Report from Dagstuhl Seminar 25052

From Research to Certification with Data-Driven Medical Decision Support Systems

Raul Santos-Rodriguez*[1], Kacper Sokol*[2], Julia E. Vogt*[3], and Sven Wellmann*[4]

1 **University of Bristol, GB.** enrsr@bristol.ac.uk
2 **ETH Zürich, CH.** kacper.sokol@inf.ethz.ch
3 **ETH Zürich, CH.** julia.vogt@inf.ethz.ch
4 **Universität Regensburg, DE.** sven.wellmann@barmherzige-regensburg.de

Abstract

This report outlines the programme and outcomes of Dagstuhl Seminar 25052 "From Research to Certification with Data-Driven Medical Decision Support Systems". Our seminar addressed the complex challenges of transferring artificial intelligence systems from research labs into real-world clinical practice. Bringing together clinicians, researchers and industry stakeholders, it explored the potential and pitfalls of deploying data-driven models in healthcare, highlighting the need for rigorous evaluation, human-centred design and responsible innovation. Key discussions included regulatory hurdles, reproducibility issues, interpretability and human–machine collaboration. Group sessions focused on evaluation frameworks and human factors in medical artificial intelligence system design. The seminar laid the foundation for a collaborative research agenda aimed at safe, effective and ethical integration of data-driven predictive models into real-life clinical workflows.

Seminar January 26–31, 2025 – https://www.dagstuhl.de/25052
2012 ACM Subject Classification Human-centered computing; Computing methodologies → Artificial intelligence; Computing methodologies → Machine learning
Keywords and phrases artificial intelligence, clinical practice, decision support systems, digital healthcare, machine learning
Digital Object Identifier 10.4230/DagRep.15.1.201

1 Executive Summary

Kacper Sokol (ETH Zürich, CH)
Raul Santos-Rodriguez (University of Bristol, GB)
Julia E. Vogt (ETH Zürich, CH)
Sven Wellmann (Universität Regensburg, DE)

Seminar Vision

Artificial intelligence has made tremendous strides across many spheres of life, however deploying this technology in safety critical domains remains challenging. This Dagstuhl Seminar focuses on clinical practice where data-driven models can streamline the work of healthcare professionals and democratise access to personalised medicine, thus have lasting positive impact on society, but also where deploying such tools without adequate foresight and safeguards can be perilous. This duality – anticipated benefits that may come along

* Editor / Organizer

From Research to Certification with Data-Driven Medical Decision Support Systems, *Dagstuhl Reports*, Vol. 15, Issue 1, pp. 201–220
Editors: Raul Santos-Rodriguez, Kacper Sokol, Julia E. Vogt, and Sven Wellmann
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

with unintended consequences – requires new technologies to be thoroughly vetted, e.g., with clinical trials and medical certification processes, before they can be deployed to avoid any harmful fallout. However, fulfilling such regulatory requirements is a lengthy and complex process plagued with many challenges, hence while prototype systems are becoming increasingly ubiquitous, they often remain indefinitely designated as research tools that can be used exclusively for research purposes. Their lacklustre adoption is compounded by pervasive reproducibility issues; history of unsafe systems being deployed prematurely; scarce data that are inherently private, difficult to collect or share, and often riddled with numerous biases; and prevalence of automation promises that never come to fruition. Such hurdles result in healthcare remaining one of the least digitised spheres of life.

A different contributing factor is predictive systems often being misconstrued as autonomous rather than social and relational, which is manifested in a counterproductive drive to match or exceed human-level performance in selected (narrowly- or ill-defined) tasks, with the aim to fully automate and replace humans. This goal has nonetheless repeatedly proven difficult to attain due to brittle predictions whose subpar fairness, interpretability and robustness as well as ambiguous accountability are concerning, especially given their potential harm. By considering the broader organisational and societal context in which data-driven systems are operationalised, we should not only strive to automate and replace (when appropriate and desirable) but also to augment and support human reasoning and decision-making to help people flourish at work, e.g., through human–machine collaboration that preserves people's agency and maintains the attribution of responsibility with them. Such a perspective promises to offer an antidote to widely reported apprehension of artificial intelligence and expedite its adoption in safety critical domains.

Seminar Topic

To address these challenges, our interdisciplinary seminar gathers a broad range of stakeholders – including clinicians, academics and researchers from industry – whose diverse expertise can contribute to charting a novel research agenda for effective and responsible adoption of artificial intelligence in medicine given the complex sociotechnical landscape outlined above. Our goal is to identify best ways of operationalising medical data-driven systems as to ensure their alignment with the needs and expectations of various stakeholders in healthcare as well as seamless integration into real-life clinical workflows, taking a human-centred perspective. Exploring these aspects of artificial intelligence is especially important given that achieving state-of-the-art performance on benchmark tasks often does not directly translate into clinical efficacy and acceptability. To support this objective, we additionally intend to scrutinise relevant evaluation procedures, medical device certification processes, practicality of clinical trials involving data-driven algorithms and clinical approvals thereof in view of compliance with various laws, rules and regulations as well as societal norms and ethical standards. Throughout the seminar we envisage identifying challenges that can be addressed with current technologies, distilling areas that require further work, and emphasising promising research directions. Finally, the event aims to galvanise an interdisciplinary community dedicated to advancing the meeting's agenda after its conclusion.

Seminar Outcomes

The seminar focused on the challenges of translating medical artificial intelligence (AI) models from research settings to real-world clinical applications. It brought together academic and

industry researchers, start-up representatives as well as practising clinicians to foster a multidisciplinary exchange of ideas. One of the key highlights of the seminar was an invited keynote by Rich Caruana from Microsoft Research. His presentation on ante-hoc interpretable models emphasised the importance of intelligibility in machine learning for healthcare. This talk sparked significant discussions among the participants and served as a catalyst for many of the conversations that followed.

