AI Health 2025

The 2nd International Annual Conference on AI in Health

10-11 November 2025 | Basel, Switzerland

Abstract Book

Bioevents
Sharing Biomed Knowledge

University of Applied Sciences and Arts Northwestern Switzerland
School of Life Sciences

AI Diagnostics: Clinical Decision Support

Experimental Evidence of Accurate Generalization in Quantum Machine Learning for Medical Image Classification

Leena Anthony[1]
School of Life Sciences, FHNW, Switzerland

Leena Anthony[1]*, Artemiy Burov[12], Clément Javerzac[1]
[1]School of Life Sciences, FHNW, Switzerland
[2]EPFL, Lausanne
*Corresponding author: leena.anthony@students.fhnw.ch

Background
Generalization from small datasets remains a key challenge in AI diagnostics, especially for medical imaging tasks where annotated data is limited [1]. Quantum machine learning (QML) offers a different inductive bias and may enable superior generalization with fewer parameters [2].

Objectives
We investigate whether quantum convolutional neural networks (QCNNs) can accurately classify breast ultrasound images using as few as 2–10 labeled examples, and compare them to a parameter-matched classical convolutional neural network (CNN).

Methods
We implemented a 6-qubit QCNN with 45 trainable parameters using Qiskit and IBM's FakeKyiv backend. Breast ultrasound images from BreastMNIST [3] were downsampled to 8×8 pixels and encoded using amplitude embedding. The circuit included mid-circuit measurement and classical control logic for pooling. Each training size (N = 2, 5, 10, 20, 40, 80) was repeated 20 times with different random seeds. A CNN with identical parameter count and input size served as baseline.

Results
The QCNN achieved over 95% accuracy across all settings, including 97% with just two samples and 99.8% ± 0.3% with ten. Generalization error remained below 0.02 and followed $1/\sqrt{N}$ scaling. The classical constrained to the same parameter count, performed worse in low-data regimes across training sizes and consistently achieved lower test accuracy overall. These results also aligned with recent QML theory on parameter reuse, sparsity, and scaling. All circuits executed within NISQ coherence constraints.

Conclusion
This study provides empirical evidence for recent generalization theory in QML [2], confirming that QCNNs can outperform classical CNNs in low-data settings. These results position QCNNs as promising tools for data-efficient clinical decision support, especially where labeled training data is scarce.

References

[1] Huang SC et al., npj Digital Medicine, 6(1):74, 2023.
[2] Caro MC et al., Nature Communications, 13(1):4919, 2022.
[3] https://medmnist.com

P-02

AI Diagnostics: Clinical Decision Support

OvarIA: A Novel System for Hybrid Screening for Ovarian Cancer

Ana Cristina Castro Castro[1]
Sugery and Cancer Research Center (CICICA-UCR), University of Costa Rica, Costa Rica

Background
Ovarian cancer is a major public health concern. A significant number of women are diagnosed with ovarian cancer each year, in advanced stages, and with poor prognosis. Ovarian cancer survival rates are generally poor, even with recent advances in chemotherapeutical and surgical treatments. Novel, molecular, early detection protocols have given significantly better results than traditional approaches. The system we present here is based on similar protocols.

Objectives
We describe a system for the combination of cell-free DNA (cfDNA) analysis, biomarker determination (such as CA-125, AFP, CA 15-3, CA 19-9, CEA, CYFRA) and Machine Learning models to improve early detection of ovarian cancer. Single-signal detection models are prone to sensitivity and specificity shortcomings; moreover, they are likely to lack predictive power during the early stages of ovarian cancer.

Methods
This abstract describes the computational architecture we are employing to process and combine data from the varied sources we utilize. It is constituted by a data ingestion engine which captures and formats input data into a single data frame format. Downstream, data frames are analyzed and processed, and analysis results (either in summarized / textual or graphical form) are produced.

Results
The system we present here analyzes multiple signals from molecular sources and provides a risk assessment of likelihood of ovarian cancer. We compare it with current gold-standard approaches and provide detailed results composed of cfDNA, biomarker and Machine Learning combined analysis.

Conclusion
By combining both cfDNA- and biomarker-based signals, we expect to see a significant increase in sensitivity over single-signal methods. Moreover, we expect continuing to apply Machine Learning techniques to develop a unified screening model will result in a more robust early detection test panel.

P-03

AI Diagnostics: Clinical Decision Support

Morphological Fingerprinting in Predicting Neurological Disease; A Deep Learning Approach

Connor Dalby[1], Austin Dibble[1], Damiano Ferrari[2], Sergio Benini[2], Alessio Fracasso[1], Donald Lyall[1], Monika Harvey[1], Lars Muckli[1], Michele Svanera[1]
[1]*School of Psychology and Neuroscience, University of Glasgow, United Kingdom*
[2]*Department of Information Engineering, University of Brescia, Italy*

Background

Neurological disorders are leading causes of disability and mortality, yet accurate diagnosis remains challenging. Structural alterations detectable with structural-MRI, including volumetric (grey matter, white matter, ventricles etc) and surface (cortical thickness, surface area, curvature) features, can encode disease-specific "morphological fingerprints" with high diagnostic potential. Previous studies have often been limited by small cohorts, with a single disease or feature focus, and methodological shortcomings in widely used software for segmentation and cortical-surface reconstruction. Fast, accurate, and scalable approaches are needed to capture these features across diverse clinical populations and translate them into predictive fingerprints for earlier, reliable diagnosis.

Aims

We aimed to i) evaluate the predictive utility of morphological fingerprints across neurological conditions, ii) establish normative trajectories across lifespan, and iii) develop a fast and accurate AI method for extracting comprehensive MRI features.

Methods

We developed two AI models for cortical surface reconstruction, thickness estimation, and segmentation, enabling estimation of 13 morphological features from a single T1w scan in under one minute. Performance in speed and accuracy was benchmarked against the state-of-the-art FreeSurfer. The framework was tested on 100,000 scans from healthy individuals and patients across 18 neurological disorders at multiple sites.

Results

Morphological fingerprints reliably distinguished 18 neurological diseases at both group and individual levels, supported by sample sizes exceeding previous studies. We show that these fingerprints define disease-specific patterns of morphological change, capturing how each feature diverges from healthy. We further establish normative-trajectories for all features across the lifespan, providing population-level benchmarks. Finally, our framework outperforms FreeSurfer in both accuracy and inference speed, resolving longstanding inaccuracies in cortical segmentation and surface reconstruction.

Conclusion

This work delivers a comprehensive demonstration of morphological fingerprinting, spanning 18 diseases, 100,000 scans, and 13 structural features. These advances move morphological profiling towards scalable tools for earlier diagnosis and population-level monitoring of neurological disease.

P-04

AI Diagnostics: Clinical Decision Support

Algorithm Design and Validation for Early Identification of Severe Mental Disorders and Suicide Risk

Diego Diaz-Milanes[1,2], Francesa Bureca[3], Raquel Remesal-Cobreros[4,5]
[1]*Health Research Institute, University of Canberra, Australia*
[2]*Department of Quantitative Methods, Universidad Loyola Andaluca, Spain*
[3]*Pediatric Emergency and Emergency Department, I.R.C.C.S. Maternal and Child Health Institute Burlo Garofolo, Italy*
[4]*Department of Clinical and Experimental Psychology, University of Huelva, Spain*
[5]*Mental Health Service, Juan Ramon Jimenez Hospital, Spain*

Background: Severe mental disorders and suicide risk represent one of the main health challenges of the 21st century, due to their high prevalence, personal suffering, family impact, and substantial economic cost. Although multiple psychometric instruments exist for assessment, their isolated predictive capacity is limited. Moreover, there is insufficient availability of trained specialists to conduct clinical evaluations in primary care, which hinders early detection in community settings. Objectives: (i) To identify the psychometric instruments with the greatest discriminative capacity among healthy individuals, patients with a severe mental disorder, and patients at suicide risk; (ii) to develop an effective and efficient predictive algorithm based on these measures, implementable by healthcare professionals in primary care. Methods: The sample included 175 participants from Spain (49.14% men; age range: 18–76 years), distributed into three groups: control, clinical, and clinical with suicide risk. A random forest model was applied for feature selection, and the most relevant variables were used to build a CART decision tree, with the Gini index as the splitting criterion. Model performance was evaluated both globally and by category through cross-validation. Results: The model identified depression, openness to experience, and psychoticism as key predictors. The decision tree showed adequate performance, with sensitivities up to .83, specificities above .75 across all groups, and AUCs ranging from .65 to .83. Overall, the model achieved an accuracy of .68 and a Cohen's kappa of .52, indicating moderate agreement. Conclusions: The developed algorithm constitutes a rapid tool, easily interpretable by community health professionals and suitable for implementation in healthcare services, with strong potential for application in primary care. Its use may optimize referrals to mental health services, reduce healthcare burden, and contribute to the early detection of suicide risk and severe disorders.

P-05

AI Diagnostics: Clinical Decision Support

AI-Supported 3D Reconstruction for Comprehensive Pediatric Nerve Assessment in Neuromuscular Disorders

Marc Robin Gruener[1], Kenan Bektas[2,3], Justus Marquetand[4,5], Simon Mayer[3], Oliver Roehrle[4], Pengfei Zhang[4], Philip J. Broser[1,6]

[1]*Neurophysiology, Children's Hospital of Eastern Switzerland, Switzerland*
[2]*Quality and Risk Management, HOCH Health Ostschweiz, Switzerland*
[3]*Institute of Computer Science, University of St. Gallen, Switzerland*
[4]*Institute for Modelling and Simulation of Biomechanical Systems, University of Stuttgart, Germany*
[5]*Neurology, University of Tuebingen, Germany*
[6]*Medical Faculty, University of Basel, Switzerland*

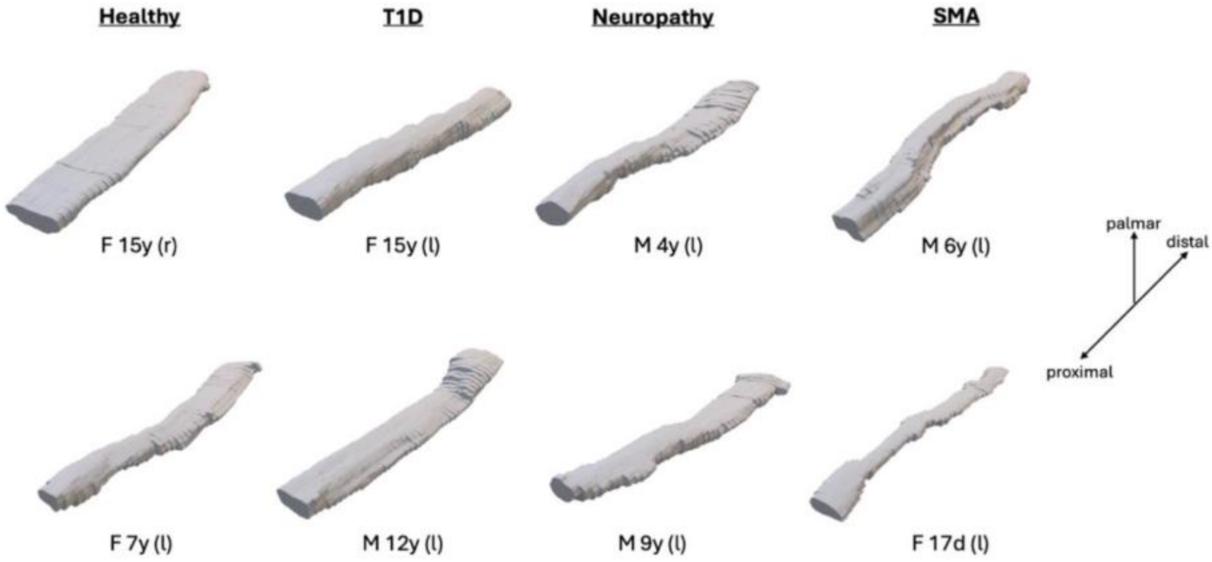
Background: Current pediatric peripheral nerve ultrasound relies on cross-sectional area measurements at one to three predetermined locations, providing limited morphological insight and failing to detect heterogeneous structural alterations along nerve trajectories in neuromuscular disorders like spinal muscular atrophy (SMA) and type 1 diabetes (T1D).

