# Off the hAIgh horse: Realistically Supporting Dermatologists with Automated Reporting

Hubert Dariusz Zając
hdz@di.ku.dk
University of Copenhagen
Copenhagen, Denmark

Niels Kvorning Ternov
niels.kvorning.ternov@regionh.dk
Copenhagen University Hospital, Herlev-Gentofte
Herlev, Denmark

Tariq Osman Andersen
tariq@di.ku.dk
University of Copenhagen
Copenhagen, Denmark

## ABSTRACT

Generative AI (genAI) is increasingly being investigated for healthcare reporting. However, current approaches often focus on automating clinical tasks like diagnosing and treatment suggestions, compounding AI inaccuracies and limiting adoption. This study investigates how automated reporting, guided by dermatologist interpretation, can integrate into clinical workflows. We developed a prototype using React.js and OpenAI API to generate structured dermatology reports, based on a preferred report structure elicited from prior work with primary care providers (PCPs). Seven dermatologists from Denmark, Sweden, and the US participated in iterative design sessions, with additional sessions planned. Preliminary findings show: (1) varied perceptions of dermatologists' roles, (2) low information density diminishes report usefulness, (3) splitting diagnostic and reporting tasks and supporting editing increases acceptance and perceived ownership over generated medical reports, (4) benefits of automated reporting in reducing cognitive load, ensuring inclusion of PCP-centred information, and improving communication tone.

## KEYWORDS

clinical documentation, communication, human-centered AI, generative AI, LLM, integration, CSCW

**ACM Reference Format:**
Hubert Dariusz Zając, Niels Kvorning Ternov, and Tariq Osman Andersen. 2025. Off the hAIgh horse: Realistically Supporting Dermatologists with Automated Reporting. In *Proceedings of (CHI '25 Workshop on Envisioning the Future of Interactive Health).* Yokohama, Japan, 5 pages.

## 1 INTRODUCTION

The challenges faced by healthcare systems today are multifaceted, driven by the ever-growing demand for services and the limited capacity of medical staff to meet these needs. This imbalance has led to mounting workloads, heightened stress, and widespread burnout among healthcare providers [15]. Artificial Intelligence (AI) advancements sparked a new wave of ideas to address these critical issues [11, 20]. Recent breakthroughs in generative AI (genAI), including Large Language Models (LLMs), have further expanded the scope of possible support [17]. Notably, the development of AI tools for generating medical reports (e-consults and consultant letters), has been broadly investigated. The goals behind automated reporting span more efficient operations, enhanced communication between specialists, general practitioners, and patients, ensuring that patients receive timely and well-coordinated care [6, 14, 17, 22].

However, successful incorporation of AI-generated text into clinical practice poses significant challenges pertaining to the factuality, accuracy, and reliability of the produced texts. To address those in radiology, Yu et al. [18] proposed RadGraph F1 as a benchmark to evaluate AI-generated reports against radiologist-reviewed standards. Building on these efforts, tools like MAIRA and its successor MAIRA-2 [1, 14] have focused on automating chest X-ray reporting, with frameworks such as RadFact introduced to assess report correctness and completeness [1]. In a qualitative study, Zaretsky et al. [22] explored the idea of improving clinical communication by using genAI to write patient-friendly discharge summaries. They found several safety-critical omissions and inaccuracies. Such factual inaccuracies, including hallucinations typical of LLMs [10], highlight the significant challenges confronting designers of genAI systems intended for safety-critical settings, especially when AI is responsible for both the diagnostic and reporting work.

In the Human-Computer Interaction (HCI) community, the use of AI in writing has also been a subject of inquiry. Researchers tackled the issue of AI-mediated communication [7], explored questions of creativity [5], and challenged the very definition of authorship [4]. For example, Lehman et al. [9] linked the text's continuous generation mode to increased editing and a weaker feeling of ownership in comparison to prompted generation. Zhou and Sterman [23] showed how imperfect intermediate AI suggestions encouraged creative rewriting, which they discussed to support ownership and creative expression. Other researchers explored more high-level writing support. For example, Dang et al. [3] developed a system that offered an external perspective through AI-generated summaries, which helped them refine their texts. Similarly, ABScribe by Reza et al. [12] supported writers in exploring AI-generated text variations, reducing workload and simplifying the revision process.

