved the lowest mean absolute error (MAE ≈ 1.0 mm), while U-Net achieved the highest segmentation accuracy (Dice > 0.8). Zero-shot phantom evaluation revealed performance degradation across models due to domain shift. With transfer learning, the hybrid U-Net + KAN reduced phantom tumour-region MAE from 1.85 mm to 1.25 mm and improved minimum-depth error to 0.72 mm, demonstrating robustness in shallow-margin estimation.

Conclusion: These findings show that hybrid U-Net + KAN models leverage complementary strengths in spatial and functional learning, improving tumour depth estimation across domains. This approach has potential to enhance intraoperative reliability of fluorescence-guided surgery.

[P56] Applications of artificial intelligence in pediatric surgical pathology: A systematic review

Eve Wang, McGill University Faculty of Medicine and Health Sciences
Sarah Wu, McGill University Faculty of Medicine and Health Sciences
Mohsen Amoei, McGill University Faculty of Medicine and Health Sciences
Elena Guadagno, The Montreal Children's Hospital, McGill University Health Centre
Karl Grenier, The Montreal Children's Hospital, McGill University Health Centre
Dan Poenaru, The Montreal Children's Hospital, McGill University Health Centre

Target Audience. The abstract targets clinicians (especially surgeons and pathologists) and clinical researchers who want to understand the current state and potential of AI in pediatric surgical pathology, with accessibility for AI scientists seeking clinical context.

Background. Artificial Intelligence (AI) techniques have the ability to transform and enhance diagnosis and treatment response predictions in pediatric surgical pathology. AI offers the potential to reduce the workload of pathologists by automating routine and labor-intensive tasks. In this systematic review, we investigate the current applications of computational (AI) pathology in pediatric surgical conditions.

Methods. Nine databases were searched from inception until January 2025 to retrieve articles looking at machine learning, AI, or virtual reality in the pathological diagnosis of pediatric surgical conditions, without language restrictions. PRISMA standards were followed, and abstract screening was performed by two reviewers, with conflicts resolved by the senior author. Original studies and reviews exploring computational pathology for diagnosing, grading, or predicting outcomes in pediatric surgical diseases were included. Extracted data included study design, type of AI model used, type of disease studied, and model performance metrics.

Results. The authors screened 3363 articles, with 34 meeting the inclusion criteria. AI applications primarily involved image-based diagnostics using convolutional neural networks (24, 70.6%), trained on whole-slide images. The most frequently studied diseases were childhood cancers (19, 55.9%), including neuroblastoma and medulloblastoma, and Hirschsprung's disease (8, 23.5%). Diagnostic support was the most common objective (13, 38.3%), followed by classification, prognostication, and treatment prediction. Nearly half (15, 44.1%) included model explainability tools, while performance metrics were heterogeneous, most often accuracy (24, 70.6%) and AUROC (21, 61.8%). Twenty-two studies (64.7%) had a high risk of bias, mainly due to small, single-center cohorts, poorly defined predictors, and outcome-informed assessments.

Conclusion. AI shows promise for pediatric surgical pathology, especially in image-based diagnosis of cancers and Hirschsprung's disease. However, high risk of bias, small cohorts, and limited validation highlight the need for rigorous, standardized studies before clinical implementation.

[P57] Large language model summarization to support communication in palliative care: a protocol

Brigitte Durieux, Department of Family Medicine, McGill University
Mohsen Amoei, Department of Experimental Surgery, McGill University
Ting Du, Research Institute of the McGill University Health Centre
Tibor Schuster, Department of Family Medicine, McGill University
Justin Sanders, Department of Family Medicine, McGill University
Dan Poenaru, The Montreal Children's Hospital, McGill University Health Centre

Title: Large language model summarization to support communication in palliative care: a protocol

Introduction: Patients affected by serious illness often accrue extensive electronic health records (EHRs), which are difficult and time-consuming to review. Inefficient information synthesis can reduce time available for relationship-building during visits, contribute to clinician burnout, and create risk through inadvertently missed information. Rising rates of serious illness necessitate interventions to reduce inefficiencies and support timely, person-centered palliative care. This project seeks to leverage generative artificial intelligence (GenAI) to improve efficiency of information transfer to palliative care physicians by providing patient summaries before initial patient consults.