Throughout the seminar, the participants engaged in a variety of discussions and presentations. Clinicians were invited to share their experiences with data-driven decision support systems, focusing on both success stories and ongoing challenges; they were also encouraged to describe their hopes and vision for the future of such tools. These clinical pitches played a central role in shaping the seminar's core themes, which included research, translation, testing, deployment, monitoring, updating and maintenance of AI systems in healthcare. Additionally, researchers delivered short presentations on their work, providing insights into the state of the art as well as open research problems in clinical AI systems. A dedicated session for start-ups offered valuable insights into the process of transforming research findings into real-life clinical tools. Among others, entrepreneurs shared their experiences with commercialisation and the regulatory hurdles they encountered. Many discussions revolved around the practical aspects of deploying AI in healthcare settings and the lessons learnt from these experiences.

The seminar also facilitated group work; two dedicated working groups were formed. The first group focused on frameworks for evaluation and (post-deployment) monitoring of clinical AI. The second group explored important criteria to consider when selecting clinical problems for which to develop AI tools; it additionally investigated human factors of medical AI systems and key approaches to improve the interaction between AI and doctors.

Overall, the seminar identified pressing challenges and opportunities in clinical AI research and deployment. Clinicians gained a deeper understanding of AI's capabilities and limitations, while researchers benefited from the exchange of strategies for overcoming integration and adoption barriers. The discussions and findings from the seminar are expected to facilitate smoother transitions from research to clinical AI prototypes, allowing such tools to be tested and deployed in hospitals. By fostering interdisciplinary collaboration, the seminar laid the groundwork for future innovations in AI-driven clinical decision support systems. The insights shared and connections formed during the event will contribute to ongoing advancements in the field and help bridge the gap between AI research and practical healthcare applications.

2 Table of Contents

Executive Summary

Kacper Sokol, Raul Santos-Rodriguez, Julia E. Vogt, and Sven Wellmann 201

Overview of Talks

Bridging the Gap Between Clinical Data and AI: Lessons From Real-World EMR Studies
Brett Beaulieu-Jones 206

Implementing Clinical Workflows in the Clinic
Michael Brudno 206

Friends Don't Let Friends Deploy Black Box Models: The Importance of Intelligibility in Machine Learning for Healthcare
Rich Caruana 207

Validation in Biomedical Imaging AI: Are We Ready for Clinical Translation?
Evangelia Christodoulou 208

Towards Deployment: Considerations Beyond Technical Performance
Jeff Clark 209

AI at the Bedside. There Must Be a Culture Change
James Fackler 209

A Few Lessons Learnt From Trying To Work With Healthcare and Related Data
Thomas Gärtner 209

AI in Healthcare: Key Human–Computer Interaction Challenges
Maia Jacobs 210

From Code to Clinic – From Bits to Bedside
Michael Kamp 210

Learning From Machine Learning – How To Deduce a Mechanism-Based Pharmacometrics Model for Serum Creatinine in Preterm Neonates From Neural Ordinary Differential Equations
Gilbert Koch 211

AI for Chronic Care Management
Yamuna Krishnamurthy 213

Rethinking Medical AI: Evaluation, Representation and Transferability
Christoph Lippert 213

All Models Are Wrong and Yours Are Useless
Florian Markowetz 214

Predictive Analytics Monitoring at the Bedside
Randall Moorman 214

Interpretability?
Rajesh Ranganath 215

AI in Paediatric Surgery and Paediatric Urology
Patricia Reis Wolfertstetter 215

My Priorities for AI in Health: Proper Evaluation and Prediction Under Intervention
Wouter van Amsterdam 216

Scaling up Clinical ML: Modalities, External Validation, Health Systems
Robin Van de Water 216

AI in Babies and Beyond, Boom or Boomerang?
Sven Wellmann 217

Working Groups

AI Monitoring in Clinical Practice
Brett Beaulieu-Jones, Evangelia Christodoulou, Thomas Gärtner, Michael Kamp, Gilbert Koch, Yamuna Krishnamurthy, Fabian Laumer, Christoph Lippert, Florian Markowetz, Randall Moorman, Rajesh Ranganath, Raul Santos-Rodriguez, Wouter van Amsterdam, Robin Van de Water, and Julia E. Vogt 218

Human Factors in Clinical AI Design and Deployment
Michael Brudno, Jeff Clark, James Fackler, Maia Jacobs, Patricia Reis Wolfertstetter, Kacper Sokol, and Sven Wellmann 219

Participants 220

3 Overview of Talks

3.1 Bridging the Gap Between Clinical Data and AI: Lessons From Real-World EMR Studies

Brett Beaulieu-Jones (University of Chicago, US)

Healthcare data, particularly electronic medical records (EMRs), present significant challenges due to their complexity, inconsistency and inherent biases. This presentation explores the implications of these issues for clinical artificial intelligence (AI) and phenotyping models, emphasising the role of clinician-initiated (CI) versus non-clinician-initiated (NCI) data in predictive modelling. Using real-world case studies, we examine how AI models interpret EMR data, the risks of confounding feedback loops in clinical decision support and the divergence between models trained on CI and NCI data. We highlight findings from large-scale EMR studies on patient risk stratification, model performance limitations and the impact of institutional effects. Additionally, we discuss the dangers of label leakage, reproducibility challenges in published predictive models and the unintended consequences of AI-based clinical alerts. The talk underscores the need for rigorous evaluation of AI models deployed in clinical settings to ensure they enhance, rather than hinder, medical decision-making.