However, incorporating genAI-based writing support can also lead to negative outcomes. A study by Singh et al. [13] demonstrated how authors make *integrative leaps*, meaning they perform extra editing work to integrate AI-generated content into their writing. A striking illustration of AI's impact on writers was identified by Bhat et al. [2]. They found that authors included AI recommendations in their texts, even when they did not agree with them, resulting in altered writing strategies and increased distractions. These findings highlight the complex interplay between incorporating AI into


*CHI '25 Workshop on Envisioning the Future of Interactive Health, April 27th, 2025, Yokohama, Japan*

the writing process, emphasising both the potential for enhanced creativity and efficiency as well as the challenges of ownership, cognitive effort, and unintended influence on ideas development.

In this study, we set out to address both aspects of using AI in clinical text generation: healthcare-focused factuality and appropriateness and HCI-focused interaction space to support ownership and reduce cognitive effort and unintended consequences, which was captured by our research question: *What are the requirements for the effective integration of automated reporting into dermatologists' clinical work to enhance communication with general practitioners and support professional practice?*

To answer it, we developed a prototype of a teledermatology system that uses zero-shot and few-shot prompting with GPT4o-mini and GPT4o to generate dermatology reports. This work directly follows a co-design study with primary care providers (PCPs) aimed at understanding their informational needs and the deficiencies in the dermatology reports they currently receive [21]. Findings from that study allowed us to create a report template used by GPTs to generate the reports and address the information gaps discovered. Importantly, rather than automating the entire diagnostic process, our approach preserves the dermatologist's interpretative expertise while generating reports based on their input and embedding them with visual links to source information. Within the same prototype, we implemented several interaction techniques using LLMs for amending the reports: *simplifying, itemising,* and *rewriting to include.* Altogether, we aimed to examine the qualities necessary to achieve clinical usefulness through assessing perceived effort, editing affordances, dealing with inaccuracies, and the overall acceptability of automated report generation as part of teledermatology practice.

We conducted an iterative design process across eight design sessions with seven radiologists from Denmark, Sweden, and the US, during December 2024 and January 2025 to create a stable version of the prototype. Preliminary findings from the sessions include:

- Contrasting perspectives on the role of dermatologists, influencing expectations for report content and format;
- Low information density, not format, was the main issue with the generated text, leading to low use of text-editing tools;
- None of the participants considered their ownership over the report diminish through the use of AI;
- Potential time and perceived effort savings compared to typing were noted, though the comparison to dictation requires further investigation.

Our findings highlight unique usability challenges in automated clinical writing, distinguishing it from more conventional text editing. We see the potential for AI to support tasks requiring little to no medical expertise. Much like transcription software has done for medical reporting before, generative AI may help streamline reporting workflows. However, more investigation is needed to understand the challenges faced in practice. By participating in the Envisioning the Future of Interactive Health workshop, we hope to exchange insights and design approaches with other researchers working on genAI-based interactive healthcare systems and receive guidance on our ongoing study with dermatologists.

## 2 METHOD

In this study, we committed to quick development iterations intertwined with design sessions with healthcare providers, inspired by our previous work on incorporating AI support across various clinical contexts in radiology [19]. Thus, after a short development period (one month), we started conducting sessions with dermatologists (two weekly). HDZ conducted the sessions online, with participation from the prototype developer and TOA (three sessions each). Every week, we updated the prototype to address issues mentioned by the participants and include new functionalities.

The sessions started with participants filling in a questionnaire about their work and AI experience, and signing an informed consent (see Table 1). Next, we moved to work with the prototype, where we simulated a teledermatology workflow (see Figure 1). We started with a quick overview of the prototype, the task, and the currently available interaction techniques to edit the AI-generated reports. Our participants were asked to simulate reporting six teledermatology cases. For each case, the participants examined the available images and clinical history and were asked to assess the diagnostic form pre-filled by the second author - a medical doctor. However, participants were required to review all the information, ensure its accuracy, and make any necessary updates. After the diagnostic form had been approved, it was used to generate a dermatology report according to the PCP-derived template. Finally, dermatologists were asked to make any necessary changes to finalise the report and, hypothetically, approve it for submission to the referring PCP. During the whole session, we asked questions related to the actions taken and our observation guide that covered the following topics: current teledermatology practice, perceived effort, psychological ownership, report structure, experiences with AI, feedback on interaction techniques, and challenges and opportunities with automated reporting. The sessions took place over Zoom, were recorded, and automatically transcribed.