Methods: On a secure server at the Research Institute of the McGill University Health Centre, we have prepared an initial data pipeline using locally hosted large language models (LLMs) for EHR data summarization. To inform and refine our process for preparing summaries (phase 1), we plan to conduct focus groups with physicians and patients at our Cancer Centre to explore information needs and safety/evaluation considerations for our use of LLMs. To validate the quality of summaries (phase 2; informed by phase 1 findings), we will conduct a blinded, cross-sectional validation study to compare LLM-generated and clinician-authored summaries using a tool validated to assess clinical documentation quality and accuracy (scores ranging (low) 9–45 (high)). Each summary will be reviewed by two blinded experts, with quality scores & feedback analyzed descriptively across groups; inter-rater reliability will be calculated via intra-class coefficient (ICC, 56 scored summaries >39.4 required sample size for an ICC between 0.6-0.8). Future work includes a hybrid effectiveness-implementation feasibility study of LLM-generated EHR summaries in our Cancer Centre, to support the design of a randomized controlled trial.

Anticipated Results: Building on the promise and success of similar tools, this project seeks to develop a GenAI-powered tool to support high-quality palliative care based on the needs and guidance of patients and physicians.

[P58] Machine Learning Models for Individualized Osteoradionecrosis Risk Prediction in Head and Neck Cancer

Mohammad Moharrami, University of Toronto
Erin Watson, Faculty of Dentistry, University of Toronto
Shao Hui Huang, Department of Dental Oncology, Princess Margaret Cancer Centre
Sreenath Madathil, Faculty of Dental Medicine and Oral Health Sciences, McGill University
John Kim, Department of Radiation Oncology, University of Toronto
Andrew McPartlin, Department of Radiation Oncology, University of Toronto
Nauman H. Malik, Department of Radiation Oncology, University of Toronto
Sonica Singhal, Faculty of Dentistry, University of Toronto
Ezra Hahn, Department of Radiation Oncology, University of Toronto
John Waldron, Department of Radiation Oncology, University of Toronto
Scott Bratman, Department of Radiation Oncology, University of Toronto
John de Almeida, University Health Network, University of Toronto
Christopher Yao, University Health Network, University of Toronto
Andrew Hope, University Health Network, University of Toronto
Carlos Quinonez, Faculty of Dentistry, University of Toronto
Michael Glogauer, Faculty of Dentistry, University of Toronto
Ali Hosni, University Health Network, University of Toronto

Introduction: Osteoradionecrosis (ORN) of the jaw is a severe and debilitating late complication among patients undergoing radiation therapy (RT) for head and neck cancer (HNC). This study aimed to develop and validate predictive models for ORN following RT in patients with HNC using time-to-event data that account for death as a competing risk, and to quantify the extent of risk overestimation that occurs when the competing risk is ignored.

Methods: In this prognostic study of patients who underwent curative RT between 2011 and 2018, with ongoing follow-up, sociodemographic, clinical, and dosimetric data were collected. The binary ORN outcome was defined by the ClinRad system (grade ≥1); all-cause mortality was the competing event. Fine-Gray regression (FGR), Random Survival Forests (RSF) with Gray's test splitting rule, and DeepHit were implemented using repeated nested stratified cross-validation. Feature selection and interpretation were guided by SHapley Additive exPlanations (SHAP). For comparison, non-competing risk models such as Cox proportional hazards (Cox PH) and standard RSF (S-RSF) with log-rank splitting rule were also trained.

Results: Of 2,466 patients, 183 developed ORN during follow-up, and 714 died. Three versions of each model were developed using 20, 10, and 5 features. The 10- and 5-feature RSF models performed best. Considering simplicity, the 5-feature model, which included tumor site, D10cc, smoking pack-years, periodontal condition, and dental insurance, was selected for production. At 60 months, Brier Score was 0.061 (95% CI: 0.060–0.063), Integrated Brier Score 0.038 (95% CI: 0.037–0.040), time-dependent AUC 0.776 (95% CI: 0.762–0.789), and C-index 0.772 (95% CI: 0.757–0.787). FGR closely followed, whereas DeepHit underperformed. Non-competing models, including the S-RSF, overestimated ORN risk, predicting an average 60-month cumulative incidence of 8.7% versus 6.8% with 5-feature RSF.

Conclusion: A parsimonious RSF model reliably estimated individualized ORN risk while avoiding overestimation from ignored competing risks. To facilitate clinical use, an interactive web application was developed and is available at https://orn-prognosis.onrender.com.