Objectives: To develop an AI-supported methodology for comprehensive 3D median nerve assessment that enables detection of disorder-specific morphological patterns and enhanced monitoring of structural changes in healthy individuals and those with neuromuscular disorders. Additionally, we plan to integrate automated robotic acquisition systems for improved standardization.

Methods: Ultrasound video acquisition was integrated with the "Segment Anything Model 2" (SAM2) deep-learning algorithm for semi-supervised frame-by-frame nerve segmentation. Using standardized 3D-printed positioning guides, we captured ultrasound video sequences and implemented physician-guided segmentation with minimal annotation requirements. In collaboration with the University of Stuttgart, we are enhancing an existing automated robotic scanning system to further standardize data collection protocols. Our approach generates detailed 3D reconstructions enabling measurement of novel metrics including shape and path characteristics (e.g. tortuosity, eccentricity, and centroid drift) along complete nerve segments. We have acquired more than 30 standardized ultrasound videos from healthy children and patients with T1D, SMA, neuropathies, and other neuromuscular disorders. Expert annotation is currently in progress.

Results: Preliminary validation demonstrates high concordance between SAM2-based segmentation and expert physician annotations with significant reductions in annotation time. The 3D reconstruction enables the visualization and quantification of morphological metrics unavailable with conventional point measurements. Initial results reveal distinct morphological pattern variations between healthy subjects and those with neuromuscular conditions.

Conclusion: SAM2-based 3D nerve reconstruction transforms pediatric neuromuscular assessment from limited point measurements to comprehensive morphological evaluation, enhancing pathological change detection while remaining feasible for routine clinical practice.


Healthy
T1D
Neuropathy
SMA
F 15y (r)
F 15y (l)
M 4y (l)
M 6y (l)
palmar
distal
proximal
F 7y (l)
M 12y (l)
M 9y (l)
F 17d (l)

P-06

AI Diagnostics: Clinical Decision Support

Multiparametric MRI and Machine Learning for Predicting Therapeutic Response to Anti-TNF Treatment in Rheumatoid Arthritis

Andreas Hess[1,2], Jürgen Rech[3], Arnd Dörfler[1], Georg Schett[3], Mageshwar Selvakumar[2]
[1]*Neuroradiology, UKER Erlangen, Germany*
[2]*I. f. Pharmacology, FAU Erlangen Nuernberg, Germany*
[3]*Department of Medicine 3 Rheumatology and Immunology, UKER Erlangen, Germany*

Background

Rheumatoid arthritis (RA) is a chronic autoimmune disorder causing severe morbidity and increased mortality. For decades, TNF-alpha is recognized as a key cytokine in RA, with direct effects on nociceptors and the central nervous system (CNS). Despite the efficacy of TNF inhibitors (TNFi), a significant number of patients do not respond to treatment.

Objectives

This study evaluated whether pre-treatment MRI metrics, analyzed using machine learning (ML) techniques, can predict clinical response to the TNF inhibitor Certolizumab -Pegol (CZP) in RA patients.

Methods

In this phase 3, multicenter, randomized, double-blind, placebo-controlled parallel-group trial, DMARD-refractory RA patients were stratified by baseline fMRI responses to painful finger joint compression (high [HV] or low [LV] voxel activity). Patients were randomized in a 2:1 ratio to receive methotrexate (MTX) plus CZP or MTX plus placebo (PL). The primary endpoint was achievement of low disease activity (LDA; DAS28 $\leq$3.2) at week 12. High-dimensional MRI data, including functional and anatomical measures, were subjected to dimensionality reduction (t-SNE) followed by supervised ML models to classify responders vs non-responders. Model accuracy was estimated using cross-validation.

Results

Out of 156 screened, 139 patients with moderate-high disease activity were enrolled and randomized. At week 12, LDA rates were 59% in HV_CZP, 42% in LV_CZP, and 26% in the PL group. Among MRI parameters, baseline pain-related fMRI activity yielded the strongest discrimination between responders and non-responders as demonstrated by a hierarchical multi-parametric t-SNE decomposition approach; ML models achieved up to 95.2% cross-validated accuracy in classifying clinical response.

Conclusion

Baseline pain-related fMRI combined with ML demonstrated strong potential as a predictive biomarker for response to CZP in RA, supporting the rational development of neuroimaging-based stratification strategies. This data-driven approach may pave the way for a more personalized and targeted therapy strategy.

Trial registration

Funded as IIT from UCB; EudraCT: 2013-000337-13; ClinicalTrials.gov: NCT01864265

P-07

AI Diagnostics: Clinical Decision Support

Leveraging AI Diagnostics to Improve Maternal Health Outcomes in Burundi Through Clinical Decision Support (CDS)

Salut Ndayishimiye[1]
Gyneco-obstetrics, Cibitoke Hospital, Burundi

Background

Maternal mortality is still a big healthcare issue in Burundi because of limited access to timely and precise diagnostics, highly skilled healthcare providers, and top-notch treatment options significantly contribute to poor maternal outcomes. Many pregnant women lack access to affordable and accurate maternal care, leading to complications and deaths. Artificial Intelligence (AI)-driven clinical decision support systems present a promising solution to bridge this healthcare gap.

Objectives

This study aims to explore the potential of AI diagnostics and CDS tools to enhance clinical decision-making in Maternofetal healthcare, particularly in rural and under-resourced settings in Burundi. The focus is on improving early detection of pregnancy-related risks, streamlining referrals, and reducing diagnostic delays.

Methods

A mixed-method review of AI clinical decision support interventions was conducted in maternal health across similar low-resource settings. Using available databases and clinical scenarios, we created a prototype AI model for risk stratification of pregnant women. This consists of patient history, basic lab results, and vital signs to predict high-risk pregnancies and recommend timely interventions.

Results

Preliminary simulations show that the AI model identified high-risk pregnancies with an accuracy of 87%, enabling early referral and preventive action. Clinicians reported improved confidence in decision-making. Cost analysis also shows potential for significant savings through early intervention and reduced complications.

Conclusion

AI-driven clinical decision support tools can be transformative in addressing maternal health challenges in Burundi. With proper training, integration into existing systems, and local data refinement, AI diagnostics can empower health practitioners, reduce maternal mortality, and make quality care more accessible and affordable.

P-08

AI Diagnostics: Clinical Decision Support

From Law to Clinic: The AI Act's Transparency and Explainability Requirements for AI Diagnostics

Syed Zulkifil Haider Shah
Department of Innovation and Digitalisation in Law, University of Vienna, Austria

The EU Artificial Intelligence Act (AI Act), in force since August 2024, classifies clinical decision support systems as high-risk AI. This creates binding obligations for developers, manufacturers, clinicians, and healthcare providers. Systems used in diagnosis, treatment planning, or prognosis must meet strict requirements of transparency, explainability, and human oversight to be legally deployed in the EU.

This presentation examines how these obligations apply in practice to clinical AI diagnostics. It explains what compliance means for manufacturers, how clinicians can integrate AI responsibly into patient care, and how patients can make sense of AI-driven decisions and exercise their rights. The focus is on translating complex legal provisions into concrete guidance for real-world healthcare.

The analysis is grounded in the EU-funded AISN project (2022–2026), which is building and validating an integrated AI platform for post-stroke rehabilitation. By combining data acquisition, clinical interpretation, brain simulation, intervention delivery, and predictive modeling, AISN provides a useful case stufy to understand how transparency and explainability requirements under the AI Act can be operationalised and embedded into system design and clinical use.

The presentation, thus, contributes timely insight into how the AI Act will shape clinical decision support, offering practical guidance for stakeholders and highlighting its impact on the future of medical care. By applying legal obligations to lived examples from clinical care, the presentation highlights both the opportunities and responsibilities arising under the AI Act, and shows how clinical AI can be introduced in a safe, transparent, and accountable way.

P-09

AI Diagnostics: Clinical Decision Support

SPOCK: Sparse-to-Precise Organs-at-Risk Contouring with Prior Knowledge for Pancreatic MR-Linac Radiotherapy with a Limited Dataset

Mehdi Shojaei[1], Bjoern Eiben[1], Jamie McClelland[3], Simeon Nill[1], Alex Dunlop[1], Arabella Hunt[2], Brian Ng-Cheng-Hin[2], Uwe Oelfke[1]

[1]*Joint Department of Physics, Institute of Cancer Research and The Royal Marsden NHS Foundation Trust, United Kingdom*

[2]*Radiotherapy, The Royal Marsden NHS Foundation Trust, United Kingdom*

[3]*UCL Hawkes Institute and Dept. of Medical Physics and Biomedical Engineering, University College London, United Kingdom*

Background: Manual Organs-at-risk (OAR) contouring for pancreatic cancer on Unity MR-Linac balanced 3DVane images can take hours. Furthermore, conventional deep-learning solutions are hindered by limited datasets, necessitating data-specific methods.

Objectives: Develop and evaluate SPOCK (Sparse-to-Precise Organs-at-risk Contouring with prior Knowledge): a method to generate full precise contours from sparse approximate contours using a conditional training approach.

Methods: Nine OARs (small bowel, duodenum, both kidneys, large bowel, liver, spinal canal, spleen, stomach) were delineated on 38 3DVane images from 14 patients(7 patients/23 images for 5-fold cross-validation (manual contours); 7 patients/15 images for testing (model outputs edited to clinical acceptability)). Additionally, two oncology consultants provided sparse, approximate contours for each OAR, roughly outlining a small number of slices per organ. This number was determined on the validation set by slice-count analysis(2-40 slices) at which performance plateaued using Dice similarity coefficient(DSC) and average surface distance (ASD)(ΔDSC0.02; ΔASD1mm). A conditioned nnU-Net was trained to predict full precise contours from a paired input image and sparse contours. During training, sparse inputs were simulated with random dilations/erosions, slice removal, and our structure-guided deformation-based augmentation(sgDefAug), with gaps filled by morphology-based interpolation. For testing, outputs were edited, if needed, to clinical acceptability. Performance was evaluated with DSC, surface DSC(sDSC), ASD, and 95th percentile Hausdorff distance(HD95). Consultants were timed for sparse contouring and output editing.

Results: Training took 14h; inference was 30s. Cross-validation yielded DSC 0.94±0.04 and ASD 1.34±0.44mm. On the test, DSC was 0.97±0.05, sDSC 0.96±0.06, ASD 0.85±1.14mm, and HD95 3.66±4.96mm. Slice-count analysis indicated ≤5 slices sufficed per OAR. Sparse contouring took 10min, editing ~23min, and total ~33min versus 8h for full annotations. As contextual benchmarking, SPOCK outperformed TotalSegmentator MRI (DSC 0.72±0.16, sDSC 0.66±0.16, ASD 6.84±7.49mm, and HD95 18.36±15.73mm).

Conclusion: By converting a handful of sparsely-contoured slices into precise OARs contours in minutes, despite limited data, SPOCK shortens contouring workflows, thereby streamlining treatment.

