Why Is it so difficult to realise medical AI?

Tariq Andersen
Associate Professor
Computer Science, University of Copenhagen

AI in Healthcare - Institute for Futures Studies
24th of May, 2023

UNIVERSITY OF COPENHAGEN

10x

increase in PubMed publications with "AI" in the title between 2018-2020

Cabitza, F. et al.(2020). Bridging the "last mile" gap between AI implementation and operation:"data awareness" that matters. *Annals of translational medicine*, 8(7).

Approved AI/ML based medical devices in the USA and Europe during 2015–2020

222 in USA (FDA)

240 in EU (CE)

Muehlematter et al. (2021). Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. The Lancet Digital Health, 3(3), e195-e203.

AL/ML is promising (in the laboratory)

npj | Digital Medicine www.nature.com/npjdigitalmed

ARTICLE OPEN

Deep learning algorithm predicts diabetic retinopathy progression in individual patients

Filippo Arcadu[1,2], Fethallah Benmansour[1,2], Andreas Maunz[1,2], Jeff Willis[3,4], Zdenka Haskova[3,4,7*] and Marco Prunotto[2,5,6,7*]

The global burden of diabetic retinopathy (DR) continues to worsen and DR remains a leading cause of vision loss worldwide. Here, we describe an algorithm to predict DR progression by means of deep learning (DL), using as input color fundus photographs (CFPs) acquired at a single visit from a patient with DR. The proposed DL models were designed to predict future DR progression, defined as 2-step worsening on the Early Treatment Diabetic Retinopathy Diabetic Retinopathy Severity Scale, and were trained against DR severity scores assessed after 6, 12, and 24 months from the baseline visit by masked, well-trained, human reading center graders. The performance of one of these models (prediction at month 12) resulted in an area under the curve equal to 0.79. Interestingly, our results highlight the importance of the predictive signal located in the peripheral retinal fields, not routinely collected for DR assessments, and the importance of microvascular abnormalities. Our findings show the feasibility of predicting future DR progression by leveraging CFPs of a patient acquired at a single visit. Upon further development on larger and more diverse datasets, such an algorithm could enable early diagnosis and referral to a retina specialist for more frequent monitoring and even consideration of early intervention. Moreover, it could also improve patient recruitment for clinical trials targeting DR.

npj Digital Medicine (2019)2:92 ; https://doi.org/10.1038/s41746-019-0172-3

INTRODUCTION

Vision loss due to diabetic eye disease is on the rise and it is expected to reach epidemic proportions globally in the next few decades. In 2017, ~425 million people worldwide had diabetes, and this number is estimated to increase to 642 million by 2040.[1] Diabetic retinopathy (DR) is the most common and insidious

The purpose of this work was to go beyond the use of DL for DR diagnostics[15–17,19] and to assess the feasibility of developing DCNNs operating on 7-field CFPs that can predict the future threat of significant DR worsening at a patient level over a span of 2 years after the baseline visit.