[P59] External Validation of the Multifocal Electroretinogram Classification Interface (MERCI), a Machine Learning Algorithm for Hydroxychloroquine Toxicity

Godfrey Wong, Kensington Vision and Research Centre
Tom, Wright, Kensington Vision and Research Centre
Brian Ballios, University Health Network

Chronic use of hydroxychloroquine (HCQ), a drug prescribed for systemic auto-immune disease, can lead to retinal toxicity and permanent central vision loss. American Academy of Ophthalmology (AAO) guidelines recommend patients receive annual ophthalmic examinations after 5 years of treatment, imposing a significant burden on both patients and the health care system.

MERCI is a machine learning model for automatic HCQ retinopathy detection (Habib, 2022) using multifocal electroretinography (mfERG). Trained on data from 1463 eyes of 748 patients, MERCI detected HCQ retinopathy with 91% sensitivity and 84% specificity.

This study validates the MERCI algorithm against a novel dataset, ensuring the algorithm can be generalizable to the wider HCQ exposed population.

255 patients referred to Kensington Health for HCQ retinopathy screening from July 2024 to August 2025 were recruited to form the novel validation dataset. Participants underwent mfERG testing, perimetry, spectral-domain optical coherence tomography, and fundus autofluorescence. Following AAO guidelines, a reference diagnosis was assigned to each patient. Patients were classified as having retinopathy if one or both eyes were abnormal. mfERG data from both datasets were also separately assessed for the presence of HCQ retinopathy using MERCI. Predictions were compared with the AAO diagnoses to assess agreement; diagnostic performance metrics were calculated to evaluate MERCI's accuracy.

Complete multimodal data required for AAO guidelines was only available for 129 patients from the original dataset. The sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) of the internal dataset are 62%, 66%, 17%, and 93% respectively. MERCI achieved 100% sensitivity with 58% specificity with the external dataset. The PPV and NPV of the external dataset are 17%, and 100.0% respectively.

This study confirms that MERCI maintains high sensitivity and NPV on unseen data. Further validation will include data from other sites to test MERCI's generalizability to a Canadian population. These findings support MERCI as an effective screening tool, reducing unnecessary testing and empowering clinicians to continue HCQ therapy in patients without retinal toxicity. By enabling scalable, objective, and accessible screening, MERCI has the potential to transform HCQ monitoring protocols and improve long-term outcomes for patients with rheumatic disease.

[P60] Exploring anthropomorphism for AI-supported decision-making in emergency psychiatry: A human-computer interaction experiment

Tosin Kasumu, Centre for Addiction and Mental Health
Shrika Vejandla, Centre for Addiction and Mental Health
Tosin Kasumu, University of Toronto
Shrika Vejandla, Queen's University
Patrycja Szkudlarek, Krembil Centre for Neuroinformatics
Laura Sikstrom, Krembil Centre for Neuroinformatics
Marta Maslej, Krembil Centre for Neuroinformatics

Artificial Intelligence (AI) is increasingly developed for applications in psychiatry, such as crisis risk assessment. While AI has potential to enhance decision-making, it may amplify inequities when clinicians rely on biased outputs that reinforce stereotypes about disadvantaged or marginalized patient groups. Understanding what drives this reliance is critical. One key factor is anthropomorphism, the attribution of human-like qualities to AI, which can heighten trust and encourage reliance on biased outputs.

We examined the impact of anthropomorphism on AI-supported risk assessment in an emergency mental health context. Participants recruited via crowdsourcing were randomized to one of three conditions: Control (n=103), wherein AI recommendations were depicted as coming from a computer system; Name (n=105), wherein AI recommendations were accompanied by a human name; and Photo (n=102), wherein AI recommendations were accompanied by a photo of a call operator. Participants read vignettes about individuals experiencing mental health emergencies, and decided whether to deploy police, or medical help. Vignettes were accompanied by AI recommendations that were biased to recommend police more for individuals depicted as Black, as compared to white.

Across conditions, participants were 48% more likely to deploy police when a vignette depicted a Black individual as compared to a white individual (OR=1.48, 95% CI: 1.24–1.75, $p < .001$), suggesting an overall effect of AI bias. Moreover, as compared to participants in the Control condition, participants in the Name condition were more likely to deploy police when a vignette depicted a Black individual (OR = 1.41, 95% CI: 1.03–1.94, $p = .031$), but not when it depicted a white individual. Exploratory analyses revealed effects of AI gender, but not race, on reliance.