3.2 Implementing Clinical Workflows in the Clinic

Michael Brudno (University of Toronto, CA)

In this presentation I will look at the challenges of implementing machine learning (ML) in a hospital setting, concentrating specifically on integrating ML into clinical workflows in a safe and effective manner. I will utilise two examples from my research: the deployment of Machine Learning Medical Directives (MLMD) for making low-risk decision in paediatric Emergency Rooms and scheduling of craniosynostosis and plagiocephaly patients for Plastic Surgery consultations based on their likely risk and urgency.

To develop MLMD we used data from the EHR system from the Hospital for Sick Children, a tertiary care hospital in the city of Toronto, Canada to train multiple ML models to predict the need for urinary dipstick testing, ECGs, abdominal ultrasounds, testicular ultrasounds, bilirubin testing and forearm X-rays using data available at triage. There was a total of 42,238 patients (54.7% boys) included in model development; mean (SD) age of the children was 5.4 (4.8) years. Models obtained high area under the receiver operator curve (0.89–0.99) and positive predictive values (0.77–0.94) across each of the use cases. The proposed implementation of MLMDs would streamline care for 22.3% of all patient visits and make test results available earlier by 165 minutes (weighted mean) per affected patient. Model explainability for each MLMD demonstrated clinically relevant features having the most influence on model predictions. In the presentation we emphasised the safety of deploying these ML models and the importance of considering clinical workflows (staff availability, importance of explaining the AI models to patients, etc.) in deployment.

In the second example we consider the scheduling of appointments of craniosynostosis and plagiocephaly patients in a plastic surgery department. Craniosynostosis is a birth defect that results in a misshapen skull due to premature bone fusion as a newborn's skull is formed. In some cases, skulls with this defect do not have adequate space for the newborn's brain to grow, which increases the chance of visual and mental development impairments; almost all cases of craniosynostosis also result in head shape abnormalities that may lead to bullying and impact individual self-perception. Craniosynostosis can be corrected by relatively non-invasive surgery before 3 months; after this age, however, patients require more complex surgery with higher morbidity. Craniosynostosis is typically diagnosed by a physical examination by a specialist, such as a paediatric plastic surgeon. Paediatricians who are not trained at identifying craniosynostosis often confuse it for plagiocephaly, a related but mostly benign condition, and typically refer patients with either condition to plastic surgery for a definitive diagnosis. However, the delay associated with the referral process can require the more complex surgical approach. We have recently demonstrated that 3D head shape reconstruction using a standalone ToF camera (3DMD system) can aid in the identification of craniosynostosis with high accuracy and allow prioritisation of referred patients. Again, deploying this tool into clinical care requires careful consideration of existing clinical workflows. While reducing the overall burden for some families, it would require patients to make multiple visits (one to have a 3D photo taken, one to see the surgeon for a comprehensive evaluation). This would potentially lead to some inequity, as patients further from the hospital would require more resources to benefit from the AI.

In discussing both cases I will emphasise the important "fall-back" mechanisms, where if a patient is not flagged by the ML system, they will still undergo standard-of-care treatment, and also consider how the presence of automation may impact clinicians who become "accustomed" to having the support, and may fail to act appropriately if the technology malfunctions.

3.3 Friends Don't Let Friends Deploy Black Box Models: The Importance of Intelligibility in Machine Learning for Healthcare

Rich Caruana (Microsoft – Redmond, US)

The conventional wisdom in machine learning has been that to achieve high accuracy you must use opaque black-box models such as deep neural nets, boosted trees or random forests, and that if you want models to be interpretable and able to explain their predictions, you have to accept a loss in accuracy. This trade-off is no longer true when working with tabular data – in the last 10 years glass-box learning methods have been developed that are just as accurate as black-box learning methods but which are fully interpretable and can explain their predictions. Applying these glass-box learning methods to healthcare data has uncovered many problems inherent in clinical data that would make models trained on the data risky to use on patients. These problems include selection bias, race and gender bias, treatment effects, other forms of statistical confounding and problems with popular methods of dealing with missing data and data coding.

None of these are new problems. What is new is how widespread these problems are, how unexpected some of them are even in high-quality well-curated data and the difficulty of correcting these problems using traditional methods. The new high-accuracy glass-box

learning methods have shown that these problems exist in every dataset. Moreover, these problems make all black-box models trained on medical data suspect because one is unable to anticipate all of the problems in advance and it is difficult to fully understand after the fact what has been learnt by complex black-box models. Glass-box learning methods not only make it easier to detect these problems, but also provide tools for correcting many of these problems by allowing clinicians to use their expertise to directly correct/edit the models when they have learnt patterns that would put patients at risk.

In the talk we examined a half dozen case studies using real medical data that show the kinds of problems that are common in medical datasets, and how we would use glass-box learning to detect and then correct these problems. The case studies serve as a wake-up call to anyone using machine learning and artificial intelligence in healthcare that if they are training and/or using models that they cannot fully understand (i.e., black-box models), then they are almost certainly putting patients at higher risk if model predictions are acted upon. In addition to providing models that are fully interpretable, some of the new glass-box learning methods not only provide methods to correct models, but also can explain their reasoning and help protect privacy. Now that glass-box learning methods are so powerful, it would be wrong to intentionally use black-box models in critical domains such as healthcare if glass-box models yielded comparable accuracy.

The talk was not about the technical details of any one glass-box learning method. Instead, it was a collection of case studies that show the dangers of using black-box models in healthcare and how glass-box methods can be used to mitigate these risks. Once the problems hidden in each dataset are uncovered, there may be multiple methods available to tackle the problems, but the key challenge is to detect the problems in the first place so that they can be corrected prior to deploying the model.