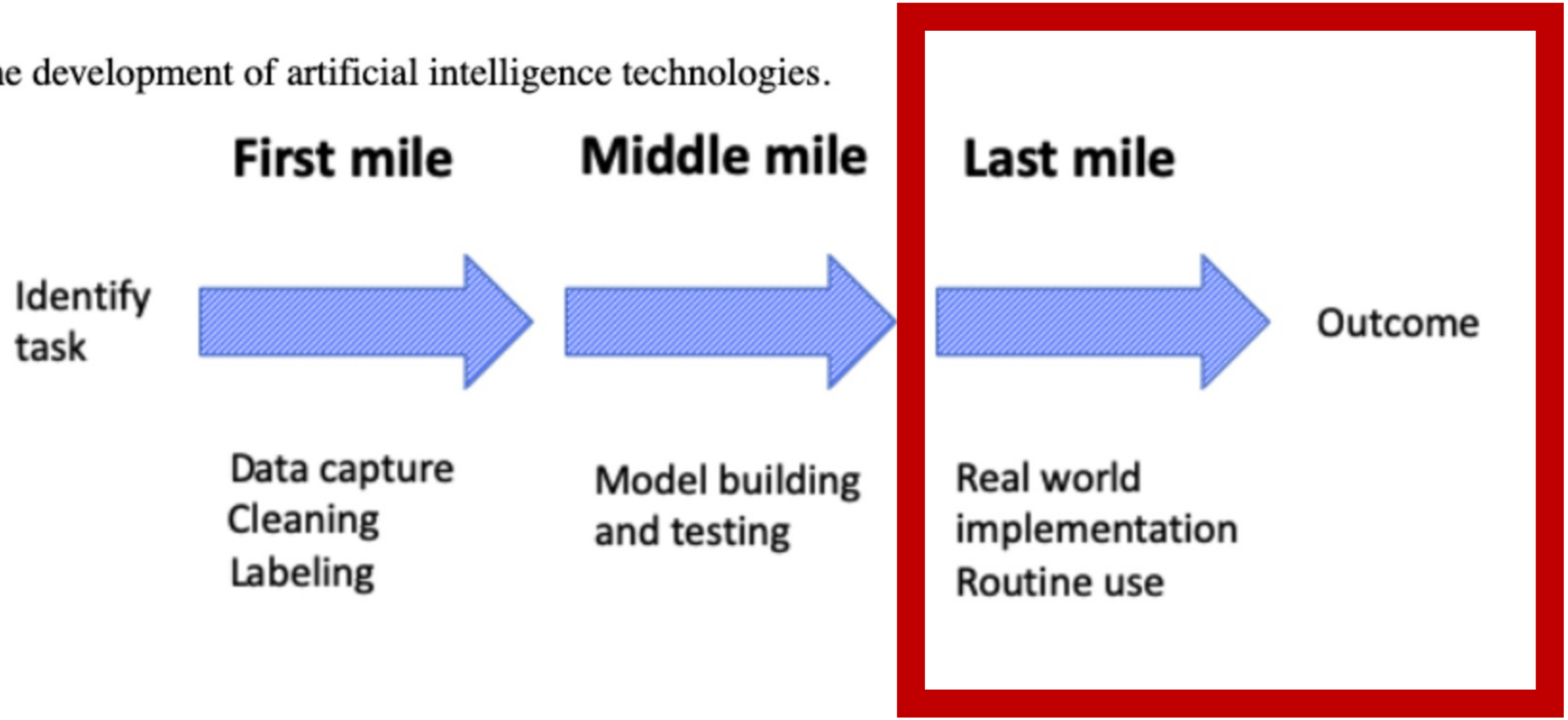
To achieve that, our DCNNs have been trained on high-

Google Research: First Deep-Learning system in the wild

Beede, E. et al. (2020). A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-12)

The Last Mile: Where AI Meets Clinical Reality

Figure 1. Three stages in the development of artificial intelligence technologies.



Coiera, E. (2019). The last mile: where artificial intelligence meets reality. *Journal of medical Internet research*, 21(11), e16323.


HJERTEALARM
NYT REDSKAB SKAL HJÆLPE LÆGERNE

AI/ML prediction of cardiac arrest (in the laboratory)

ESC European Society of Cardiology
Europace (2019) **21**, 268–274
doi:10.1093/europace/euy257

CLINICAL RESEARCH
Sudden death and ICDs

Predicting electrical storms by remote monitoring of implantable cardioverter-defibrillator patients using machine learning

Saeed Shakibfar[1], Oswin Krause[1], Casper Lund-Andersen[2], Alfonso Aranda[3], Jonas Moll[1], Tariq Osman Andersen[1], Jesper Hastrup Svendsen[2,4], Helen Høgh Petersen[2†], and Christian Igel[1]*

[1]Department of Computer Science, University of Copenhagen, Universitetsparken 5, Copenhagen 2100, Denmark; [2]Department of Cardiology, The Heart Centre, Rigshospitalet, University of Copenhagen, Copenhagen, Denmark; [3]Medtronic, Medtronic Bakken Research Center, Maastricht, The Netherlands; and [4]Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark

Received 9 July 2018; editorial decision 6 October 2018; accepted 10 October 2018; online publish-ahead-of-print 30 November 2018

Aims Electrical storm (ES) is a serious arrhythmic syndrome that is characterized by recurrent episodes of ventricular arrhythmias. Electrical storm is associated with increased mortality and morbidity despite the use of implantable cardioverter-defibrillators (ICDs). Predicting ES could be essential; however, models for predicting this event have never been developed. The goal of this study was to construct and validate machine learning models to predict ES based on daily ICD remote monitoring summaries.