Textual anthropomorphism, conveyed through names, language, or conversational style, has been found to heighten perceptions of trust, competence, and social presence, which can increase engagement and decision-making accuracy. Yet, when AI is biased, our findings suggest that textual anthropomorphism can amplify bias during decision-making, by increasing reliance on biased AI outputs. These findings highlight a need for considering how design-related factors impact AI-assisted decision-making equity in high-stakes settings, like healthcare.

[P61] Bridging AI and Clinical Flow: a Mixed Methods Evaluation of an XGBoost-Powered Queuing Tool in the Emergency Department (EDQ)

Matthew Yip, Temerty Faculty of Medicine, University of Toronto
Muhammad Mamdani St. Michael's Hospital, Unity Health Toronto
Derek Beaton, St. Michael's Hospital, Unity Health Toronto,
Neal Kaw, St. Michael's Hospital, Unity Health Toronto
Sam Vaillancourt, Temerty Faculty of Medicine, University of Toronto
Shaun Mehta, Temerty Faculty of Medicine, University of Toronto

Target Audience: AI researchers, data scientists, and clinician innovators in healthcare operations, particularly those focused on clinical workflow optimization, applied machine learning, and human–AI interaction in acute care environments.

Background: Queuing theory has long been proposed to improve patient flow, length of stay (LOS), and quality of care in emergency departments (EDs). Yet, real-world applications of AI-based queuing systems remain limited. This study evaluated the impact of an AI-powered queuing tool, EDQ, designed using XGBoost to reduce LOS in the ambulatory section of a Toronto ED by suggesting an optimal patient order and identifying patients who can be dispositioned after assessment by a physician.

Methods: A mixed-methods analysis was conducted on deployment of EDQ from April to November 2024. Quantitatively, EDQ's impact was examined at both encounter and aggregate levels. Encounter-level LOS before and after implementation was compared using generalized linear models using gamma distribution and the log link on a time-matched dataset controlling for seasonality. This analysis was adjusted for patient age, sex, and housing status. At weekly and monthly levels, interrupted time series analysis was performed. Qualitatively, semi-structured interviews were conducted with five flow coordinators who had used the tool, and responses were analyzed thematically.

Results: Encounter-level analysis using a Gamma regression showed no significant change in length of stay (LOS ratio 0.98, 95% CI 0.96–1.00). Interrupted time-series analyses similarly demonstrated no significant immediate level change (–2.74 minutes, $p = 0.27$) or change in post-intervention trend (0.04 minutes/week, $p = 0.42$). Qualitative interviews revealed skepticism toward algorithmic prioritization, perceived disruption to existing workflows, and limited confidence in the model's output.

Conclusion: Although EDQ showed no measurable improvement in LOS, this real-world deployment underscores critical human and workflow factors influencing the adoption and perceived value of AI tools in dynamic clinical environments. Future research should integrate principles of implementation science and human-centered design to bridge the gap between algorithmic performance and clinical usability.

[P62] Stratifying Task Importance Across ALS Bulbar Severity Using Multimodal AI

Vishnu Akundi, Electrical and Computer Engineering Department, University of Toronto
Ervin Sejdic, Electrical and Computer Engineering Department, University of Toronto
Lorne Zinman, Department of Medicine, Sunnybrook Health Sciences Centre
Agessandro Abrahao, Department of Medicine, Sunnybrook Health Sciences Centre
Yana Yunusova, Rehabilitation Sciences Institute, University of Toronto

Target Audience
Clinicians (neurologists and speech-language pathologists) interested in multimodal modeling, explainable AI, and neurological disease assessment, AI researchers, and biomedical engineers.

Introduction
Bulbar ALS assessment currently relies on lengthy, burdensome protocols [1,2]. AI methods enable automatic, objective, and remote evaluations. Our prior work showed that reduced task subsets retain diagnostic accuracy [3]. Here, we examine whether task importance shifts across disease severity levels.

Methods
We analyzed multimodal recordings from 83 ALS patients, with severity stratified into four categories (Normal, Mild, Moderate, Severe/Profound) using the ALS-Bulbar Dysfunction Index–Remote (ALSBDI-R) [4]. Patients performed the following speech (e.g., passage reading, syllable repetition (PA, TA, KA) and non-speech (e.g., jaw opening) tasks. Kinematic features were derived from Mediapipe facial landmarks, and acoustic features from segmented audio [5]. Using our machine learning pipeline, task importance is quantified using Leave-One-Group-Out (LOGO) analysis, with a focus on comparing relative contributions of speech versus non-speech tasks across each of the four severity categories [3,6].