3.4 Validation in Biomedical Imaging AI: Are We Ready for Clinical Translation?

Evangelia Christodoulou (DKFZ – Heidelberg, DE)

Reliable validation of machine learning (ML) algorithms remains a critical challenge, particularly in biomedical image analysis, where chosen performance metrics often fail to reflect domain interests. To address this, we introduce Metrics Reloaded, a comprehensive framework guiding researchers in selecting problem-aware validation metrics. Developed by an international consortium, it employs a structured problem fingerprint to capture key aspects influencing metric selection. Additionally, we highlight a crucial limitation in current performance reporting: the widespread neglect of performance variability. Analysing 221 MICCAI 2023 segmentation papers, we find that over 50% do not assess variability, and only 0.5% report confidence intervals (CIs). To bridge this gap, we propose an approximation method that reconstructs CIs using unreported standard deviation values, revealing that reported performance differences often lack statistical significance. Together, these contributions aim to enhance ML validation practices, ensuring more reliable and clinically relevant algorithm evaluation.

3.5 Towards Deployment: Considerations Beyond Technical Performance

Jeff Clark (IngeniumAI – Bath, GB)

When developing a decision support system, it is tempting to focus most of your energy on the core technology: the technical innovation, which we believe will have a positive impact on the healthcare system once deployed. In this talk I touch upon many of the other required facets, which must be pursued in parallel, as you move from a research project towards deployment. This includes technical considerations concerned with safe deployment such as prospective performance and drift, but also many factors beyond the performance of the core technology, including but not limited to: initiating a quality management system, regulatory evidence and documentation, route to market strategy, human factors and healthcare economics. None of these other factors can be ignored, and will be pivotal to the success of deploying your innovation.

3.6 AI at the Bedside. There Must Be a Culture Change

James Fackler (Johns Hopkins University – Baltimore, US)

Focused on paediatric critical care, I believe there are no current uses of machine learning (ML) or artificial intelligence (AI) in general that have reached the bedside. However, I remain optimistic that AI will have a profound impact on patient care in the next five years (or ten at the longest).

To leverage AI at the bedside will require a substantial medical culture change. Because knowledge will be "ubiquitous", the traditional hierarchy where the doctor (or in academic medicine, the attending physician) is the final arbiter of truth and the sole source of a care plan, will be upended. Patients will have access to the same knowledge as do the doctors. The role of the senior clinician will become one who understands what AI "knows" and more importantly what AI does not (or cannot) know. Individuals on the care team (e.g., nurses, junior physicians, pharmacists) will develop the same relationship with AI and knowledge but will do so within the "niche" expertise.

3.7 A Few Lessons Learnt From Trying To Work With Healthcare and Related Data

Thomas Gärtner (Technische Universität Wien, AT)

In my talk, I reported on a variety of experiences from (so far) mostly unsuccessful attempts of applying machine learning algorithms to healthcare and related data. My first experience was with time series of oxygen levels taken during brain surgeries and was available for

a very small number of patients only. My next experience was on images of eyes with implanted lenses for cataract patients and could be solved sufficiently well without the use of sophisticated machine learning algorithms. My most recent experience is with clinical studies and involves long discussions about NDAs and IPRs with legal departments.

3.8 AI in Healthcare: Key Human–Computer Interaction Challenges

Maia Jacobs (Northwestern University – Evanston, US)

The use of artificial intelligence (AI) for improving medical decision-making has garnered great excitement in recent years. Yet, despite growing enthusiasm and increased research, real-world clinical impact has been slow. Often, abandonment of these tools in clinical settings is not related to algorithmic performance, but rather due to inattention towards the technologies' design and implementation. To understand and address these challenges, I will share two of my lab's research projects, which use user-centred and participatory design methods to incorporate both providers' and patients' perspectives into clinical decision support systems.

3.9 From Code to Clinic – From Bits to Bedside

Michael Kamp (Universitätsmedizin Essen, DE)

The integration of artificial intelligence (AI) into clinical practice requires not only technological advancements but also rigorous methods to ensure reliability, privacy and interpretability. At the University Hospital Essen (UK Essen), one of the world's leading smart hospitals, the Institute for AI in Medicine (IKIM) develops and deploys AI systems within a large-scale data infrastructure based on Europe's largest FHIR server. This enables advanced machine learning applications while maintaining strict data governance.

This talk will present research from the Trustworthy Machine Learning group, focusing on three core challenges in medical AI: privacy-preserving federated learning, where we move beyond standard model aggregation techniques to improve learning from distributed clinical data; statistical performance guarantees, leveraging theoretical insights from loss surface analysis to better understand generalisation in deep learning; and (federated) causal discovery, which aims to disentangle causal relationships in medical datasets to improve model interpretability and robustness.

By combining these approaches, we work toward AI models that are not only predictive but also scientifically grounded and reliable in clinical decision-making. The talk will discuss recent advancements in federated learning, causal inference and generalisation theory, along with their implications for AI applications in healthcare.

3.10 Learning From Machine Learning – How To Deduce a Mechanism-Based Pharmacometrics Model for Serum Creatinine in Preterm Neonates From Neural Ordinary Differential Equations

Gilbert Koch (Universitäts-Kinderspital beider Basel, CH)

Introduction

Machine learning (ML) is an emerging field in pharmacometrics (PMX) [1], providing methods for a variety of PMX tasks, including data preparation [2], data analysis and data modelling. One ML approach gaining special attention in PMX are neural ordinary differential equations (NODEs) [3, 4, 5, 6]. Although an NODE is basically an ordinary differential equation (ODE), the difference is that the right-hand side of the ODE is not described with mechanism-based functions, as it is typically done in PMX, but it consists of neural networks (NNs). Consequently, these NNs learn the dynamics observed in the training data. However, there are some major criticisms regarding NODEs, including that (i) they are "black box" models, (ii) they have poor extrapolation capabilities, e.g., for unseen dose ranges, due to their structure, and (iii) they do not include prior clinical knowledge. In this work, a reverse modelling approach is presented that leverages the learnt knowledge by a NODE to deduce a mechanism-based model allowing to additionally include clinical knowledge. This enables to overcome the criticism of NODEs mentioned above and to make them a more viable approach in the field of PMX.