V-V Interval (ms)
VT Rx 1 Burst
PHILIPS
lenovo

Risk of ventricular fibrillations or tachycardia within 30 days

1

❌ Alarm Raised

2

44.0% chance of ventricular tachycardia or fibrillation within 30 days

Observations increasing likelihood of arrhythmia:

observations
ventricular tachycardia episodes in zone 1 previous week

Observations decreasing likelihood of arrhythmia:

observations
no electrical storm on day
no ventricular tachycardia episodes in zone 1 on day
no atps on day
no ventricular tachycardia episodes in zone 1 previous day
no ventricular tachycardia episodes previous week

3

AI/ML Usefulness: Dependent on Clinical Context

"The AI-alarm's prediction is something that might make me react a little more aggressively. If our patient schedule is fully booked, both today and tomorrow, and the day after tomorrow, but on Friday we have time. Then I kind of have to make a trade off if I really want to spare him a shock" (electrophysiologist)

Why is ML/AI **uniquely difficult** to design, develop and implement in clinical contexts?

Clinician-Facing AI in the Wild: Taking Stock of the Sociotechnical Challenges and Opportunities for HCI

HUBERT D. ZAJĄC, University of Copenhagen, Denmark
DANA LI, Copenhagen University Hospital, Denmark
XIANG DAI, University of Copenhagen, Denmark
JONATHAN F. CARLSEN, Copenhagen University Hospital, Denmark
FINN KENSING and TARIQ O. ANDERSEN, University of Copenhagen, Denmark

Artificial Intelligence (AI) in medical applications holds great promise. However, the use of Machine Learning-based (ML) systems in clinical practice is still minimal. It is uniquely difficult to introduce clinician-facing ML-based systems in practice, which has been recognised in HCI and related fields. Recent publications have begun to address the sociotechnical challenges of designing, developing, and successfully deploying clinician-facing ML-based systems. We conducted a qualitative systematic review and provided answers to the question: "How can HCI researchers and practitioners contribute to the successful realisation of ML in medical practice?" We reviewed 25 eligible papers that investigated the real-world clinical implications of concrete clinician-facing ML-based systems. The main contributions of this systematic review are: (1) an overview of the technical aspects of ML innovation and their consequences for HCI researchers and practitioners; (2) a description of the different roles that ML-based systems can take in clinical settings; (3) a conceptualisation of the main activities of medical ML innovation processes; (4) identification of five sociotechnical interdependencies that emerge from medical ML innovation; and (5) implications for HCI researchers and practitioners on how to mitigate the sociotechnical challenges of medical ML innovation.

CCS Concepts: • **Human-centered computing → HCI theory, concepts and models;**

Additional Key Words and Phrases: Systematic review, machine learning, artificial intelligence, clinician-facing systems, health, real-world, implementation, conceptual framework

ACM Reference format:

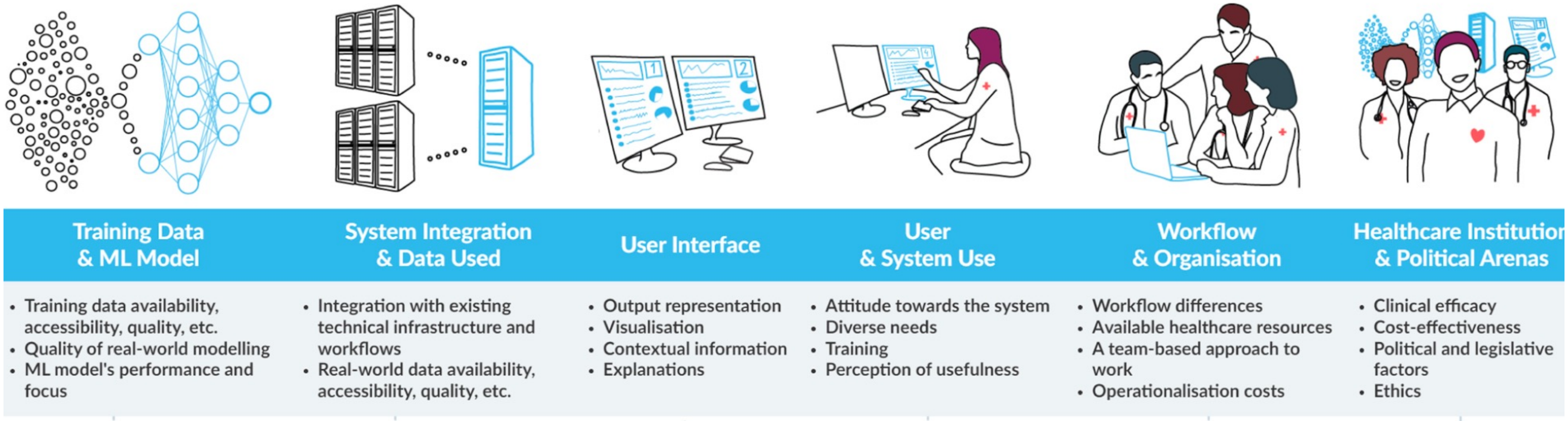
Hubert D. Zając, Dana Li, Xiang Dai, Jonathan F. Carlsen, Finn Kensing, and Tariq O. Andersen. 2023.

9.672 -> 25

Hubert D. Zając, Dana Li, Xiang Dai, Jonathan F. Carlsen, Finn Kensing, and Tariq O. Andersen. 2023. Clinician-Facing AI in the Wild: Taking Stock of the Sociotechnical Challenges and Opportunities for HCI. *ACM Trans. Comput.-Hum. Interact.* 30, 2, Article 33 (2023), 39 pages. https://doi.org/10.1145/3582430

CHALLENGES:
5 SOCIOTECHNICAL INTERDEPENDENCIES

Hubert D. Zając, Dana Li, Xiang Dai, Jonathan F. Carlsen, Finn Kensing, and Tariq O. Andersen. 2023.
Clinician-Facing AI in the Wild: Taking Stock of the Sociotechnical Challenges and Opportunities for HCI.
ACM Trans. Comput.-Hum. Interact. 30, 2, Article 33 (April 2023), 39 pages.
https://doi.org/10.1145/3582430


Training Data & ML Model
• Training data availability, accessibility, quality, etc.
• Quality of real-world modelling
• ML model's performance and focus
System Integration & Data Used
• Integration with existing technical infrastructure and workflows
• Real-world data availability, accessibility, quality, etc.
User Interface
• Output representation
• Visualisation
• Contextual information
• Explanations
User & System Use
• Attitude towards the system
• Diverse needs
• Training
• Perception of usefulness
Workflow & Organisation
• Workflow differences
• Available healthcare resources
• A team-based approach to work
• Operationalisation costs
Healthcare Institution & Political Arenas
• Clinical efficacy
• Cost-effectiveness
• Political and legislative factors
• Ethics

“That was a lot of work to type in all that sh-t and generate that number, and that’s not that helpful.”

Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable AI: Fitting intelligent decision support into critical, clinical decision-making processes. In Proceedings of the Conference on Human Factors in Computing Systems (CHI’19). Association for Computing Machinery, New York, NY, 1–11.
DOI:https://doi.org/10.1145/3290605.3300468

Clinicians:

“The quality of the EHRs collected in Chinese hospitals were much worse than those of the MIMIC dataset”

Zhuochen Jin et al. 2020. CarePre: An intelligent clinical decision assistance system. ACM Trans. Comput. Healthc. 1, 1 (3 2020). DOI: https://doi.org/10.1145/3344258

1 Training Data & ML Model <-> System Use

Poor training data generated “concerns” during in use

- Quantity, consistency, and comprehensiveness
- Data not considered during modelling (such as individual patient information, social stressors e.g. unemployment or loss of a family member)

Inadequate ML models led to “confusion”

- Does not “understand” the clinician’s job
- General and simplistic predictions did not provide new insights
- Undermined its perception of usefulness

“For me to have to track them both down to give them that information would be burdensome and that’s what would get in the way of flow in the [emergency department].”

Sahil Sandhu et al.. 2020. Integrating a machine learning system into clinical workflows: Qualitative study. J. Medic. Internet Res. 22, 11 (11 2020). DOI:https://doi.org/10.2196/22421

2 System Integration & Data Used <-> Workflow

Poor performance when integrated increased the workload

- **Availability of data** during production led to poor performance
- **Poor timing** of ML output decreased its usefulness
- **Additional time spend** and feeling "overburdened"
- Over-utilisation of healthcare resources
- Difficulties with aligning with team-based work

Nurses said that they became very careful about documentation due to the thoughts that the data they entered will be used to infer risks of falling

Insook Cho and Insun Jin. 2019. Responses of staff nurses to an EMR-based clinical decision support service for predicting inpatient fall risk. In Studies in Health Technology and Informatics, Vol. 264. IOS Press, 1650–1651. DOI:https://doi.org/10.3233/SHTI190579

3. User Interface <-> User & System Use

Missing or poor explanations negatively affected the use of the systems

- **Trust deteriorated** due to poor explainability and black-box issues
- **Poor presentation** of the ML output created interpretation issues

Missing contextual patient information next to the ML output decreased its usefulness

Clinical accountability challenged by insufficient explainability

Interactive ML models ("too captivating") led to **confirmation bias**

“We (doctors) spend years in school to learn how to make [a diagnosis based on those [traditional] statistical tools and diagrams… your tool is obviously more informative but we just need more time to get familiar with it”

Zhuochen Jin et al. 2020. CarePre: An intelligent clinical decision assistance system. ACM Trans. Comput. Healthc. 1, 1 (3 2020). DOI: https://doi.org/10.1145/3344258

“Just like with all other new technology based on machine learning: the first 2 months I sit and read through to see what I have, but in month 3, I will look at the [ML] output alone. Because then I trust that it has pulled out what is appropriate [...]”

Matthiesen, S. et al. (2021). Clinician preimplementation perspectives of a decision-support tool for the prediction of cardiac arrhythmia based on machine learning: near-live feasibility and qualitative study. JMIR human factors, 8(4), e26964.

4. User & System Use <-> ML-based System

Clinicians’ attitudes and feelings about machine learning affected usefulness

- “Fear of overstepping,” “feeling uncomfortable,” “resistance to change”, and “sceptical attitudes”

Diverse clinical roles and diverse needs

Lack of end-user training and promotion

- Unfamiliarity with the ML-based systems
- Insufficient computer literacy among clinical end-users

Some nurses felt the need to "warn" patients that they would need to travel should a referral be given. Given the far distance and inconvenience of getting to Pathum Thani Hospital, 50% of patients at clinic 4 opted out of participating in the study

Emma Beede et al. 2020. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI'20). ACM, New York, NY, 1–12. DOI:https://doi.org/10.1145/3313831.3376718

5. Healthcare Institution & Political Arenas <-> ML-based System

Medical professional/academic arenas affected the scope of possibilities

- Weak consensus on definitions of clinical diagnosis
- Clinical guidelines differed across clinical sites
- Hard to reach mutual agreement about ML way of modelling disease

Demonstrating clinical efficacy/cost-effectiveness were critical for acceptance and adoption (Alignment with existing local reimbursement)

Political and legislative factors were decisive for the success

- Adherence to external regulation emerged as an issue during deployment in clinical environments.