Results
Certain tasks, such as PA actively harmed performance by 12% and are prime candidates for removal. Extending this framework, we are now in the process of stratifying LOGO-derived task importance by severity, which will allow us to identify how the value of speech and non-speech tasks changes across disease progression.

Discussion/Conclusion
By stratifying task importance across disease stages, this study seeks to uncover which tasks provide the most information at different stages of disease. This approach reframes ALS assessment from a static set of tasks toward an adaptive, stage-specific protocol, informed by interpretable AI. Through this study, we aim to reduce patient burden, improve sensitivity across the disease continuum, and move clinical assessments toward personalized, AI-guided monitoring.

[P63] Efficiency Improvements at a Single Medical Genetics Clinic in Toronto Using PocketMD as an EMR for Virtual Genetics Consultations

Andrew Shuen, McGill University, Genaissance

Background: Clinical genetics services in Ontario face significant administrative burdens due to extensive documentation requirements, including patient intake, genetic risk assessments, and genetic testing result reporting. Non-urgent referrals now have wait times exceeding two years in some Ontario clinics. Traditional genetics consultations typically last 60–90 minutes, with an additional 30 minutes for charting and 10–15 minutes to complete genetic test requisition forms. These prolonged workflows limit clinician availability for direct patient interaction, potentially impacting patient care quality. Initiatives such as Ontario Health's ""Patients Before Paperwork"" emphasize the importance of efficient documentation solutions.

Methods: We evaluated PocketMD, a mobile and web-based application featuring an AI-driven summarizer that rapidly converts virtual visit transcripts into structured consultation and SOAP notes within seconds. Comprehensive patient intake questionnaires were electronically administered via Jotform and integrated directly into the clinic's BIRD documentation system. PocketMD automated label generation for genetic test requisitions and shifted the responsibility of communicating consultation results to patients, eliminating clinician-generated faxes. A prospective evaluation was conducted involving 100 patient visits over three months, comparing consultation durations before and after implementing PocketMD.

Results: Implementation of PocketMD reduced average consultation duration by approximately 69% for initial consultations and 62% for follow-up visits across 100 patient visits. Specifically, genetic consultations were completed in an average of 28.3 minutes (SD ± 14.9 minutes) for initial visits and 23.2 minutes (SD ± 9.2 minutes) for follow-up visits. Genetic testing requisition time decreased from 10 minutes to 8 minutes, saving approximately 2 minutes per genetic test requisition (Ontario genetic tests), achieved through automated label generation. Electronic intake questionnaires integrated with BIRD substantially reduced clinician charting time. The clinician reported improved documentation accuracy and enhanced patient communication without compromising patient satisfaction.

Discussion: PocketMD directly addresses key inefficiencies in clinical genetics workflows through rapid AI-generated SOAP notes, streamlined patient intake, automated label generation, and reduced administrative tasks. The significant reduction in consultation duration increased clinician availability for direct patient interaction, improved documentation accuracy, and maintained high patient satisfaction levels.

[P64] Comparative Evaluation of Convolutional Neural Networks (CNNs) and Vision Transformer (ViT) Learning Models for Classifying Alzheimer's Disease in Small-Datasets

Avin Sharma, Imaging & Behavior, Krembil Brain Institute, University Health Network
Amulya Bhagirath, Princess Margaret Cancer Research Center, University Health Network
Imran Khan, Department of Computer and Mathematical Sciences, University of Toronto

Target Audience:
Neurology clinics, neuroimaging researchers/clinician-scientists, and machine-learning researchers assessing model performance in neurodegenerative classification.

Introduction:
Early, accurate Alzheimer's disease (AD) detection using structural MRI-scans is integral for diagnosis and evaluating prognosis. The efficacy of analyzing the scans using learning models is uncertain in small neuroimaging datasets. Convolutional Neural Networks (CNNs) are the standard for medical imaging because of their efficiency and strong inductive bias for local feature extraction. Contrarily, Vision Transformers (ViTs) use self-attention to understand global context. This study aims to compare the accuracy of a CNN (ResNet-18) versus a transformer (ViT-Small) at distinguishing AD from a healthy brain control in small MRI datasets on accessible hardware.