Table 1. Time required for sparse contouring and subsequent editing of SPOCK outputs across nine OARs) in seven test patients, performed by two experienced oncology consultants. The columns report the mean ± std times for sparse contouring, editing, and overall. Most OARs could be sparsely contoured and edited in under 5 minutes.

Organs	Sparse Contouring Time (s)	Editing Time (s)	Overall Required Time (min:s)
Duodenum	126.07 ± 8.92	192.2 ± 89.52	05:18 ± 01:30
L. Kidney	33.45 ± 5.58	7.1 ± 14.40	00:41 ± 00:15
R. Kidney	24.83 ± 2.58	25.9 ± 23.65	00:51 ± 00:24
L. Bowel	146.44 ± 10.53	331.1 ± 208.54	07:58 ± 03:29
Liver	76.43 ± 5.57	150.1 ± 192.05	03:47 ± 03:12
S. Bowel	101.88 ± 11.33	547.3 ± 288.62	10:49 ± 04:49
Spinal Canal	30.87 ± 3.62	4.6 ± 18.4	00:35 ± 00:19
Spleen	36.705 ± 3.83	40.5 ± 121.71	01:17 ± 02:02
Stomach	88.05 ± 12.87	55.8 ± 55.13	02:24 ± 00:57
Total	596.90 ± 42.48	1354.6 ± 173.28	32:31 ± 02:58

Table 2. Geometric accuracy of SPOCK and TotalSegmentator MRI across nine abdominal OARs for 7 test patients. The ground truth for SPOCK was defined as the clinician-edited contours, whereas TotalSegmentator MRI was evaluated against the same reference. This means the two approaches are not directly comparable; however, the results provide context for how state-of-the-art methods perform on Unity MR-Linac datasets. Despite the inherent similarity between unedited and edited SPOCK outputs, the performance gap between SPOCK and TotalSegmentator MRI is substantial. For each metric, better performance is highlighted in bold. TotalSegmentator failed to accurately contour the flexible OARs such as bowels, and even missed some OARs in some patients over all slices.

Organ	DSC		sDSC		ASD [mm]		HD95 [mm]	
	TotalSegMR	SPOCK	TotalSegMR	SPOCK	TotalSegMR	SPOCK	TotalSegMR	SPOCK
Duodenum	0.35 ± 0.19	**0.96 ± 0.04**	0.29 ± 0.18	**0.96 ± 0.04**	16.80 ± 8.72	**0.57 ± 0.56**	47.28 ± 20.45	**2.76 ± 2.68**
L. Kidney	0.84 ± 0.02	**0.99 ± 0.01**	0.82 ± 0.06	**0.99 ± 0.00**	2.08 ± 0.47	**0.13 ± 0.25**	8.05 ± 2.13	**0.43 ± 0.72**
R. Kidney	0.86 ± 0.02	**0.99 ± 0.00**	0.85 ± 0.04	**0.99 ± 0.01**	1.83 ± 0.27	**0.09 ± 0.15**	6.90 ± 2.23	**0.53 ± 1.31**
Large Bowel	0.59 ± 0.26	**0.91 ± 0.13**	0.48 ± 0.30	**0.89 ± 0.15**	9.18 ± 7.22	**2.14 ± 3.26**	24.53 ± 16.87	**10.42 ± 15.98**
Liver	0.91 ± 0.01	**0.99 ± 0.01**	0.77 ± 0.03	**0.98 ± 0.02**	2.57 ± 0.48	**0.36 ± 0.34**	9.64 ± 6.49	**2.34 ± 2.60**
Small Bowel	0.51 ± 0.30	**0.92 ± 0.10**	0.41 ± 0.26	**0.89 ± 0.13**	11.76 ± 10.45	**2.13 ± 2.46**	38.00 ± 29.99	**1.00 ± 13.39**
Spinal Canal	0.83 ± 0.23	**0.95 ± 0.10**	0.93 ± 0.25	**0.97 ± 0.07**	5.61 ± 18.58	**0.40 ± 0.77**	10.58 ± 33.79	**1.51 ± 3.39**
Spleen	0.80 ± 0.22	**0.99 ± 0.00**	0.70 ± 0.23	**0.97 ± 0.05**	7.83 ± 20.36	**1.03 ± 1.52**	14.44 ± 33.88	**0.95 ± 1.51**
Stomach	0.76 ± 0.16	**0.96 ± 0.04**	0.68 ± 0.13	**0.95 ± 0.06**	3.90 ± 2.91	**0.80 ± 0.99**	11.83 ± 7.76	**3.16 ± 3.04**
Mean	**0.72 ± 0.16**	**0.97 ± 0.05**	**0.66 ± 0.16**	**0.96 ± 0.06**	**6.84 ± 7.49**	**0.85 ± 1.14**	**18.36 ± 15.73**	**3.66 ± 4.96**

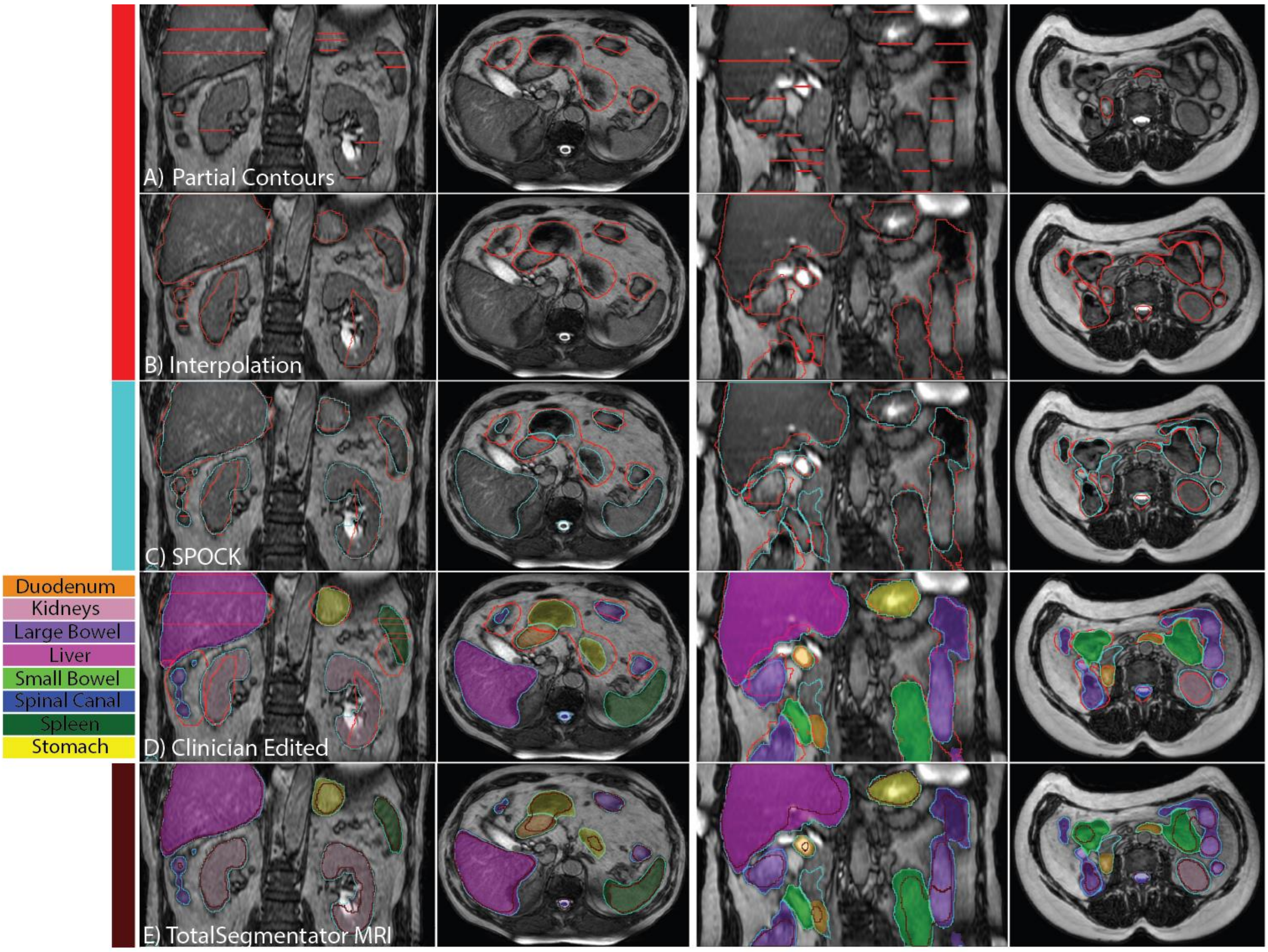


Figure 1. End-to-end workflow of SPOCK and qualitative comparison against clinician editing and TotalSegmentator MRI on Unity MR-Linac images for two representative test patients. Each row shows paired coronal (left) and axial (right) views; overlays use consistent colours across rows. Row A — sparse clinician input (red): Partial OAR contours drawn on some slices only, illustrating the limited annotations provided for conditioning (few slices per organ rather than full-FOV). Row B — morphology-based interslice interpolation (red): Dense contours obtained by interpolating directly from the sparse annotations, filling gaps between labelled slices. Row C — SPOCK prediction (cyan): Conditioned nnU-Net output given the MR image and sparse contours, showing refined, full-FOV OAR delineations that improve upon the interpolation results (Row B) and aligning well with OAR boundaries. Row D — Clinician-edited SPOCK contours (filled colours): Final ground truth used for evaluation, produced by expert editing of the SPOCK outputs to correct residual errors and standardise boundaries across all OARs. In both patients, edits relative to SPOCK predictions are small, typically subtle boundary refinements, indicating close agreement between SPOCK output and the final, edited contours. Row E — Contextual benchmark (burgundy): TotalSegmentator MRI overlaid for comparison, with performance assessed against the same clinician-edited contours from row D. In these examples, TotalSegmentator misses many boundaries and, in places, entire OARs, for example, the large bowel, resulting in visibly incomplete or imprecise structures. The right-hand patient has a single kidney, included to demonstrate robustness of the conditioned pipeline to asymmetric anatomy and organ absence.