Methods

As endurance test, a dataset consisting of serum creatinine concentration measurements ($n = 4{,}026$) from extremely low birth weight neonates ($n = 217$) with marked renal maturation processes was applied [7]. The low-dimensional NODE approach was utilised [6] where the right-hand side of the NODE consists of two types of NNs specifically tailored to PMX. The first NN takes the state as input, reflecting the autonomous behaviour of the dynamics. The second NN takes explicit time as input, reflecting behaviour of the dynamics that change over time, e.g., maturation processes. First, the serum creatinine measurements were fitted with the low-dimensional NODE in the non-linear mixed-effects context in Monolix and a covariate analysis was performed. Second, the learnt dynamics of the NNs were visualised in derivative versus state or time plots [6]. Based on visual inspection of these plots, PMX functions were selected that described the shape of the trajectories in these plots. Third, these PMX functions were combined to deduce a mechanism-based model that is capable to characterise the dynamics of serum creatinine concentrations. Fourth, this deduced mechanism-based model was further refined with clinical knowledge about the influence of body weight on the volume of distribution. As last step, this deduced final mechanism-based model was fitted to the data, a covariate analysis was conducted with the previously gained information from the NODE-covariate analysis and simulations were performed.

Results

The developed low-dimensional NODE was capable of learning complex dynamics of serum creatinine in preterm neonates with good measures of precision and bias (mean squared error MSE = 0.023 and relative mean prediction error RMPE = 1.471). In comparison to the

previously published model [7], the NODE model provided similar data fitting and simulated similar GA-dependent reference values. Further it was able to identify the most important covariates found in the previously published model. Based on the visualised trajectories in the derivative versus state or time plots, a linear function for the NN characterising the state and an Emax function for the NN describing time were chosen. Remarkably, the deduced mechanism-based model had a similar structure as the previously published serum creatinine model [7]. In addition, clinical knowledge was included, i.e., volume of distribution for serum creatinine was assumed to be 7 dL/kg, resulting in the final mechanism-based model with similar measures of precision and bias as the NODE model (MSE = 0.025, RMPE = -2.17). It should be noted that NODE-based ML approach dramatically reduced time effort associated with the development of a mechanism-based model describing serum creatinine dynamics in neonates.

Conclusion

A mechanism-based model was successfully deduced from the dynamics learnt by the NODE. Structure of the deduced mechanism-based model was in accordance with a previously published, conventionally developed model for serum creatinine concentration in preterm neonates. Hence, we demonstrated the potential that initially learning the dynamics by an NODE is expected to accelerate development of mechanism-based models, particularly in paediatrics.

References

1 Alexander Janssen, Frank C. Bennis, and Ron A. A. Mathôt. *Adoption of machine learning in pharmacometrics: An overview of recent implementations and their considerations*. Pharmaceutics, 14(9), p.1814, MDPI, 2022

2 Dominic Stefan Bräm, Uri Nahum, Andrew Atkinson, Gilbert Koch, and Marc Pfister. *Evaluation of machine learning methods for covariate data imputation in pharmacometrics*. CPT: Pharmacometrics & Systems Pharmacology, 11(12), p.1638–1648, Wiley Online Library, 2022

3 Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. *Neural ordinary differential equations*. Advances in Neural Information Processing Systems, 31, 2018

4 James Lu, Kaiwen Deng, Xinyuan Zhang, Gengbo Liu, and Yuanfang Guan. *Neural-ODE for pharmacokinetics modeling and its advantage to alternative machine learning models in predicting new dosing regimens*. iScience, 24(7), Elsevier, 2021

5 Christopher Rackauckas, Yingbo Ma, Julius Martensen, Collin Warner, Kirill Zubov, Rohit Supekar, Dominic Skinner, Ali Ramadhan, and Alan Edelman. *Universal differential equations for scientific machine learning*. arXiv preprint arXiv:2001.04385, 2020

6 Dominic Stefan Bräm, Uri Nahum, Johannes Schropp, Marc Pfister, and Gilbert Koch. *Low-dimensional neural ODEs and their application in pharmacokinetics*. Journal of Pharmacokinetics and Pharmacodynamics, 51(2), p.123–140, Springer, 2024

7 Tamara van Donge, Karel Allegaert, Verena Gotta, Anne Smits, Elena Levtchenko, Djalila Mekahli, John van den Anker, and Marc Pfister. *Characterizing dynamics of serum creatinine and creatinine clearance in extremely low birth weight neonates during the first 6 weeks of life*. Pediatric Nephrology, 36, p.649–659, Springer, 2021

3.11 AI for Chronic Care Management

Yamuna Krishnamurthy (Phamily – New York, US)

In this talk I presented how we, at Phamily, are empowering chronic care management with artificial intelligence (AI). Phamily is a healthcare start-up with a vision to provide value-based care to chronic care patients. Our goal is to provide a system that brings physicians, nurses, care managers and patients together for continued conversations about the care that the patients need and how they can be addressed by the medical staff in between office visits. AI is the much needed assistant to the care managers that can help them quickly do chart reviews, draw up care plans and engage the patients. It can also assess short- and long-term patient risks for early detection and timely intervention that can save lives and prevent exorbitant costs for all involved.