Methods:
All 436 OASIS-1 open-source dataset patients were included, and each had a T1-weighted MRI and Clinical Dementia Rating (CDR) score: CDR = 0 (healthy) and CDR ≥ 0.5 (AD). Scans were resized to 224×224, intensity-normalized, and made into 2D axial slices. Data splits were by subject (70% train, 15% validation, 15% test) to avoid splice-level leakage. Models used PyTorch and were trained on an Apple M4 10-core GPU with AdamW, a cosine learning-rate scheduler, and cross-entropy loss. ResNet-18 trained for 20 epochs. ViT-Small, pretrained on ImageNet-1k, was calibrated for 30 epochs.

Results:
The average age was 77 for AD patients and 51 for healthy controls, with 100 AD and 336 healthy. The testing set had 55 patients and 48 slices each, n = 2640 slices. Both models trained stably and produced calibrated probabilities for AUROC; accuracy was defined in the testing set as percentage of correctly identified MRI scans. ResNet-18 achieved 81.7% accuracy and 0.87 AUROC, while ViT-Small had 73.9% accuracy and 0.71 AUROC (Fig 1). Models took about 8-10 minutes per epoch. DeLong's test (AUROC) and paired bootstrap resampling (accuracy) confirmed significant differences ($p < 0.05$).

Discussion/Conclusion:
ResNet-18 had higher AUROC and accuracy than ViT-Small with subject-level splits, providing evidence that in clinically applicable limited-data scenarios, CNNs remain better suited for neuroimaging tasks on consumer-grade hardware. Next steps include addressing limitations through multi-site validation (OASIS-3/ADNI), 3D slice evaluation and learning curve analysis.

[P65] Development and Pilot Evaluation of an AI-Enabled Mobile Application for Postoperative Recovery: The POP Study Protocol

Tanya Abenaim, Faculty of Medicine and Health Sciences, McGill University
Dan Poenaru, Faculty of Medicine and Health Sciences, McGill University

Low health literacy makes it difficult for many patients to understand and act on discharge instructions after ambulatory surgery, creating barriers to safe recovery. Mobile health (mHealth) applications powered by artificial intelligence (AI) have emerged as potential solutions, yet most face limited scalability due to clinician-dependent customization and monitoring. Meanwhile, patients increasingly use general-purpose AI chatbots for postoperative guidance, risking hallucinated or unsafe recommendations. Furthermore, no validated instruments currently assess postoperative self-management behaviours or AI use, constraining rigorous evaluation.
To address these gaps, we propose developing (1) a patient-reported experience measure (PREM) for postoperative self-management and AI utilisation, and (2) POP, an AI-enabled mHealth application that employs retrieval-augmented generation (RAG) with multi-agent validation to transform discharge instructions into plain-language recovery pathways without clinician burden.

We will conduct a three-phase sequential mixed-methods study. Phase 1 will develop and validate the PREM through scoping review, qualitative interviews, expert panel review, cognitive testing, and psychometric validation in adults undergoing ambulatory surgery. Phase 2 will employ participatory co-design with patients of diverse literacy and digital backgrounds, alongside multidisciplinary clinicians, to refine usability, safety, and workflow requirements. POP will use RAG grounded exclusively in patient-specific discharge instructions to convert multimodal inputs into structured recovery pathways. A three-agent validation pipeline will assess outputs: Agent 1 (extraction accuracy), Agent 2 (clinical safety), and Agent 3 (plain-language readability); flagged outputs will undergo review. Phase 3 will conduct a pilot hybrid type III implementation–effectiveness trial, enrolling adults undergoing elective ambulatory surgery and randomized 1:1 to standard discharge instructions with or without POP. Participants in the intervention arm will upload, photograph, or record instructions, receiving daily, gamified recovery guidance. Primary outcomes at 2 and 4 weeks will assess feasibility (recruitment, retention, engagement, fidelity); secondary outcomes will evaluate usability (System Usability Scale), acceptability (interviews), and efficacy compared to control via comprehension and PREM-assessed self-management behaviours.

This protocol integrates validated measurement with scalable AI design. The PREM will deliver the first standardized tool for evaluating postoperative self-management and AI use. By combining RAG, multi-agent validation, and participatory design, POP offers a safe, scalable, and patient-centred model for surgical aftercare.