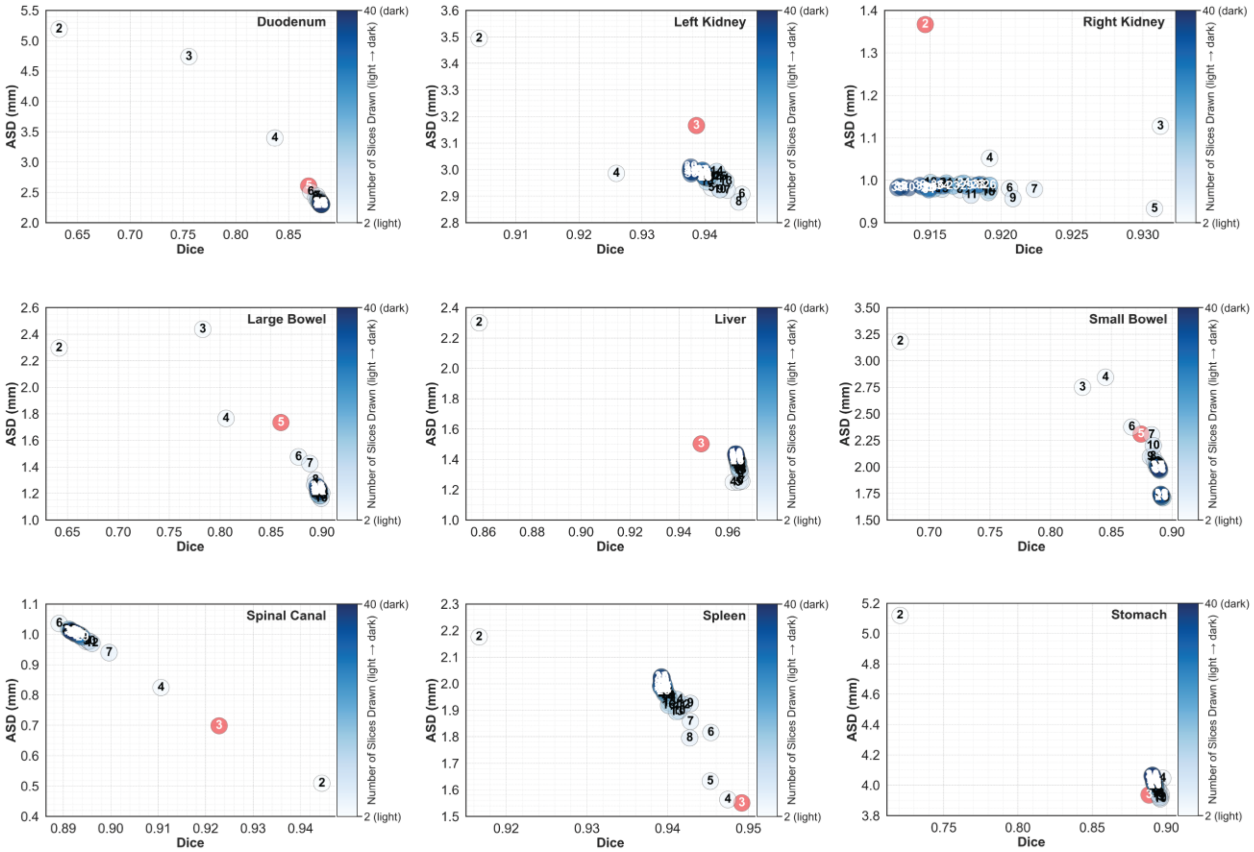


Figure 2. Effect of the number of contoured slices on geometric accuracy across nine OARs in the validation dataset. DSC (x-axis) and ASD (y-axis) are plotted for instances ranging from 2- 40 contoured slices. A stopping criterion was applied to identify the minimum number of slices beyond which further improvements were negligible, defined as ΔDSC < 0.02 and ΔASD < 1 mm within the following 10 instances. The corresponding threshold points are highlighted as red markers. The numbers inside each marker indicate the number of slices contoured, while the colour bar (right) provides a visual scale of slice counts from 2 (light) to 40 (dark).

0.02 and ΔASD 1 mm within the following 10 instances. The corresponding threshold points are highlighted as red markers. The numbers inside each marker indicate the number of slices contoured, while the colour bar (right) provides a visual scale of slice counts from 2 (light) to 40 (dark)." width="1030" height="872" /

AI Diagnostics: Clinical Decision Support

Rethinking Completeness: AI-Driven Patient Stratification in Complete Spinal Cord Injury

Olga Taran[1,2], Catherine Jutzeler[1,2], Sarah Brueningk[3]
[1]*Department of Health Sciences and Technology, ETH Zurich, Switzerland*
[2]*SIB, Swiss Institute of Bioinformatics, Switzerland*
[3]*Department of Radiation Oncology, Inselspital, Bern University Hospital and University of Bern, Switzerland*

Background: Spinal cord injury (SCI) is a rare but devastating condition, characterized by substantial heterogeneity in severity and recovery outcomes. Effective patient stratification, i.e., grouping patients by initial severity or expected recovery patterns, is essential for prognosis, treatment planning, and clinical trial design. However, patients with complete SCI, defined by the absence of sensory or motor function in the sacral segments, pose many challenges. While they appear uniformly impaired at baseline, some later transition to incomplete status and follow diverse recovery trajectories. This hidden heterogeneity motivates AI approaches to provide additional insights into clinically meaningful subgroups.

Objectives:

- Investigate whether data-driven methods can uncover meaningful subgroups among patients with complete SCI.
- To explore whether these subgroups align with distinct functional recovery patterns.

Methods: Data from the European Multicenter Study about Spinal Cord Injury (EMSCI) is used, with neurological assessments based on the International Standards for Neurological Classification of SCI serving as input features for the analysis. Two approaches are studied: classical clustering approaches with predefined numbers of clusters (e.g., k-means, Leiden algorithm), and an unsupervised deep clustering approach that infers subgroup structure directly from the data without any prior assumptions about the number of clusters.

Results: Unsupervised deep clustering identified three clinically meaningful subgroups with distinct recovery patterns not apparent from baseline classification. Tetraplegic patients showed moderate potential for improvement in both upper and lower limbs. Among paraplegic patients, one subgroup had minimal recovery prospects, while the other showed substantial lower limb improvement. These findings reveal hidden heterogeneity and provide actionable insights for prognosis and treatment planning.

Conclusion: Findings highlight the need for more refined patient stratification within the complete SCI population. The observed variability challenges the current definition of "completeness". Incorporating AI-driven stratification into clinical practice could support more personalized treatment strategies and improve the design of interventional studies.

AI Diagnostics: Clinical Decision Support

Beyond the Black Box: Policy Frameworks for AI Explainability in High-Stakes Medical Decisions

Sandya Venugopal[1]
Global Healthcare and Biomedical Consultancy, Venugopals Consulting Services, USA

Background: Current AI diagnostic systems often operate as "black boxes," creating fundamental tension between performance and interpretability in clinical settings. Common presentations like headache, chest discomfort, and pediatric bruising can range from benign to life-threatening, making AI explainability crucial for clinical decision-making. While regulatory bodies worldwide are developing frameworks for AI accountability, existing policies inadequately address the explainability-performance tradeoff in high-stakes medical decisions.

Objectives: To analyze current policy approaches to AI explainability in healthcare across major regulatory jurisdictions and develop a practical framework for balancing algorithmic transparency with clinical efficacy using real-world diagnostic scenarios.

Methods: Comparative policy analysis of regulatory frameworks (FDA guidance, UK MHRA guidelines, EU AI Act) applied to three hypothetical clinical cases: neurological presentation (headache), cardiovascular presentation (chest discomfort), and pediatric oncological presentation (bruising in a child). Each case scenario was designed to test different aspects of explainability requirements, examining how current regulatory frameworks would address transparency needs across varying clinical contexts, time pressures, and patient populations.

Results: Current policies show significant gaps in addressing context-dependent explainability needs across different clinical scenarios and patient populations. Analysis reveals three distinct regulatory approaches with varying effectiveness depending on diagnostic complexity, time sensitivity, and patient vulnerability. A risk-stratified explainability framework is proposed that adjusts transparency requirements based on clinical context, differential diagnosis severity, and available human oversight.

Conclusion: A tiered policy approach to AI explainability can preserve innovation while ensuring appropriate accountability across diverse clinical presentations. This framework provides actionable guidance for regulators and healthcare systems navigating the transparency-performance tradeoff in AI diagnostic implementation.

P-12

AI Diagnostics: Clinical Decision Support

Evaluating Host Gene Expression Severity Signatures Against the Phoenix Criteria for Pediatric Sepsis

Aspasia Vozi[1,2,3], Lachlan Coin[4,5], Catherine Jutzeler[1,2], Kristen Gibbons[6], Luregn J Schlapbach[3,6], RAPIDS Study Group[4,5,6]
[1]*Department of Health Sciences and Technology (D-HEST), ETH Zurich, Switzerland*
[2]*SIB, Swiss Institute of Bioinformatics, Switzerland*
[3]*Department of Intensive Care and Neonatology, and Children's Research Centre, University Children's Hospital Zurich, University of Zurich, Zurich, Switzerland*
[4]*Department of Microbiology and Immunology, University of Melbourne at The Peter Doherty Institute for Infection and Immunity, Australia*
[5]*Department of Clinical Pathology, University of Melbourne, Australia*
[6]*Children's Intensive Care Research Program, Child Health Research Centre, The University of Queensland, Australia*

Background and Aims: The Phoenix Sepsis Score (PSS) operationalizes life-threatening organ dysfunction in children. There is a need for reliable biomarkers predicting disease severity. This study assesses the performance of gene expression signatures in identifying children with sepsis based on PSS.
Methods: Transcriptomic data from 907 children evaluated for infection/sepsis were analyzed from the prospective RAPIDS multicenter cohort study in Queensland, Australia. A cohort-derived 10-gene severity signature was assessed for identifying patients with PSS≥2 within 24 hours of sampling. Performance was evaluated using the area under the receiver operating characteristic curve (AUC) with 95% confidence intervals (CI) and compared existing signatures.
Results: The 10-gene signature achieved AUCs of 0.93 (CI: 0.90-0.96) and 0.86 (CI: 0.79-0.92) in predicting patients with PSS≥2 in the discovery (n=595) and validation (n=312) cohorts, respectively. Other previously published signatures achieved AUCs of 0.83-0.93 and 0.77-0.81 for the discovery and validation cohorts, respectively.
Conclusions: The 10-gene severity signature demonstrates high diagnostic performance in identifying children with sepsis based on the new Phoenix sepsis criteria. Future trials should explore whether severity biomarkers can provide actionable insights at the bedside.

P-13

AI Diagnostics: Clinical Decision Support

Proteomic Characterization of Dysregulated Host Response in Blood Culture-Proven Pediatric Sepsis

Aspasia Vozi[1,2,3], Philipp K A Agyeman[4], Christoph Berger[5], Eric Giannoni[6], Sarah C Brüningk[7], Catherine R Jutzeler[1,3], Luregn J Schlapbach[2], on behalf of the Swiss Pediatric Sepsis Study[2]
[1]*Department of Health Sciences and Technology (D-Hest), ETH Zurich, Switzerland*
[2]*Children's Intensive Care Research Program, Child Health Research Centre, University Children's Hospital Zurich, University of Zurich, Zurich, Switzerland*
[3]*SIB, Swiss Institute of Bioinformatics, Switzerland*
[4]*Department of Pediatrics, Inselspital, Bern University Hospital, University of Bern, Switzerland*
[5]*Division of Infectious Diseases and Hospital Epidemiology, Children's Research Center, University Children's Hospital Zurich, Switzerland*
[6]*Clinic of Neonatology, Department Woman-Mother-Child, Lausanne University Hospital and University of Lausanne, Switzerland*
[7]*Center for Artificial Intelligence in Medicine (CAIM), University of Basel, Switzerland*

Background: Sepsis is a leading cause of death in children. The dysregulated host response underlying sepsis remains poorly characterized, limiting development of predictive biomarkers and targeted therapies in children.

Methods: We analyzed blood samples from 378 children with blood culture proven bacterial infections using four Olink proteomic panels that measured 358 proteins. Blood samples were collected within 7 days of blood culture sampling. The primary outcome was presence of organ dysfunction (OD) (Phoenix Sepsis Score ≥2). Secondary outcomes included recovery from multiple OD by day 7 and mortality. Differential protein expression, network analysis, machine learning modeling for biomarker discovery, and clustering were employed.

Results: 245 children (66%) developed OD; among these, 178 (73%) recovered and 21 (6%) died. We identified 61 proteins differentially expressed between children with/without OD, 72 proteins between recovered/non-recovered patients, and 64 proteins between survivors/non-survivors. A 7-protein signature distinguished patients developing OD (AUC 0.89 [0.77–0.93]). Three proteomic phenotypes emerged: Phenotype 1 (25%) comprised older children with high mortality despite mild presentation, characterized by cellular homeostasis and chromatin-modifying proteins; Phenotype 2 (58%) included moderately older children with less severe presentation and better outcomes; Phenotype 3 (17%) consisted of young infants with severe respiratory-focused presentation, featuring lymphocyte chemotaxis and GPCR ligand-binding proteins.