3.12 Rethinking Medical AI: Evaluation, Representation and Transferability

Christoph Lippert (Hasso-Plattner-Institut, Universität Potsdam, DE)

Medical artificial intelligence (AI) systems promise to revolutionise clinical practice, but questions about their real-world effectiveness and interoperability remain largely unanswered. In this talk, I addressed two fundamental questions: *Do we need to evaluate medical AI?* and *Do we need ontologies?*

In the first part, I discussed insights from our study on commercial AI systems for tuberculosis detection [1], where we found that key information – such as training population details – is often lacking or opaque. This undermined model applicability, especially in global health settings, and required us to conduct extensive pilot testing to adapt a commercial algorithm for use in a South African community-based screening initiative.

In the second part, I turned to the role of ontologies in medical AI. Based on our recent work [2], I argued that representations learnt by large language models (LLMs) offer a superior and more scalable alternative to traditional medical ontologies. Our GRASP model embeds medical codes into a unified semantic space using LLMs, enabling cross-system and cross-country transferability of EHR-based prediction models – even without harmonised data models.

Taken together, the talk advocates for greater transparency in evaluation and a shift from static ontologies to dynamic, data-driven language representations for advancing trustworthy and generalisable medical AI.

References

1 Jana Fehr, Stefan Konigorski, Stephen Olivier, Resign Gunda, Ashmika Surujdeen, Dickman Gareta, Theresa Smit, Kathy Baisley, Sashen Moodley, Yumna Moosa, Willem Hanekom, Olivier Koole, Thumbi Ndung'u, Deenan Pillay, Alison D. Grant, Mark J. Siedner, Christoph Lippert, Emily B. Wong, and Vukuzazi Team. *Computer-aided interpretation of chest radiography reveals the spectrum of tuberculosis in rural South Africa*. npj Digital Medicine, 4(1), p.106, Nature Publishing Group UK London, 2021

2 Matthias Kirchler, Matteo Ferro, Veronica Lorenzini, FinnGen, Christoph Lippert, and Andrea Ganna. *Large language models improve transferability of electronic health record-based predictions across countries and coding systems*. medRxiv, 2025–02, Cold Spring Harbor Laboratory Press, 2025

3.13 All Models Are Wrong and Yours Are Useless

Florian Markowetz (University of Cambridge, GB)

Most published clinical prediction models are never used in clinical practice and there is a huge gap between academic research and clinical implementation. In this talk I propose ways for academic researchers to be proactive partners in improving clinical practice and to design models in ways that ultimately benefit patients.

3.14 Predictive Analytics Monitoring at the Bedside

Randall Moorman (University of Virginia – Charlottesville, US)

Predictive analytics monitoring is a new field of research and development that is ready for clinical implementation. The precepts are that there are detectable signatures of illness in continuous monitoring data and that detection of these signatures can lead to early detection, early treatment and improved outcomes.

I describe a successful example. Sepsis in premature infants is a common and pernicious problem but, if the culprit infection is diagnosed early, antibiotic treatment averts severe morbidity and mortality. We found more than 20 years ago that there was a robust signature of illness, reduced variability and transient decelerations of heart rate, that appeared hours before clinical presentation. A very large randomised trial showed that infants with a display of a risk index based on these abnormal heart rate characteristics had improved all-cause survival. While the signature was detected by visual inspection of many heart rate records by clinicians, the same signature was detected by machine learning and deep learning methods.

There are challenges to this new field. Before modelling begins, data sets may not have well-annotated target events and the data may reflect the clinicians and not the patients. When modelling, deep learning may be no better than machine learning, bias in the data may lead to bias in the model and the model output may not be explainable to clinicians. After deployment, effective implementation is difficult and the model will not work the same if the data shift.

3.15 Interpretability?

Rajesh Ranganath (NYU Courant Institute of Mathematical Science, US)

Joint work of Aahlad Puli, Nhi Nguyen, Rajesh Ranganath

Feature attributions attempt to highlight what inputs drive predictive power. Good attributions or explanations are thus those that produce inputs that retain this predictive power; accordingly, evaluations of explanations score their quality of prediction. However, evaluations produce scores better than what appears possible from the values in the explanation for a class of explanations called encoding explanations. Probing for encoding remains a challenge because there is no general characterisation of what gives the extra predictive power. We develop a definition of encoding that identifies this extra predictive power via conditional dependence and show that the definition fits existing examples of encoding. This definition implies, in contrast to encoding explanations, that non-encoding explanations contain all the informative inputs used to produce the explanation, giving them a "what you see is what you get" property, which makes them transparent and simple to use. Next, we prove that existing scores (ROAR, FRESH, EVAL-X) do not rank non-encoding explanations above encoding ones, and develop STRIPE-X, which ranks them correctly. After empirically demonstrating the theoretical insights, we use STRIPE-X to show that despite prompting a large language model (LLM) to produce non-encoding explanations for a sentiment analysis task, the LLM-generated explanations encode.

3.16 AI in Paediatric Surgery and Paediatric Urology

Patricia Reis Wolfertstetter (KH Barmh. Brüder Klinik St. Hedwig – Regensburg, DE)

Artificial intelligence and machine learning models are promising tools for the further development of paediatric surgery and paediatric urology. They can be used for optimising treatment and patient stratification preoperatively, during operation and postoperatively. Up to now, our work focused on paediatric appendicitis. First, predictive models were developed and validated on a dataset acquired from 430 children and adolescents aged 0–18 years, based on a range of information encompassing history, clinical examination, laboratory parameters and abdominal ultrasonography. Logistic regression, random forests and gradient boosting machines were used for predicting the three target variables: diagnosis, treatment and severity. Furthermore, we presented interpretable machine learning models for predicting the diagnosis, management and severity of suspected appendicitis using ultrasound images. Our approach utilised concept bottleneck models (CBM) that facilitate interpretation and interaction with high-level concepts understandable to clinicians. We extended CBMs to prediction problems with multiple views and incomplete concept sets. Our models were trained on a dataset comprising 579 paediatric patients with 1,709 ultrasound images accompanied by clinical and laboratory data. The developed models are deployed as an open access easy-to-use online tool (for tabular data and ultrasound images).