[P66] A systematic review of ambient scribe use for physician documentation: A look at primary care

Dylan Grimm, Queen's University Department of Family Medicine

1. Background

Ambient scribes have emerged to reduce documentation burden from electronic medical records (EMRs). We conducted a PRISMA-guided systematic review of ambient scribe use in real clinical settings to identify evaluated outcomes and impacts, focusing on studies involving primary care clinicians (family medicine, internal medicine, pediatrics) and ambulatory providers.

2. Approach

A proximity search strategy was applied in Embase, MEDLINE, and Web of Science (Table 1). Two authors independently screened titles, abstracts, and full texts. We included studies where ambient scribes were deployed in real encounters with physicians, generating notes without prior editing and measuring specific outcomes. Exclusions were studies limited to non-physician providers, scripted or simulated settings, or pre-authored notes.

3. Results

Of 2,571 screened studies (Figure 1), 19 met inclusion criteria. Most used pre–post designs, with one randomized controlled trial. Outcomes centered on provider experience (stress, burnout, cognitive burden), productivity (time in EMRs, after-hours work), and note characteristics (length, quality). Quality assessment was conducted using the NHLBI Study Quality Assessment Tools. Findings included reduced after-hours EMR time, improved satisfaction and efficiency, decreased burnout, and shorter documentation times in some contexts (Table 2).

4. Conclusions

Initial evidence suggests that AI scribes can have beneficial impact on physician workflow, work-life balance without sacrificing note quality. However, evidence on ambient scribes remains limited. Most studies were conducted in a single country, restricting generalizability. Only one randomized trial exists, limiting certainty regarding true impact. Most outcomes studied relate to providers, with little evidence on patient-level effects. Future work should address methodological gaps, include diverse settings, and evaluate whether ambient scribes improve not only clinician workload but also patient experience and care quality.

[P67] Evaluating a Locally Deployed 20B-Parameter LLM for Automated Screening in Systematic Reviews

Paulo Henrique Moreira Melo, Faculty of Medicine and Health Sciences, McGill University
Elena Guadagno, Faculty of Medicine and Health Sciences, McGill University
Dan Poenaru, Faculty of Medicine and Health Sciences, McGill University

Background:
Recent advances in large language models (LLMs) have opened opportunities to automate evidence synthesis tasks such as literature screening and data extraction. However, in systematic review (SR) workflows, independent screening by two human reviewers is still regarded as the methodological benchmark, and AI-based automation remains experimental. Large language model (LLM) reliability depends heavily on prompt design. Because SRs prioritize sensitivity, ensuring no relevant studies are excluded, prompt wording must carefully balance sensitivity and specificity. We explore here the use of LLMs to assist human screening of literature abstracts in the evidence review of healthcare publications.

Methods:
We deployed gpsoss:20b locally. Each study's title, abstract, and the predefined inclusion/exclusion criteria from a completed SR were provided as input. The model produced binary include/exclude decisions, which were benchmarked against the final inclusion list after full-text review. Two prompting strategies were compared: a conservative "when in doubt, exclude" setting and a sensitivity-oriented "when in doubt, include" setting. Performance metrics included sensitivity, specificity, precision, and overall accuracy.

Results:
A total of 3,352 abstracts were screened. The conservative configuration showed limited performance, with a sensitivity of 34%, as the model failed to include 21 of the 32 studies ultimately accepted in the SR. In contrast, the sensitivity-oriented configuration achieved a sensitivity of 94%, specificity of 98%, precision of 25%, and overall accuracy of 97%. During screening, it excluded 3,234 abstracts, of which only two (0.06%) were incorrectly rejected. Compared with human reviewers during the abstract-screening phase, the model also correctly excluded 39 abstracts that the 2 human reviewers had initially included, but later removed after full-text assessment, demonstrating its potential to reduce early over-inclusion. The time required for screening all abstracts was approximately 26 hours for each human reviewer and 5.58 hours by the AI (4.7-fold faster).

Conclusion:
Early evidence shows that our model demonstrated high sensitivity and specificity, supporting its use as a first-pass screening tool to accelerate systematic reviews while maintaining methodological rigor. In such a collaborative human-AI process, false positives would be resolved during subsequent human adjudication.