Conclusions: This first large-scale targeted Olink proteomic study in pediatric sepsis demonstrates proteomics' potential for identifying disease mechanisms and enabling precision diagnostics, despite limitations including sample size, sampling timing, and lack of external validation.

P-14

AI Diagnostics: Clinical Decision Support

Enhancing Skin Diagnostics through Generative Reconstruction of Polarization Features

Sławomir Wilczyński[1], Beata Sarecka-Hujar[1], Paweł Popielski[2]
[1]*Department of Basic Biomedical Science, Medical University of Silesia in Katowice, Faculty of Pharmaceutical Sciences, Poland*
[2]*Institute of Biomedical Engineering, University of Silesia in Katowice, Faculty of Science and Technology, Poland*

Background: Polarized-light imaging is increasingly important in medical imaging, especially in dermatology and aesthetic medicine. It reveals key skin features—wrinkles, vascular changes, discolorations, micro-injuries, and early signs of infection—that remain invisible in standard non-polarized photographs.

Objectives: We aim to develop an AI tool that transforms standard photographs into images corresponding to cross-polarized exposure. The approach targets use in screening, clinical decision support, and treatment-progress analysis in aesthetic medicine.

Methods: We propose a conditional generative framework (pix2pix) designed to preserve anatomical structure while reconstructing polarization-specific skin details. We established a standardized workflow covering acquisition of non-polarized and polarized images, anonymization and quality control, image pairing, and semi-automatic cropping. To date, we have collected approximately 12,000 paired frames from 186 participants with diverse skin phototypes, enabling training of a stable prototype. A web-based demo service performs inference in under one second on a typical computer; on a GPU, processing time drops to tens of milliseconds.

Results: Preliminary evaluation shows high concordance of generated images with references (mean SSIM $\approx$ 0.89) and a clear improvement in the visibility of vascular structures and subtle discolorations compared with input photos. In pilot clinical tests, specialists reported better readability of fine skin features.

Conclusion: The proposed solution has the potential to broaden access to polarized imaging without expensive hardware and to substantially improve image-based diagnostics in clinical practice. Ongoing work focuses on expanding the dataset, generalizing to additional anatomical locations, and conducting formal clinical validation using standardized rating scales and inter-rater agreement.

P-15

AI Diagnostics: Clinical Decision Support

GenMasterAI: A Local Large Language Model–Based Agentic and Reasoning software for Democratizing Genetic Analysis

Jing Zhai[1,2], Sergey Oreshkov[1,2], Nelly Pitteloud[1,2], Federico Santoni[1,2]
[1]*Service of Endocrinology, Diabetology and Metabolism, Lausanne University Hospital, Switzerland*
[2]*Faculty of Biology and Medicine, University of Lausanne, Switzerland*

Background
Interpreting genetic variants remains a major challenge in clinical and research genomics, as it requires the integration of medical knowledge, diverse annotations, phenotype data, and standardized classification guidelines. Current approaches are often time-consuming, manual, and prone to error, underscoring the urgent need for improvement in the rapidly expanding field of medical genetics.
Objectives
We previously developed GenMasterTable, a desktop application for large-scale genomic data analysis. Building on this foundation, our goal is to integrate GenMasterAI, a privacy-preserving, local language model–powered agent that automates variant annotation, evidence retrieval, and interpretation, generating geneticist-ready reports on standard computers.
Methods
The agent processes data provided by standard pipelines, automatically extracting for each variant relevant genetic information along with in-silico predictions such as MAF, CADD, SpliceAI, REVEL, and AlphaMissense. This information is then integrated with the patient's phenotype to generate an overall evaluation of the variant's pathogenicity. To achieve this, quantized large language models fine-tuned with retrieval-augmented generation (RAG) and low-rank adaptation (LoRA), and deployed in GGUF format, were incorporated on standard laptops and PCs. We benchmarked ten quantized SLMs/LLMs using synthetic datasets to assess their performance across different metabolic diseases.
Results
GenMasterAI integrates quantized LLMs/SLMs with an agentic layer that queries ClinVar, PubMed, OMIM and InterVar APIs, applies ACMG criteria, and incorporates recent literature, together with user-provided phenotypes through an embedded prompt-based chatbot. The system generates structured variant reports, combining annotation, reasoning, and evidence retrieval. Large models provided the most accurate reasoning, while lighter models achieved competitive outputs with faster runtimes on local hardware. Smaller raw models underperformed in multi-source reasoning therefore ongoing work focuses on fine-tuning small models to improve usability on resource-limited devices.
Conclusion
GenMasterAI advances secure, off-cloud genomic analysis by enabling LLM/SLM-powered reasoning on geneticists'/clinicians' desktops. This approach balances accuracy, latency, and feasibility, offering a practical and privacy-preserving path toward democratizing genetic diagnostics.

AI Therapeutics: Drug Discovery and Development

A Simulation-based Evaluation Framework for Federated Learning in Drug Discovery under Non-IID Conditions

Yeonwoo Choi[1], Minseok Choi[1]
College of Electronics & Information, Kyung Hee University, South Korea

Background

Federated Learning (FL) is a pivotal technology in drug discovery, enabling partners like pharmaceutical companies and academic institutions to collaboratively train AI models without sharing proprietary data, as exemplified by the MELLODDY project. A major challenge for FL is statistical heterogeneity (non-IID data), which is common in these multi-partner collaborations. However, robust methodologies for simulating and evaluating FL platforms under such realistic conditions are lacking.

Objectives

This study aims to: (1) Propose a novel simulation-based framework to evaluate FL platforms for drug discovery. (2) Utilize this framework to demonstrate that the FedGF algorithm is more effective than the conventional Federated Averaging (FedAvg) in non-IID settings for ADMET prediction.

Methods

We simulated non-IID environments for 10 clients using three publicly available ADMET datasets (BBBP, Tox21, SIDER). Two data partitioning strategies were employed: (1) a K-medoids clustering approach based on Tanimoto distance between molecular fingerprints to create extreme non-IID scenarios, and (2) a Dirichlet distribution (α=0.1) to partition data based on an optimal number of clusters determined by the Elbow Method, creating tunable heterogeneity. The performance of FedGF was benchmarked against FedAvg using a Multi-Layer Perceptron (MLP) model. An IID setting (α=10) was used as a control.

Results

As hypothesized, FedAvg`s performance degraded significantly in non-IID settings. In contrast, FedGF proved highly robust, consistently outperforming FedAvg across both extreme and tunable non-IID scenarios. Notably, its performance in these challenging settings was often superior even to that of FedAvg in the ideal IID scenario, highlighting its practical advantages.

Conclusion

Our findings validate the proposed simulation framework as an effective tool for evaluating FL algorithms. We also demonstrate that FedGF is a more suitable algorithm for real-world drug discovery applications involving heterogeneous data, offering a path to more robust and accurate collaborative AI models, ultimately accelerating the pace of therapeutic innovation.

P-17

AI Therapeutics: Drug Discovery and Development

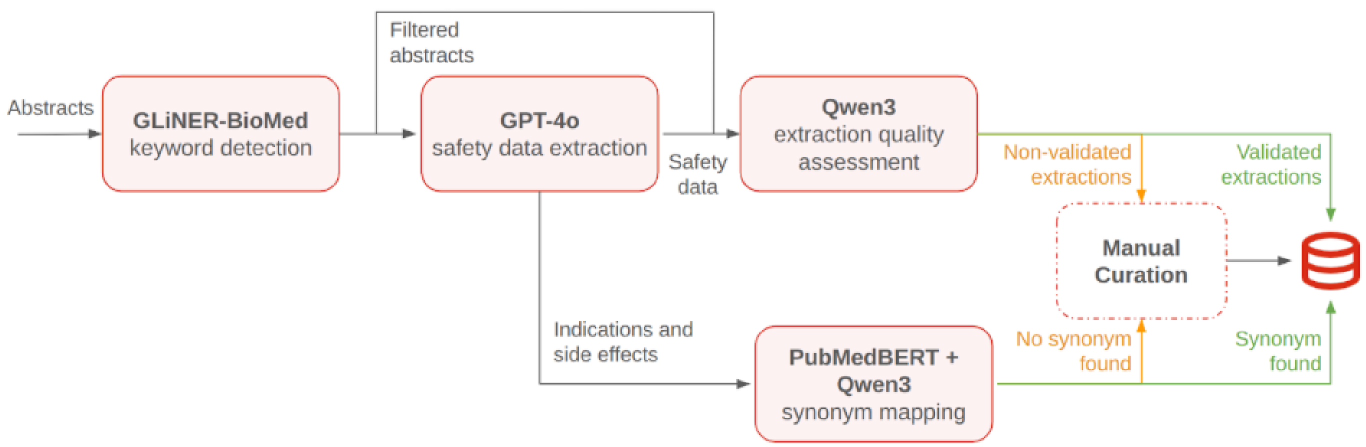
Generative AI-Based Sequential Multi-Model Pipeline for Extracting High-Quality Safety Data from Congress Abstracts

Maria Martinez-Sevillano[1,2], Nikola Panajotovikj[1,2], Mariona Isern[1], Ricard Garcia-Serna[1], Elisabet Gregori-Puigjane[1], Jordi Mestres[1,2]
[1]*Translational Sciences, Chemotargets, Spain*
[2]*Institut de Quimica Computacional i Catalisi, Universitat De Girona, Spain*

Background: Early assessment of a therapy's adverse events and efficacy is crucial for its long-term success. A key source of relevant insights on novel therapeutic approaches and their clinical evaluation are the abstracts presented at safety-related congresses. However, these data are highly unstructured and their extraction requires manual curation, which significantly delays clinical safety analyses and predictions.

Objectives: This project aims to (1) develop a generative AI-based approach to extract and validate structured safety data from congress abstracts and to (2) notably reduce the manual curation effort to build a reliable clinical evaluation database.

Methods: The pipeline leverages the GLiNER-BioMed model to filter the abstracts without mention of therapies or diseases. The remaining abstracts are processed with OpenAI's GPT-4o model to extract structured safety data from the text. The extractions are evaluated by a reasoning Qwen3 model to ensure accuracy and consistency. PubMedBERT and Qwen3 are used to map diseases to their synonyms. High-quality extractions are added to the database and the remainder are further assessed manually. This pipeline was validated against a set of 470 manually curated abstracts.



Results: The NER model step filtered the abstracts to further process by 59.4% with a recall of 94.4%. The OpenAI model discarded 65% of the remaining abstracts and retrieved structured safety data from the rest. After the data extraction and its corresponding quality assessment, 55.4% of the abstracts were considered high-quality, with an accuracy of 88.9%. 67.4% of the side effects were automatically mapped to their preferred term. Overall, the estimated manual curation time reduction was 93.8%.

Conclusions: The pipeline successfully retrieved high-quality safety data from relevant abstracts and classified the models' output by reliability. This allows to focus manual curation efforts on complex texts while automatically accepting validated extractions, notably expediting the data processing.