We are following abductive thematic analysis [16], as our previous engagements with primary care providers on eliciting preferred dermatology report structure and experiences with designing interactive AI-based systems cannot be reasonably decoupled from the analysis. Based on those experiences, after transcription and familiarisation, we created the initial codebook consisting of the following concepts: perceived usefulness, workflow (with and without AI), challenges and barriers, ownership and responsibility, opportunities, and other (to include any other relevant excerpts).

## 3 FINDINGS

While our work is still ongoing, in this section, we will explore the preliminary themes conceptualised during the analysis.

*Contrasting perspectives on the role of dermatologists, influencing expectations for report content and format.* We observed two general approaches to reporting: one emphasising the referring PCP's access to domain expertise and the other highlighting dermatologists' active participation in the diagnostic process. This contrast was well captured by P4, who stated, *The expertise of the dermatologist is in the diagnosis and the [recommended] action. How that is presented to the GP doesn't change the [dermatologist's] expertise.* In contrast, P3 framed reporting as a more collaborative effort, saying, *I want*

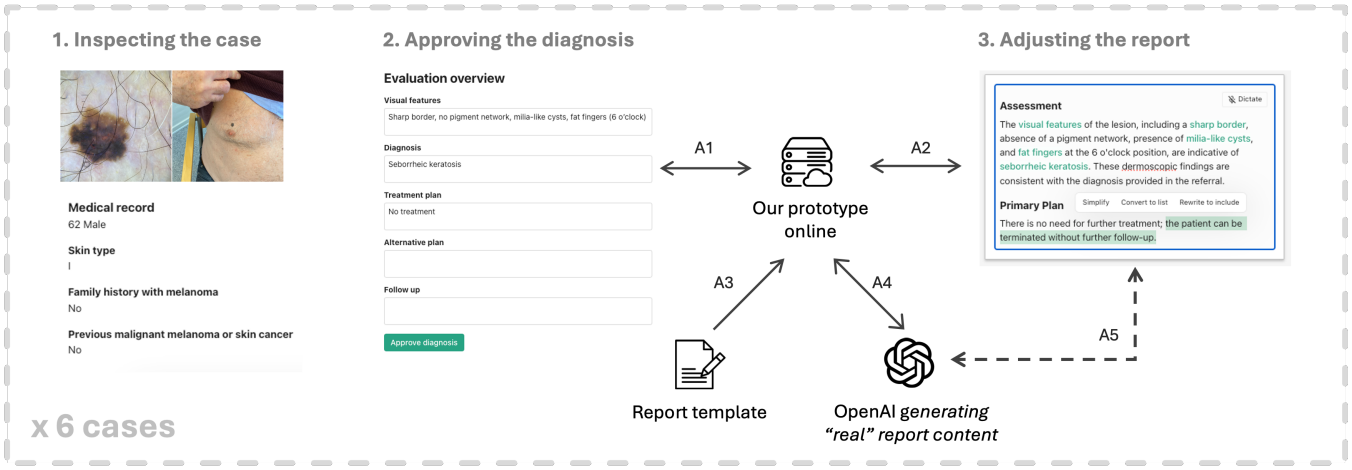


**Figure 1: Design session steps with screenshots from case 1.**

| # | Practice experience | Completed consults | Place of practice | AI at work | AI attitude | Country |
|---|---|---|---|---|---|---|
| P1 | 20y | 1k | University hospital | NA/NT | 4 | Sweden |
| P2 | 15y | 5k | University hospital | NA/T | 5 | Sweden |
| P3 | 16y | 5k | University hospital | A/SU | 5 | USA |
| P4 | 16y | 1k | University hospital | A/SU | 5 | Denmark |
| P5 | 30y | 1k | University hospital | NA/NT | 4 | Denmark |
| P6 | 23y | 5k | Private practice & University hospital | A/SU | 4 | Denmark |
| P7 | 20y | 5k | Private practice | NA/T | 3 | Denmark |

**Table 1: Overview of participants, their years of experience as dermatologists, approximate number of written dermatology consults, place of practice, AI availability at the workplace: [NA – not available | A – available]/[NT – not tested any | T – tested some systems | SU – sporadic use], expectation of AI support at work [0 – not at all, 5 – transformative effect], and the country of practice.**

*it to sound very collaborative. I don't want to sound like I'm a specialist telling primary care doctors what to do. I want this to be like a learning loop, a team.* These perspectives illustrate the potential challenges when introducing a PCP-centric reporting structure and the opportunity for automated reporting to increase the usefulness of medical reports for the end recipients.