[P68] Evaluating End-User Feedback to Optimize AI Integration into the Operating Room

Ariana Walji, University Health Network and the University of Toronto
Caterina Masino, University Health Network
Wagner Souza, University Health Network
Stephanie Williams, University Health Network
Spencer Gable-Cook, University Health Network
Jimmy Qiu, University Health Network
Amin Madani, University Health Network

Introduction: Recent years have seen a surge in the development of intraoperative artificial intelligence (AI) guidance tools to support safe surgical decision-making and enhance surgical outcomes. Despite the excitement towards implementing these technologies, there is a paucity of literature examining human factors influencing AI adoption and successful translation into the operating room. This study addresses the gap by collecting end-user feedback on two AI deployment platforms through User Acceptance Testing. The findings aim to inform design and deployment decisions to improve integration and adoption into surgical workflows.

Methods: User Acceptance Testing was conducted to evaluate usability, ergonomics, and overall user experience of two AI deployment platforms (Cloud-based and Edge-based) in the operating room. The study involved 23 end-users (surgeons, residents, fellows, and operating room nurses) and was carried out in a high-fidelity simulated operating room. Participants tested the GoNoGoNet AI model on each platform as a use case. Usability of each platform was assessed using task completion rates and the System Usability Scale. Ergonomics was assessed for surgeons, fellows, and residents using the NASA Task Load Index and Borg CR-10 rating scales. Overall experience was gathered using open-ended survey responses. Wilcoxon Signed-Rank Test was used to assess significance for non-normally distributed data and a paired t-test for normally distributed data.

Results: Task completion rates revealed significantly greater usability for AI deployment with the Cloud platform ($p<0.001$). Conversely, System Usability Scores did not reveal significant differences. Significantly greater Physical and Frustration scores were reported with the Cloud platform ($p = 0.0004$ and 0.0036 respectively), as well as significantly more exertion reported in the neck region ($p=0.0067$). End-user feedback surveys revealed overall positive experiences across both platforms.

Discussion: Feedback from end-users on usability, ergonomics, and overall experience provides valuable insights for institutions aiming to identify the most suitable deployment infrastructure for their needs and to guide iterative improvements to each AI deployment platform. These insights can enhance user receptiveness and seamless incorporation of AI tools into routine surgical practice.

[P69] Membership Disclosure Evaluation for Synthetic Clinical Text Generated by Dynamic Few-Shot In-Context Learning

Sen Li, University of Ottawa
Khaled El Emam, University of Ottawa
Fida Dankar, CHEO Research Institute

Target Audience
Healthcare data custodians, clinical NLP researchers, privacy/auditing practitioners, and developers deploying LLMs for clinical text synthesis.

Introduction
LLMs are increasingly used to generate synthetic clinical notes for data sharing, but membership inference attacks (MIAs) can reveal whether specific real notes appeared in prompts or training data. Prior studies often examine static few-shot prompts, i.e., fixed exemplar notes that remain the same across queries. Our focus is on dynamic few-shot in-context learning (ICL), where the k exemplar notes in the prompt are selected on the fly for each generation rather than being fixed. Our goal is to examine whether dynamic few-shot prompting reduces membership disclosure compared with static few-shot prompting under this realistic, text-only auditing setting.

Methods
For each target adversarial sample, we compute a vector of sentence-level embedding similarities between the target sample and the released synthetic notes. Three similarity-based features—maximum similarity, mean similarity, and mean top-3 similarity—are derived for each target sample. Each of these features can then be used to cluster the adversarial set into members and non-members through unsupervised learning. We report F1 and an adjusted metric $M=(F1-Fmax)/(1-Fmax)$, where Fmax denotes the F1 score achieved by the naive adversarial strategy of classifying all records as members.

Results
Preliminary findings suggest: (1) dynamic ICL greatly lowers membership disclosure compared to static ICL (F1reduced from 0.89 to 0.44, M reduced from 0.8 to 0.01); (2) within dynamic ICL, 2–3 shots further reduce leakage (0.4% in F1, 0.6% in M) vs. 1-shot and (3) The maximum similarity feature yields the most robust membership classifier compared with the mean and top-3 similarity features.

Discussion/Conclusion
Dynamic few-shot ICL offers a practical mitigation for membership disclosure in synthetic clinical text, with modest shot increases (2–3) delivering additional benefit. Our text-only, model-agnostic auditing provides custodians with deployable tools and an interpretable metric to decide release readiness. Ongoing experiments will expand datasets, LLMs, and auxiliary data sizes, and assess robustness to domain drift and prompt formatting changes.