P-18

AI Therapeutics: Drug Discovery and Development

AI-based Identification of Virus-Infected Cells Using Label-Free Microscopy for advanced Drug Research

Juliane Pfeil[1], Corinna Siegmund[2,3], Eva Müller[3], Shakhnaz Akhmedova[1], Tobias Tertel[4], Alexandra Löwe[1], Anne Kauter[5], Michael Laue[5], Vu Thuy Khanh Le-Trilling[2,3], Sandra M. Bütow[1], Christian Sieben[6], Mirko Trilling[2,3], Roland Schwarzer[3], Nils Körber[1]

[1]*Centre for Artificial Intelligence in Public Health Research, Robert-Koch-Institute, Germany*

[2]*Institute for Virology, University of Duisburg-Essen, Germany*

[3]*Institute for the Research on HIV & AIDS-associated Diseases, University of Duisburg-Essen, Germany*

[4]*Institute for Transfusion Medicine, University of Duisburg-Essen, Germany*

[5]*Centre for Biological Threats and Special Pathogens, ZBS 4: Advanced Light and Electron Microscopy, Robert-Koch-Institute, Germany*

[6]*Nano Infection Biology, Helmholtz Centre for Infection Research, Germany*

Background

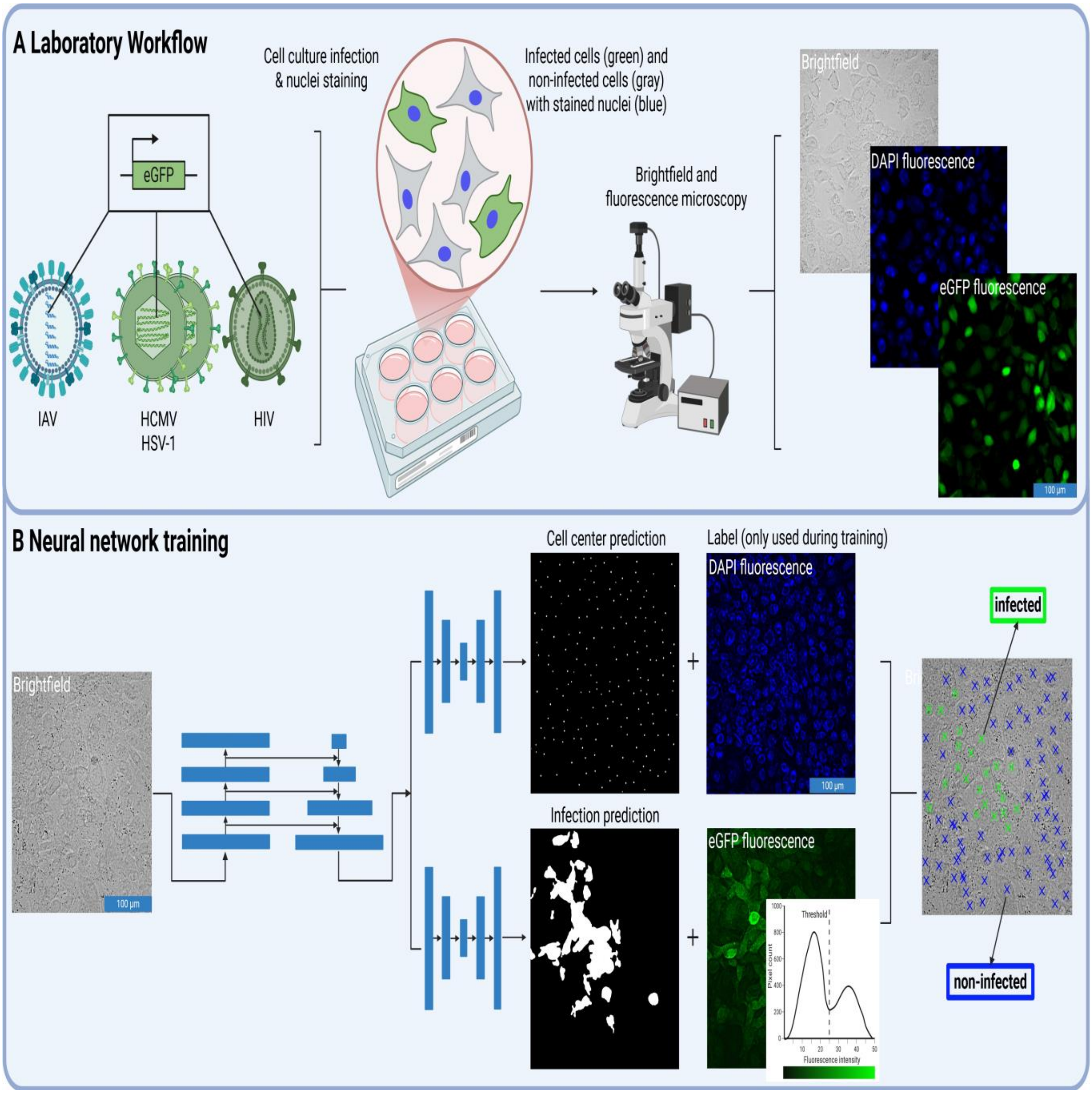
Identifying virus-infected cells usually involves observing cytopathic effects or using conventional assays like immunofluorescence staining. These methods are often subjective, time-consuming, and expensive. They also rely on specific antibodies and cell fixation, which makes them unsuitable for live-cell imaging and for characterizing new viruses. A precise method for identifying infected cells is critical for understanding viral dynamics and developing new therapies.

Objectives

The aim of this work is to introduce and validate a novel tool for the detection of virus-infected cells with artificial intelligence (AI) and microscopy (VAIruScope). By analyzing standard brightfield (BF) and phase contrast (PC) microscopy images, VAIruScope eliminates the need for fluorescent markers and other conventional labeling methods.

Methods

The process for VAIruScope`s development involved several steps. First, cells were infected with eGFP reporter viruses, and paired BF/PC and fluorescence images were captured (A). The eGFP signal served as the ground truth to identify infected cells. A transformer-based neural network was trained on this data to predict cell centers and classify infection status using only the label-free images (B).



Results

VAIruScope was tested on five cell lines infected with various viruses, including Influenza A, HCMV, HSV-1 and HIV. The tool achieved high classification accuracies, reaching up to 98 %. Its performance was also confirmed with a wild-type virus using electron microscopy. VAIruScope was successfully tested to measure drug-induced changes in viral infectivity, demonstrating its potential for use in drug discovery.

Conclusion

VAIruScope is a powerful, label-free method for identifying virus-infected cells using standard microscopy. It offers major advantages over traditional staining techniques by enabling precise, real-time cell identification without fluorescent labels. This tool is ideal for studying the dynamic processes of viral infection and has wide-ranging applications in accelerating drug screening, vaccine development, and research into viral pathogenesis.

P-19

AI Therapeutics: Drug Discovery and Development

Lightweight RAG Pipeline for Biomedical Research and Clinical Trial Question-Answering

Xialan Dong[1], Kaliel Wilson[1], Tracie Locklear[1], Ping-An Li[1], **Weifan Zheng**[1]
Department of Pharmaceutical Sciences, North Carolina Central University, USA

Background: Retrieval-augmented generation (RAG) improves large language model (LLM) responses by grounding answers in relevant source documents, which is especially valuable for navigating dense biomedical literature such as PubMed abstracts. *Objectives*: The goal of this work was to implement and evaluate a working RAG pipeline in Google Colab capable of reading biomedical and clinical corpora, embedding and indexing text, retrieving relevant context, and generating grounded answers. *Methods*: A plain text file of PubMed abstracts was uploaded, tokenized, and segmented into overlapping 512-token chunks using a HuggingFace tokenizer. Each chunk was embedded using a SentenceTransformer model and stored in a FAISS index. User queries were embedded with the same model, and the top-k most relevant chunks were retrieved. These chunks were concatenated into a contextual prompt, which was combined with the user query and passed to an LLM (e.g., a HuggingFace model) for answer generation. *Results*: The pipeline successfully created embeddings, built the FAISS index, and retrieved contextually relevant text passages. In a test case querying "stroke and ferroptosis," the system returned relevant chunks and produced a coherent answer aligned with the retrieved abstracts. *Conclusion*: This proof-of-concept demonstrates that a lightweight RAG pipeline can be effectively deployed for biomedical and clinical text exploration. The workflow is flexible, supporting both local HuggingFace models and OpenAI APIs, reproducible for domain-specific question-answering, and adaptable for research, teaching, and training applications, including clinical trial recruiter and participant Q&A.

P-20

AI Prospects: Digital Twins, Citizen Care, One Health

Generating and Evaluating Realistic Synthetic Healthcare Data: A Comprehensive Evaluation

Mieke L. M. Bentvelzen[1], Laura Jochim[2], Simon C. Mastbergen[1], Paco M.J. Welsing[1], Said el Bouhaddani[2]

[1]*Department of Rheumatology & Clinical Immunology, University Medical Center Utrecht, Netherlands*

[2]*Department of Data Science & Biostatistics, div. Julius Centrum, University Medical Center Utrecht, Netherlands*

Background

Healthcare data underpins biomarker discovery, personalised treatment, and drug development. However, access to data is hindered by high costs, as well as regulatory and logistical limitations. These challenges limit opportunities to validate and replicate findings or develop, train and test (machine-learning) models under realistic conditions. Synthetic data generation (SDG) is a promising solution. Yet, while various strategies exist to evaluate SDG performance on specific aspects, there is limited guidance on a holistic performance evaluation for healthcare data1,2. Moreover, most studies on tabular datasets evaluate large-N datasets with moderate/low-p, leaving small-sample, high-dimensional scenarios, common in healthcare settings, largely underexplored.

Objectives

This study evaluates and compares four representative SDG methods across healthcare datasets with varying sample sizes and dimensionalities. The aim is to provide practical guidance on the strengths and limitations of these methods in different data dimensionalities. Additionally, a comprehensive evaluation framework is proposed to systematically assess synthetic data quality.

Methods

We compared four representative tabular SDG methods: synthpop, vine copulas, conditional tabular GAN (ctGAN), and variational autoencoders (VAE) across three healthcare datasets: a large-sample, low-dimensional EHR dataset (N=9662,p=42); a moderate-sample, higher-dimensional TCGA breast cancer proteomic dataset (N=841,p=222); and a small-sample, higher-dimensional proteomics dataset from a phase III, multicentre, randomized clinical trial (TOFA-PREDICT; N=80,p=350). Performance was evaluated at univariate, bivariate, and multivariate levels (Table 1). The complete workflow will be implemented in an R package.

Results

Primary results showed that Synthpop excelled in large-N, low-p settings, while vine copula slightly outperformed in moderate-N, high-p scenarios; both effectively preserved distributions and correlations. ctGAN required extensive tuning and computational resources to achieve comparable utility.

Conclusion

Vine copula is promising for moderate-sample, high-dimensional data, and this may extend to more challenging small-sample, high-dimensional settings. Overall, our evaluation highlights the need for and provides practical guidance on systematically generating and evaluating realistic synthetic healthcare data.

Table 1. Metrics for univariate, bivariate, and multivariate evaluation of synthetic data

Dimension	Category	Metric / Method
Univariate	Descriptive statistics	Difference in mean Difference in variance Difference in skewness Difference in kurtosis
	Information measures	Wasserstein distance
	Hypothesis testing	Cramér–von Mises tests Anderson–Darling
	Predictive evaluation	Standardized propensity mean squared error (pMSE)
Bivariate	Correlation structure	Absolute Pearson's r difference Correlation heatmap (visualisation)
Multivariate	Classification based performance/evaluation	Logistic regression and Random Forest (AUC) (pMSE)
	Multivariate distribution distance	Maximum Mean Discrepancy
	Constraint violation detection	Multivariate anomaly detection
	Global structure	PCA

P-21

AI Prospects: Digital Twins, Citizen Care, One Health

Reconstruction of 3D Shapes from Magnetic Data for Orthopedic Orthoses using Deep Learning

Sarah Chouchene[1], Manon Lambert[1], Corentin Féry[2], Simon Lemoigne[2], Thomas Quirin[2], Joseph Lam-Weil[1], Joris Ravaglia[1], Ralph Schumacher[3], Joris Pascal[2], Morgan Madec[1]
[1]*Laboratoire ICube, UMR 7357 (University of Strasbourg / CNRS), France*
[2]*Institute for Medical Engineering and Medical Informatics, School of Life Sciences FHNW, Switzerland*
[3]*Bellwaldtec, Bellwaldtec, Switzerland*

Context:

The design of customized orthotics relies heavily on the accurate reconstruction of 3D shapes. Traditional methods require expensive optical scanners or complex manual measurements that can injure patients. A promising alternative is to use magnetic sensors positioned on bandages and stuck to the surface of the shapes, allowing for more accessible and portable data acquisition.