3.17 My Priorities for AI in Health: Proper Evaluation and Prediction Under Intervention

Wouter van Amsterdam (University Medical Center Utrecht, NL)

Artificial intelligence systems in healthcare must ultimately support safe and effective decision-making. In this talk, I argue that evaluation should extend beyond predictive performance on held-out data to the real-world setting – treating deployment itself as an intervention. I highlight how misaligned evaluation metrics can lead to harmful self-fulfilling prophecies, especially in treatment decision support.

Next, I argue researchers should build models for "prediction under intervention" (sometimes referred to as counterfactual prediction): estimating what would happen under different treatment options rather than expected outcomes under historical regimes. I draw on methods from causal inference and off-policy evaluation, and reflect on how emerging regulatory frameworks (EU and FDA) and available randomised control trial data can support a more rigorous approach.

3.18 Scaling up Clinical ML: Modalities, External Validation, Health Systems

Robin Van de Water (Hasso-Plattner-Institut, Universität Potsdam, DE)

Scaling medical machine learning (ML) requires integrating multiple modalities, improving model validation practices and establishing robust infrastructure to handle massive amounts of clinical data.

Scaling Up to Different Modalities

In visceral surgery, postoperative complications often arise in nursing wards, where real-time monitoring is limited. While ML-powered predictive systems show promise in the intensive care unit (ICU), their effectiveness diminishes outside of it due to data shortages, leaving patients at risk. To address this, we propose an integrated approach that combines patient data from preoperative, intraoperative, ICU and nursing ward stages, while introducing high-resolution continuous vital sign monitoring in a hybrid nursing environment [2]. This system enhances early detection of complications like surgical site infections and bile leakage, demonstrating the importance of high-quality wearable data. Our findings suggest that hybrid monitoring can significantly improve ML-based early warning models in clinical settings, enhancing patient outcomes.

Scaling Up External Validation

One of the biggest challenges in scaling clinical ML is ensuring reproducibility and transparency across datasets. With ICU models, it is difficult to verify claims of superior performance due to lack of access to datasets as well as unclear cohort definitions and preprocessing steps. To address this, we introduce YAIB, a modular framework designed to support reproducible

clinical ML experiments with multiple open-access ICU datasets [3]. YAIB provides an end-to-end solution for model evaluation, including predefined tasks like mortality and sepsis, and highlights the critical role of dataset selection, cohort definition and preprocessing in model performance. By offering a unified benchmarking tool, YAIB paves the way for more transparent, comparable ML research in clinical settings.

Scaling Up to Entire EHR Systems

For ML to be deployed at scale across entire health systems, we need an efficient infrastructure to process both retrospective and prospective clinical data. To meet this need, we developed the Medical Event Data Standard (MEDS) [1], a flexible low-level ML standard that integrates seamlessly with existing electronic health record processing and modelling tools. MEDS accelerates the training of predictive models on several clinical tasks. As a proof of concept, we built an ETL pipeline to enable model development using the recently released NWICU dataset. Additionally, we are working to convert data from the Mount Sinai AIR MS PHI OMOP database into MEDS, making it possible to train models for various clinical endpoints and develop foundation models. With ongoing efforts to verify data quality and define cohorts, we aim to enhance model robustness and further advance the scalability of ML in healthcare.

References

1 Bert Arnrich, Edward Choi, Jason Alan Fries, Matthew B. A. McDermott, Jungwoo Oh, Tom Pollard, Nigam Shah, Ethan Steinberg, Michael Wornow, and Robin van de Water. *Medical event data standard (MEDS): Facilitating machine learning for health*. ICLR 2024 Workshop on Learning from Time Series For Health, 2024

2 Robin van de Water, Axel Winter, Max M. Maurer, Felix August Treykorn, Daniela Zuluaga, Bjarne Pfitzner, Igor M. Sauer, and Bert Arnrich. *Combining hospital-grade clinical data and wearable vital sign monitoring to predict surgical complications*. ICLR 2024 Workshop on Learning from Time Series For Health, 2024

3 Robin van de Water, Hendrik Nils Aurel Schmidt, Paul Elbers, Patrick Thoral, Bert Arnrich, and Patrick Rockenschaub. *Yet another ICU benchmark: A flexible multi-center framework for clinical ML*. Proceedings of the Twelfth International Conference on Learning Representations, 2023

3.19 AI in Babies and Beyond, Boom or Boomerang?

Sven Wellmann (Universität Regensburg, DE)

Birth is one of the most critical moments in a person's life. Birth marks the transition from life in the womb (pregnancy) to life outside the womb. The infant's survival and growth depend fundamentally on basic support for months and years. Many disorders affecting the nervous system lifelong originate in early life and in particular in perinatal complications such as neonatal encephalopathy, preterm birth, neonatal sepsis or jaundice.

We will learn how medical examinations of newborn babies are routinely performed immediately after birth and during the first weeks out of the womb, how vital signs indicate healthy body functions and how subtle clinical signs can point towards serious problems that require more complex examination and possibly subsequent treatment. Building on

this, we will discuss areas in the care of pregnant women and newborns where improved diagnostics through the use of computer algorithms could contribute to reducing morbidity and mortality.

The introduction of prediction algorithms and decision support tools based on methods of so-called artificial intelligence (AI) has started in neonatology and paediatrics. We will discuss first use cases and possible benefits in reducing healthcare gaps. However, we will also shed a light on potential risks that may harm the baby's well-being despite the current boom in AI.