Ethics considerations affected the overall perceived usefulness

Training Data & ML Model

- Training data availability, accessibility, quality, etc.
- Quality of real-world modelling
- ML model's performance and focus

System Integration & Data Used

- Integration with existing technical infrastructure and workflows
- Real-world data availability, accessibility, quality, etc.

User Interface

- Output representation
- Visualisation
- Contextual information
- Explanations

User & System Use

- Attitude towards the system
- Diverse needs
- Training
- Perception of usefulness

Workflow & Organisation

- Workflow differences
- Available healthcare resources
- A team-based approach to work
- Operationalisation costs

Healthcare Institution & Political Arenas

- Clinical efficacy
- Cost-effectiveness
- Political and legislative factors
- Ethics

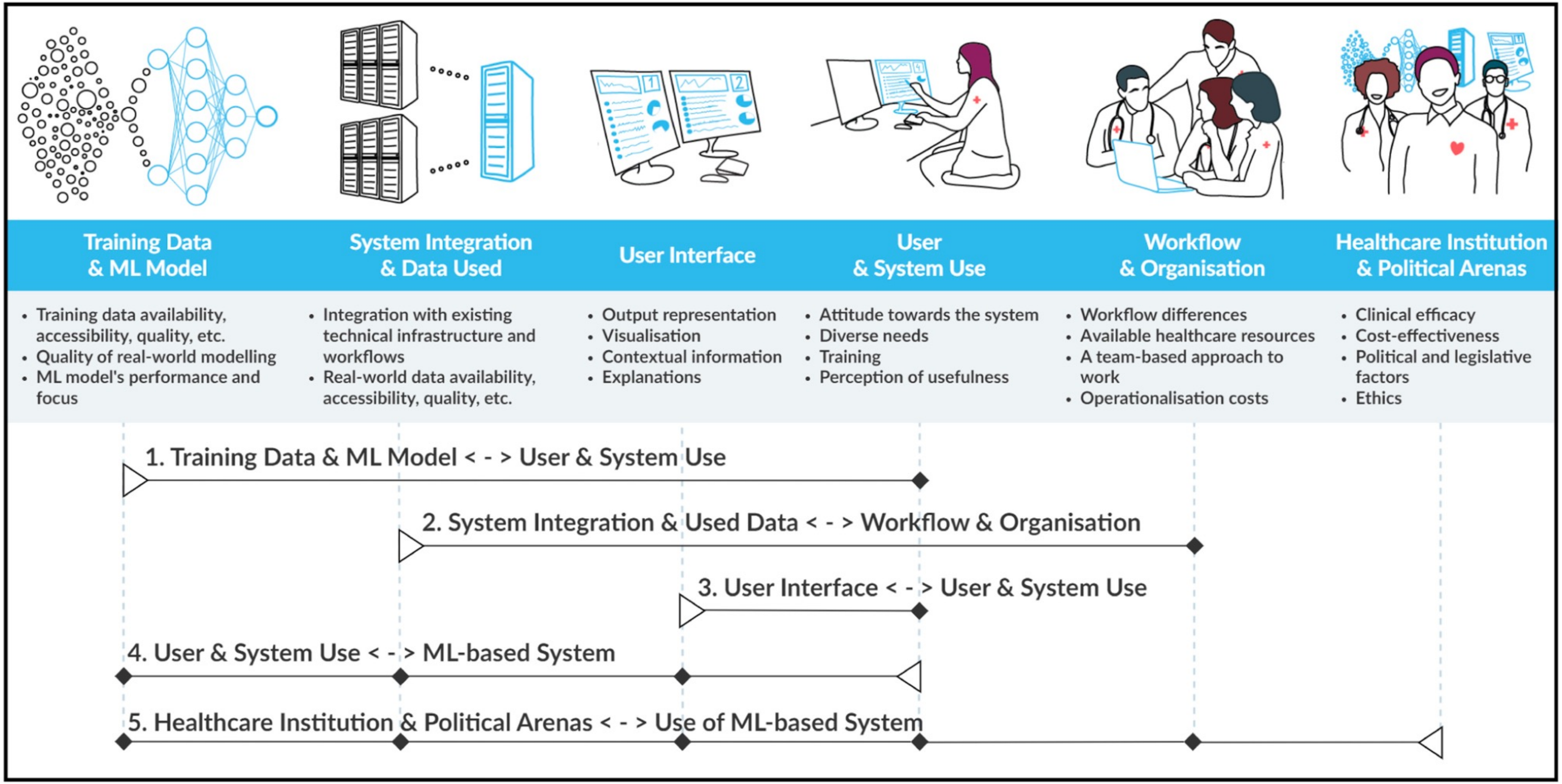


Fig. 6. Five sociotechnical interdependencies of clinician-facing ML-based systems' innovation.

Frodi, D. et al. (2021). Rationale and design of the SafeHeart study: Development and testing of a mHealth tool for the prediction of arrhythmic events and implantable cardioverter-defibrillator therapy. *Cardiovascular digital health journal*, 2(6), S11-S20.