Objective:

Develop an artificial intelligence approach capable of reconstructing 3D point clouds from magnetic sensor data in order to generate models that can be used for 3D printing of orthopedic orthoses.

Methods:

A deep learning model is trained using supervised learning, with input from magnetic sensors position and output from optical scan considered as the references. Evaluation is performed on new shapes in order to measure the model`s generalization ability.

Results:

The model achieves an average distance error of 4.9 mm between reconstruction and reference on previously unseen shapes. However, the model does not generalize perfectly and still exhibits inconsistencies in some cases. For that, areas for improvement are in pipeline, such as adding parameters related to morphology and orientation, as well as combining AI with traditional rigid registration methods

Conclusion:

This study shows that it is possible to reconstruct 3D shapes from magnetic data using deep learning, with encouraging accuracy for a first step. The integration of additional constraints and hybrid approaches paves the way for functional and generalizable prototypes, including for shapes with complex deformations.

P-22

AI Prospects: Digital Twins, Citizen Care, One Health

A Novel Digital Twin Model to Analyze Effects of Pathogenic Variants in mTOR Pathway Genes

Valerio Conti[1], Sara Sommariva[2], Sara Pisini[1,3], Silvia Berra[4], Cristiana Pelorosso[1], Federico Benvenuto[2], Marta Lazzaretti[2], Cristina Campi[2,4], Michele Piana[2,4], Simona Balestrini[1,3], Matteo Lenge[1], Renzo Guerrini[1,3]

[1]*Department of Neuroscience and Human Genetics, Meyer Children's Hospital IRCCS, Italy*

[2]*Department of Mathematics, University of Genoa, Italy*

[3]*Department of Neuroscience, Pharmacology and Child Health, University of Florence, Italy*

[4]*Life Science Computational Laboratory, IRCCS San Martino Hospital, Italy*

Background

The mammalian target of rapamycin (mTOR) pathway is a key regulator of cell growth, proliferation, survival, autophagy, transcription, and protein synthesis. Pathogenic variants in mTOR pathway genes have been associated with different overgrowth syndromes, including cancer, as they drive tumor cells to grow and divide uncontrollably. In the nervous system, hyperactivation of the mTOR pathway leads to a phenotypic continuum of malformations of cortical development (MCD), ranging from focal cortical dysplasia type II (FCDII) to megalencephaly (MEG) and dysplastic megalencephaly (DMEG).

Objectives

Given the complexity of the mTOR pathway and the number of genes it involves, understanding how a pathogenic variant in a specific gene alters downstream signaling is crucial for identifying novel candidate genes, disease biomarkers, and potential targets for personalized therapies.

Methods

We developed a novel digital-twin model designed to investigate how pathogenic variants in mTOR pathway genes affect the expression and phosphorylation of downstream effectors. This model comprises two main components. The first is an in silico predictor that uses chemical reaction networks (CRNs) to interpret and predict how mutations affecting specific genes alter the whole molecular dynamics profile of the proteins within the mTOR pathway. The second is a biological tool based on human neurons differentiated from induced pluripotent stem cells (iPSCs), reprogrammed from the fibroblasts of a patient carrying a pathogenic variant in the MTOR gene.

Results

Here, we present the first release of the in silico tool of the digital-twin model and the generation of cortical neurons derived from a unique patient carrying the MTOR p.Thr1977Ile variant in skin cells.

Conclusions

The digital-twin model we developed has the potential to identify and validate mTOR pathway changes induced by the MTOR p.Thr1977Ile variant. Once validated, it could serve as a standardized model to test the effects of other variants affecting mTOR pathway genes.

P-23

AI Prospects: Digital Twins, Citizen Care, One Health

Gold Standard Benchmarking for Mental Health Services Efficiency: Evidence from Andalusia (Spain)

Diego Diaz-Milanes[1,2], Carlos Benjamin Cordova Portillo[1], Carlos R. Garcia-Alonso[1,2]
[1]*Department of Quantitative Methods, Universidad Loyola Andalucia, Spain*
[2]*Health Research Institute, University of Canberra, Australia*

Background: Measuring efficiency in mental health systems is challenging due to uncertainty, heterogeneity, and the limitations of traditional single-indicator benchmarks. Standard Data Envelopment Analysis (DEA) may misclassify units and overlook variability. Artificial intelligence (AI) methods can enhance robustness by combining stochastic simulation and rule-based interpretation. Objectives: This study aimed to: (1) apply a Gold Standard (GS) benchmark to evaluate the Andalusian mental health ecosystem; (2) analyze the impact of integrating day care and hospital services relative to outpatient-only specifications; and (3) assess model performance. Methods: The Efficient Decision Support–Mental Health (EDeS-MH) framework integrates a Monte Carlo simulation engine, a fuzzy logic inference engine, and DEA under variable returns to scale. Service data were transformed into composite staffing indicators across outpatient, day care, and hospital care. DEA models were run with and without the GS benchmark in outpatient-only and multi-setting scenarios. Efficiency contrasts between standard DEA (RTE) and GS-DEA (GS-RTE) were classified into five effect types: amplification (both improve, GS-RTE shows larger gains), attenuation (both improve, GS-RTE shows smaller gains), GS-only gains, RTE-only gains, and no gains. Results: Amplification dominated, occurring in 96.5% of simulated units and 98.5% of areas. GS-only gains accounted for 3.2% of units and 1.5% of areas, while other effect types were rare (0.3%). Consistency analysis showed 76.9% of areas were always classified as amplification, and 96.9% reached this classification in at least 75% of replications. Mean efficiency differences confirmed the benchmark effect: average gains were 0.255 under RTE and more than doubled to 0.534 under GS-RTE. Benchmarking reduced misclassification and improved sensitivity and specificity compared with standard DEA. Conclusion: EDeS-MH with GS benchmarking provides a sensitive and accurate framework for efficiency evaluation. By integrating multiple care settings, it supports robust, AI-driven resource allocation and policy planning in Andalusia.

P-24

AI Prospects: Digital Twins, Citizen Care, One Health

Fuzzy-Logic Bayesian Network to Map 20-Minute Neighbourhood Pathways to Cardiometabolic Risk in Adelaide (Australia)

Diego Diaz-Milanes[1,2], Luis Salvador-Carulla[2], Mark Daniel[3], Carlos R. Garcia-Alonso[1,2]
[1]*Department of Quantitative Methods, Universidad Loyola Andalucia, Spain*
[2]*Health Research Institute, University of Canberra, Australia*
[3]*Department of Population Health, Dasman Diabetes Institute, Kuwait*

Background: Cardiometabolic risk (CMR) is shaped by the built environment through daily behaviours such as physical activity and diet. The 20-Minute Neighbourhood (20MN) concept offers a pragmatic, place-based lever to improve access to health-supportive amenities. We applied an AI approach to capture complex, uncertain interdependencies between buitl and natural environment and CMR. Objectives: To model causal pathways linking 20MN-relevant built-environment features with CMR in an Australian urban context using Bayesian Network Analysis (BNA), integrating expert knowledge with cohort and geospatial data. Methods: Wave 1 of the North-West Adelaide cohort enrolled 4,056 adults. The analytic sample for this study comprised 974 adults with complete data (56.3% women; 55.9 years on average, SD=10.6). Thirty-one destination types were grouped into eight 20MN domains (food facilities, gastronomic, public local facilities, education, local health, green/blue spaces, recreation, mobility/transport). Geographic indicators were calculated within a 1,600 m road-network buffer (approximately a 20-minute walk). An expert panel provided relevance and direction-of-effect ratings that were encoded in a fuzzy-logic Bayesian Network with CMR as the outcome node. We ran Monte Carlo scenario experiments (27,000–56,000 iterations) and assessed convergence. Results: When local access was limited or ignored, the green/blue domain concentrated in unhealthy ranges (Q1=32.2, Q3=40.6 on a 0–100 scale; SD=9.51), with ~99.7% of probability mass in unhealthy or very-unhealthy states. A habit-dominant variant (27,000 iterations) remained unhealthy (SD=4.15). Under high local availability, distributions shifted toward healthier profiles: one scenario showed a healthy mode with Q1=47.06 and Q3=67.06 (SD=11.47), including a very-healthy tail of ~14%. A second high-access test produced similar results (SD=11.04). Conclusions: An expert-informed, fuzzy-logic BNA indicates that improving 20MN access—particularly to green and blue spaces—shifts neighbourhood profiles from unhealthy toward healthy or very-healthy, consistent with reduced CMR. This interpretable AI framework supports simulation of urban-design policies for cardiometabolic prevention.

AI Prospects: Digital Twins, Citizen Care, One Health

Bridging the Knowledge Gap: An AI Chatbot for Migraine Education and Support

Andrea Franchini[1], Oscar Lithgow Serrano[1], Giulia Mallucci[2], Alberto Cordella[2], Chiara Zecca[2,3], Fabio Rinaldi[1]
[1]*Natural Language Processing and Information Retrieval, Dalle Molle Institute for Artificial Intelligence, Switzerland*
[2]*Neurocenter of Southern Switzerland, Ente Ospedaliero Cantonale, Switzerland*
[3]*Faculty of biomedical Sciences, Universit della Svizzera Italiana, Switzerland*

Background: Migraines is a debilitating neurological disease affecting over 1 billion people globally, where more than 90% of those affected report significant interferences in their daily lives. Despite its prevalence, migraine is often misdiagnosed. Given the limited time available between neurologist and patients, it is often difficult to provide patients with sufficient information on the desease.
Objectives: To bridge this knowledge gap, the NSI neurological group has partnered with IDSIA to explore the advantages of an AI-driven chatbot. The aim is to support neurologists in providing patients with reliable information about migraine. The platform will be available in the patient's own language (Italian), and a clinical trial is currently being conducted, which will allow to assess the efficacy of a patient–neurologist-chatbot solution compared to a typical patient–neurologist only session.
Methods: Our solution focuses on large language models, for their competency in natural language comprehension and generation, specifically GPT 4o and Deepl (translation). Our solution comprises a web-based interface, a translation layer (Italian-English-Italian), retrieval augmented generation to retrieve relevant pre-validated migraine documents, and a static prompt inspired from NSI's examples of patient interactions.
Results: The experiment is currently ongoing, yet our three partner neurologists, considered the model sufficiently accurate and effective in dispensing knowledge, generally on par with a human neurologist. Across the 14 patients currently active on the platform, the mean conversation lengths (number of utterances) is 10.5, mean length of uninterrupted conversations (time delta 5m) is 2.18 and mean number of new conversation started by a patient is 4.39.
Conclusion: The patients' involvement with the chatbot suggests a positive effect of our system as a supporting solution for migraine education. Patients are not just initiating contact but are actively interacting with the chatbot. Upon trial completion we will be able assess more advanced aspects such as user query relevancy and expertise evolution.