4 Working Groups

4.1 AI Monitoring in Clinical Practice

Brett Beaulieu-Jones (University of Chicago, US), Evangelia Christodoulou (DKFZ – Heidelberg, DE), Thomas Gärtner (Technische Universität Wien, AT), Michael Kamp (Universitätsmedizin Essen, DE), Gilbert Koch (Universitäts-Kinderspital beider Basel, CH), Yamuna Krishnamurthy (Phamily – New York, US), Fabian Laumer (Scanvio Medical AG, CH), Christoph Lippert (Hasso-Plattner-Institut, Universität Potsdam, DE), Florian Markowetz (University of Cambridge, GB), Randall Moorman (University of Virginia – Charlottesville, US), Rajesh Ranganath (NYU Courant Institute of Mathematical Science, US), Raul Santos-Rodriguez (University of Bristol, GB), Wouter van Amsterdam (University Medical Center Utrecht, NL), Robin Van de Water (Hasso-Plattner-Institut, Universität Potsdam, DE), and Julia E. Vogt (ETH Zürich, CH)

Our working group has identified a fundamental challenge in post-deployment monitoring of clinical artificial intelligence (AI) systems: a misalignment between incentives and resources that undermines effective oversight. Those with the greatest interest in ensuring AI safety and efficacy – clinicians, researchers and patients – often lack the necessary funding, technical infrastructure and institutional support. Meanwhile, AI vendors and large healthcare systems, which have these resources, frequently lack strong incentives to engage in long-term monitoring.

To address this issue and align stakeholder interests, introducing regulatory mandates, standardised monitoring metrics and financial incentives may be necessary. Creating clear reporting requirements, real-world performance evaluations and publicly accessible monitoring databases appears particularly promising for enhancing transparency and trust in clinical AI tools. Achieving these objectives will likely require tight collaboration between regulators, developers and healthcare providers, with the goal of establishing best practices and ensuring continuous oversight. Without such measures, AI-driven healthcare solutions risk inconsistent safety and effectiveness, ultimately limiting their long-term benefit to patient care.

4.2 Human Factors in Clinical AI Design and Deployment

Michael Brudno (University of Toronto, CA), Jeff Clark (IngeniumAI – Bath, GB), James Fackler (Johns Hopkins University – Baltimore, US), Maia Jacobs (Northwestern University – Evanston, US), Patricia Reis Wolfertstetter (KH Barmh. Brüder Klinik St. Hedwig – Regensburg, DE), Kacper Sokol (ETH Zürich, CH), and Sven Wellmann (Universität Regensburg, DE)

Our working group reviewed key criteria for selecting clinical problems where artificial intelligence (AI) can have the most meaningful impact; we also focused on human factors that are fundamental to ensuring safe and effective deployment of AI in clinical practice. One important point is the role of biological plausibility. Should AI operation always align with known and accepted medical knowledge, or can models be trusted even if these mechanisms are unclear? Additionally, clinician trust tends to depend on both accuracy and explainability of AI, but how much weight should be given to each remains an open question.

Another key consideration is how to integrate AI into clinical workflows. Can AI systems be adapted to existing medical workflows, or should they be designed to drive (beneficial) workflow changes over time? Crucially, AI could play a role in operational improvements – such as staffing predictions and workflow optimisation – while balancing feasibility and impact. Also, the characteristics of the clinical challenge for which an AI solution is envisaged need to be considered. For example, should AI development focus on supporting (and possibly automating) routine decisions, allowing clinicians to devote their attention to more complex cases? Moreover, we explored opportunities for AI chatbots in collection of patient history, provision of feedback to clinicians and general decision support; nonetheless, questions remain about how to define their limits and handle sensitive topics.

From the human factors perspective, ensuring graceful AI failure modes and designing intuitive handover protocols are of paramount importance so that operators can easily identify such cases and handle them appropriately. A related issue is the need for AI to recognise when it encounters unfamiliar cases and transition control back to human oversight without disrupting provision of care. Additionally, where and how AI-generated alerts should appear in a clinician's workflow remains an open question. Their role is also unclear: should they be advisory, mandatory or something in-between?

Past failures of deployed clinical AI systems highlight the risks of over-reliance on these tools, especially without clear understanding of their limitations. Moreover, we need to consider professional and cultural barriers. For example, how can we ensure that AI benefits all types of healthcare professionals, from nurses to senior physicians? In this context, another important open question is how AI can facilitate teamwork, particularly in patient handoff between clinicians; should AI simply provide information, or should it actively suggest next steps?

Participants

- Brett Beaulieu-Jones
University of Chicago, US
- Michael Brudno
University of Toronto, CA
- Evangelia Christodoulou
DKFZ – Heidelberg, DE
- Jeff Clark
IngeniumAI – Bath, GB
- James Fackler
Johns Hopkins University – Baltimore, US
- Thomas Gärtner
Technische Universität Wien, AT
- Maia Jacobs
Northwestern University – Evanston, US
- Michael Kamp
Universitätsmedizin Essen, DE
- Gilbert Koch
Universitäts-Kinderspital beider Basel, CH
- Yamuna Krishnamurthy
Phamily – New York, US
- Fabian Laumer
Scanvio Medical AG, CH
- Christoph Lippert
Hasso-Plattner-Institut, Universität Potsdam, DE
- Florian Markowetz
University of Cambridge, GB
- Randall Moorman
University of Virginia – Charlottesville, US
- Rajesh Ranganath
NYU Courant Institute of Mathematical Science, US
- Patricia Reis Wolfertstetter
KH Barmh. Brüder Klinik St. Hedwig – Regensburg, DE
- Raul Santos-Rodriguez
University of Bristol, GB
- Kacper Sokol
ETH Zürich, CH
- Wouter van Amsterdam
University Medical Center Utrecht, NL
- Robin Van de Water
Hasso-Plattner-Institut, Universität Potsdam, DE
- Julia E. Vogt
ETH Zürich, CH
- Sven Wellmann
Universität Regensburg, DE