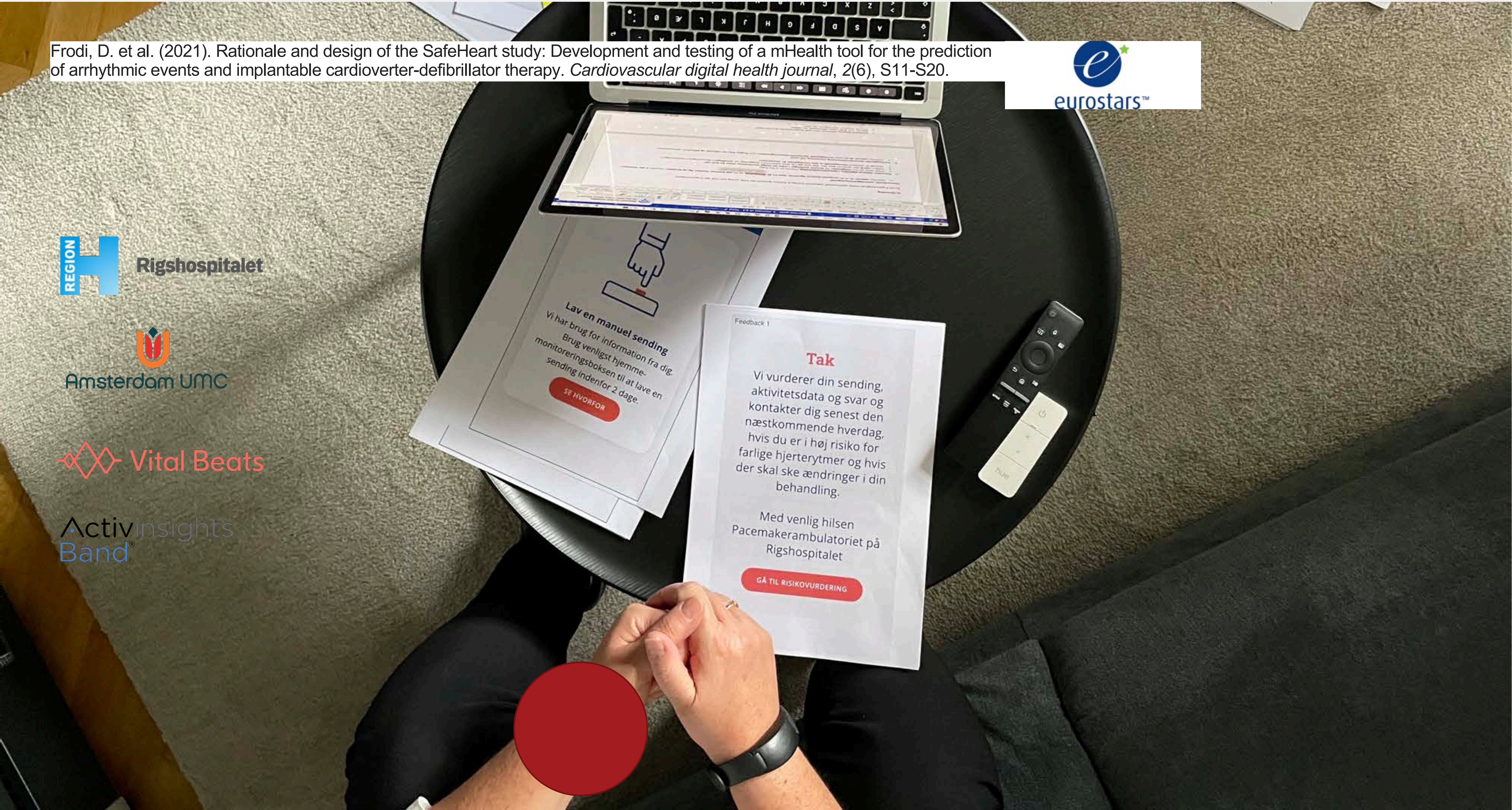
eurostars™

REGION H Rigshospitalet

Amsterdam UMC

Vital Beats

ActivInsights Band®



AI4XRAY
Design Interventions in Kenya & Denmark

Evaluate input data

Provide an impression

Reflect on the AI output

Envision future clinical use

ACM Transactions on Computer-Human Interaction

Human-centred AI in healthcare: Challenges appearing in the wild

Artificial intelligence (AI) holds great promise to improve our healthcare systems. Current AI-based systems already support drug development, triage, screening, diagnosis, and patient follow-up; and the potential is to completely reconfigure the way our healthcare systems work. In the face of such strong potential, many scientists have devoted their research to the development of AI-based systems, and PubMed has seen a tenfold increase in the number of publications mentioning AI, in the past two years alone. Nevertheless, and despite the strong optimism, few AI-based systems have been integrated in everyday care.

Going the "last mile" with AI-based systems will require not only robust algorithms, but also dealing with implementation challenges and guaranteeing that AI-based system fit the needs and practices of patients, carers, clinicians, and clinical researchers. In other words, we need Human-Computer Interaction and a Human-Centred perspective to help unleash the AI potential in healthcare.

The HCI community has often been seen as contributing at the margins of the creation of AI-based systems for healthcare. As the majority of research studies focused on demonstrating technical feasibility and performance of algorithms, they could draw on retrospective datasets and refrain from involving users. However, as AI-based systems start to integrate healthcare systems, there is a pressing need for exploring questions related to: i) What constitutes Human-Centred AI in healthcare? ii) How to design AI-based interactive systems for healthcare? and iii) How to deploy and evaluate AI-based systems in practice and what are the sociotechnical and ethical implications for the human end-users?