P-26

AI Prospects: Digital Twins, Citizen Care, One Health

AI-Assisted Evaluation of Work Capacity and Functional Limitations In Large-Scale Medical Records: Opportunities And Challenges

Ioannis Kokkinakis[1], Konstantinos Tzartzas[1], Murielle Bochud[2], Bernard Favrat[1]
[1]*Department of Vulnerabilities and Social Medicine, Unisante, University Center for Primary Care and Public Health & University of Lausanne, Switzerland*
[2]*Department of Epidemiology and Health Systems, Unisante, University Center for Primary Care and Public Health & University of Lausanne, Switzerland*

Background

Medical assessment of work capacity and functional limitations requires the analysis of large, heterogenous, and chronologically dispersed medical files. This raises concerns about time-efficiency, inter-expert variability, cost, reproducibility, and equity. Artificial intelligence (AI) adapted to annotate, analyze and synthetize complex medical documentation, offers promising tools to support the medical expertise process.

Objectives

We aim to explore the potential benefits and challenges of AI-assisted evaluation of large-scale medical records, focusing on work capacity and functional limitations in the context of medical expert assessments.

Methods

We conducted a narrative review of literature (2020–2025) using both traditional database searches (Pubmed, Endbase) and AI-assisted synthesis (Microsoft Copilot, GPT-4) to identify and summarize key publications on AI integration in complex clinical documentation, disability assessment, and expert evaluations. We focused on large language models (LLMs), natural language processing (NLP), and AI-supported tools for transcription, annotation, analysis, synthesis and clinical evaluation. All AI-generated content was critically reviewed by the authors for scientific accuracy and integrity.

Results

Out of 55 articles identified as potentially relevant based on title and abstract, 12 were selected for full text review. Studies suggest that AI can significantly reduce the time required to process extensive medical records (15-40%), improve consistency in expert assessments, summarize clinical narratives, extract relevant information and support decision-making. Challenges remain regarding data quality, model transparency, risk of bias and legal/ethical issues. Combining human expertise with AI-driven insights seems to provide better results in clinical contexts. Furthermore, studies show promising results concerning functional assessment, with a focus on the International Classification of Functioning, Disability and Health (ICF).

Conclusion

AI has the potential to revolutionize work capacity evaluation by supporting the expert in the management of voluminous medical records. While promising, its integration into medico-legal and insurance contexts requires robust scientific validation, ethical considerations, and interdisciplinary collaboration.

P-27

AI Prospects: Digital Twins, Citizen Care, One Health

Beyond English: Medical LLMs in German Preoperative Patient Education

Noah Leuenberger[1], Amr Saad[2], Daniel Perruchoud[1], Susanne Suter[1]
[1]*Institute of Data Science, University of Applied Sciences Northwestern Switzerland, Switzerland*
[2]*Executive Management, PatientEd Saad, Switzerland*

Background: Large Language Models (LLMs) offer significant potential for patient-facing applications, such as preoperative education. However, many state-of-the-art LLMs are predominantly trained on English data, posing a challenge for creating reliable, non-English conversational chatbots for specialized medical domains. This is particularly relevant in preoperative patient education, where conversational chatbots could handle routine patient inquiries, alleviate the conversational burden and streamline staff workflows.

Objectives: This study investigated the feasibility of using German-prompted LLMs for a preoperative patient-education chatbot. The primary goal was to design and implement a robust, automated evaluation pipeline to systematically benchmark LLMs and prompting strategies in a reproducible manner.

Methods: We created a modular pipeline and web interface to assess three LLMs via three prompting strategies: In-Context Learning (Few-Shot/Zero-Shot) and Retrieval-Augmented Generation (RAG). Additionally, we evaluated a system translating German input into English and back. The evaluation was based on 34 question-answer pairs, generating 306 model responses. A dual-track process combined an LLM-judge with a human evaluation, where a medical expert annotated a subset of responses across five metrics.

Results: The choice of prompting strategy was found to have a significant impact on performance. In-Context Learning approaches consistently outperformed RAG, indicating that the primary bottlenecks were related to retrieval and indexing rather than the inherent capabilities of the models. Furthermore, a systematic overestimation of quality by the LLM-judge compared to medical expert assessment was observed, particularly regarding safety and appropriateness. We observed trade-offs in correctness and latency between German-prompted and translation-based approaches.
Conclusion: This work successfully established a solid technical foundation and a reliable, scalable framework for the rapid-cycle evaluation of LLMs in a medical non-English context. The findings underscore the critical role of the evaluation pipeline in identifying key challenges, paving the way for the future development and validation of clinically applicable conversational systems.

AI Prospects: Digital Twins, Citizen Care, One Health

Integrating Expert Guidelines with LLMs for Personalised Meal-Level Recommendations

Ying-Chieh Liu[1]
Department of Industrial Design, Chang Gung University, Chinese Taipei

Background: Healthy eating is essential for well-being, promoting growth and lifelong health while reducing the risk of obesity and chronic diseases. In our previous research, we developed a rule-based system in collaboration with domain experts that encodes dietary guidelines to generate personalised dietary intake recommendations. The system is applied to use cases defined by specific physiological profiles based on metabolic syndrome indicators. Recently, technological advances and evolving user expectations motivate conversational and context-aware, generative-AI-driven personalisation.

Objectives: To identify a practical deployment path, we prototyped and compared two approaches—both employing prompting and retrieval-augmented generation (RAG)—that differ only in fine-tuning status: (1) a general-purpose GPT model (non-fine-tuned); and (2) a comparable base model that we fine-tuned.

Methods: The use case is based on a high-risk scenario characterised by elevated triglycerides (TG). Given meal inputs (e.g., lunch and dinner), the two approaches were implemented on the OpenAI platform under the same prompt and the same retrieval reference (one curated text file of guideline-aligned dish substitutions, e.g., for “deep-fried chicken”). To enable reproducible, scalable comparison under time and resource constraints, automated metrics included keyword/phrase coverage of predefined terms derived from expert dietary guidelines.

Results: The fine-tuned approach demonstrated some improvements on the predefined automated metrics; however, benefits may be limited or mixed in other aspects. Further research is needed, including domain-expert evaluation and expansion to additional use cases.

Conclusion: This study explores a practical approach that integrates our previous rule-based system with an LLM-enabled, context-aware prototype that analyses each user’s everyday meal intake. Overall, this approach shows potential to scale personalised dietary support. Future work will expand the fine-tuning dataset and evaluate the recommendations via domain-expert heuristic assessment. At a higher level, this work provides a practical path from rule-based tools to guideline-consistent LLM systems, enabling scalable personalisation without compromising clinical alignment.

P-29

AI Prospects: Digital Twins, Citizen Care, One Health

Accelerating National Health Strategy Reviews with AI: Operationalizing the WHO HSforHS Framework

Lina Yu[1], Peter Mala[1], Celine Hazbun[2], Nirmal Kandel[1]
[1]*Health Emergency Preparedness Department, World Health Organization, Switzerland*
[2]*Information and Digital Technology Department, World Health Organization, Switzerland*

Background

WHO developed the Health Systems for Health Security (HSforHS) framework to strengthen the foundational health system elements for health emergency preparedness. Embedding these elements in National Health Policies, Strategies, and Plans (NHSPs)—which guide national health priorities and tie to fiscal budgets—is essential. To this end, WHO institutionalized NHSPs review to identify the gaps in foundational health system elements and provide recommendations for their integration; but traditional reviews are time-consuming and resource intensive. To address this, we applied an AI-enabled tool to streamline the process.

Objectives

This work presents an AI-enabled tool designed to accelerate NHSPs reviews by operationalizing the WHO HSforHS framework. The tool aims to address inefficiencies in manual reviews, reduce technical workload and cost, and generate timely insights for countries to update their NHSPs.

Methods

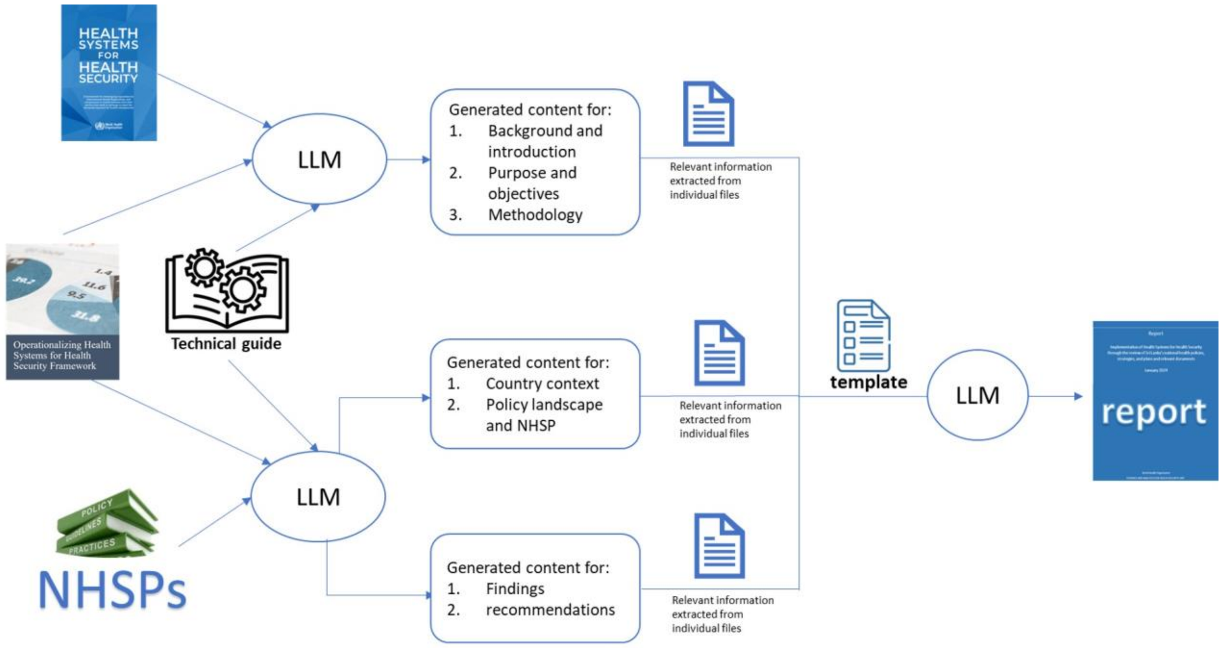
Use tailored large language models and natural language processing, the tool scans NHSPs and relevant documents for foundational elements of health systems, identifies gaps and produces a structured draft review in under ten minutes. Experts then refine the draft for technical accuracy and contextual relevance. Development of the prototype tool cost approximately USD 20,000, but it can be applied globally, offering a high return on investment.

Results

In test case with of Sri Lanka and Kenya's NHSPs review, the AI-generated drafts aligned 85-90% with manual reports, cutting review time from 3–4 weeks of intensive work by experts to minutes through an "upload-and-click" process. This not only saves time, and human resources, but also enables scalability—allowing reviews of all available NHSPs for a global overview, which is not feasible with traditional methods.

Conclusion

The AI-enabled tool transforms NHSP reviews by enhancing efficiency, quality, and scalability while saving cost, enabling countries and partners to achieve timely, evidence-based planning while optimizing resources.


HEALTH SYSTEMS FOR HEALTH SECURITY
Operationalizing Health Systems for Health Security Framework
Technical guide
NHSPs
LLM
LLM
Generated content for:
1. Background and introduction
2. Purpose and objectives
3. Methodology
Generated content for:
1. Country context
2. Policy landscape and NHSP
Generated content for:
1. Findings
2. recommendations
Relevant information extracted from individual files
Relevant information extracted from individual files
Relevant information extracted from individual files
template
LLM
report


Country
Sri Lanka
Drag & drop files or Browse
Supported formates: PDF
Generate AI Summary
AI generated-
Review_Report_Sri_Lanka_NHSP.doc
(120 kB)
Reports generated by this tool are subject to validation by WHO technical officers and should not be shared externally or with Member States.
Download
Important: Please download the report and store it on your computer. The report will be erased if you restart the process.