Guiding the next generation of HCI work on AI for healthcare are the contributions of the community to Explainable AI, as well as the studies on accountability, transparency, fairness, and ethics. The few ethnographic studies describing AI-based systems use in practice will also be useful in illuminating the assessment of these type solutions in context. Still, important work is yet to come as designing Human-AI interactions is extremely complex due to the variety of workflows triggered by different datasets, the current lack of iterative prototyping tools, and the difficulties of communicating AI capabilities to the design team.

Topics:

- Ethnographies that unpack the use, appropriation, and other sociotechnical aspects of AI-based systems in healthcare, in self-care, clinical care, or clinical research;
- Expectations, perspectives, and misalignments between users and developers;
- Theoretical discussions of key concepts appearing in the Human-Centred AI literature for healthcare, including explainable AI, accountable AI, as well as fairness or ethics in AI;
- Theoretical discussions (re-)visiting key HCI concepts in the space, including patient-clinician interaction, shared decision-making, or self-care;
- Applications and designs that explore and advance Human-Centred AI in healthcare;
- Methodologies, methods, approaches, or adaptations needed for creating appropriate human-AI interactive systems;
- Reviews of existing research on the design, integration, and/or evaluation of ML and AI technology in healthcare.

Special Issue Editors

Tariq Osman Andersen, University of Copenhagen and Vital Beats ApS

Francisco Nunes, Fraunhofer Portugal AICOS

Lauren Wilcox, Google and Georgia Tech

Thanks ✋
Questions?
tariq@di.ku.dk