*Low information density, not format, was the main issue with the generated text, leading to low use of text-editing tools.* The most common complaint about the generated reports was their excessive length. We identified two key reasons for this issue. First, P6 highlighted the relationship between the information entered in the diagnostic form and the length of the generated report, *I think the report was a little too long, compared to the information I gave.* We observed that the report paragraphs often remained consistent in length regardless of the amount of input provided, sometimes resulting in unnecessary filler text. The second reason stemmed from the model's extreme caution, which frequently led to additional suggested checks and follow-ups. These recommendations were often unnecessary, as noted by P2, *It's too long. I can appreciate that you need to put some of these, like regular skin checks, but... [a] 6-month follow-up? I wouldn't recommend here.* These findings suggest that the precautionary and stylistic measures embedded in generative AI, may be limiting their applicability when used by domain experts.

*None of the participants considered their ownership over the report diminished through the use of AI..* Dermatologists considered automated reporting using genAI just another tool, as P1 explained, *It's a system that puts nice text for me.* This could have been linked to the fact that AI was not responsible for diagnosing the patient, instead, the expertise remained with the specialist. Moreover, our participants always had the ability to manually compose the reports to their liking, as per P2, *I feel like I'm the owner, as I [entered] the keywords, and especially [because] I removed all the yapping.* These inputs suggest that separating the diagnostic and reporting aspects of clinical practice may ease the acceptance of automated medical text generation.

*Potential time and perceived effort savings compared to typing were noted, though the comparison to dictation requires further investigation.* The workflows of teledermatologists varied, and so did their experiences with automated reporting. On the one hand, genAI can expedite report writing while at the same time maintaining content relevancy for the end recipients. This was noted by P6, *It*

*would take me maybe five minutes to write that [nowadays, without the system,] and this report to the general practitioner is much more useful actually, generated based on the keywords.* On the other hand, when dermatologists used other supporting technologies, like speech-to-text dictation software, the immediate benefits on the author's side were not so apparent. *I would be afraid that … it would be too time-consuming if I had to read it through several times and edit it. I think the difficult thing is that we usually have such short reports to the GP, and it doesn't take long in the first place [using speech to text],* as recalled by P7. This indicates that automated reporting solely to increase dermatologists' efficiency may not always bring the desired value. Instead, HCI and health researchers should explore additional benefits, such as enhancing communication, reducing cognitive load, and improving report usefulness for specialists, PCPs, and patients.

## 4 DISCUSSION AND FUTURE WORK

The preliminary findings of this study offer insights into the intricate dynamics of integrating AI-based automated reporting into dermatologists' clinical practices that affect not only the direct users of the system but also the report recipients and the patients [21]. The reported issue of low information density in AI-generated reports resonates with the concerns raised by Zaretsky et al. [22] regarding the potential for safety-critical omissions and inaccuracies in AI-generated summaries. This convergence underscores a crucial challenge in the application of AI for medical reporting: ensuring the generated content is not only accurate but also sufficiently detailed and relevant for clinical decision-making.

Moreover, the study's observations on dermatologists' perceptions of ownership and control in AI-assisted reporting expands the broader HCI research on the impact of AI on writing processes. Specifically, the finding that dermatologists did not report a diminished sense of ownership when using AI to generate reports echoes the insights of Lehman et al. [8]. This suggests that when AI is implemented as a supportive tool that augments rather than replaces professional input, it can be integrated into clinical workflows without compromising the clinician's sense of agency and responsibility. Retention of the expertise work, in our case, diagnosing and recommending treatment, appears to be a critical factor in fostering a sense of ownership and enabling clinicians to incorporate AI assistance in a way that aligns with their professional judgment and expertise.

This manuscript explored preliminary findings from our ongoing work on automated reporting in teledermatology. In the next step, we plan to investigate dermatologists' perceived efficiency and ownership when using automated reporting, as well as barriers and opportunities ahead. By presenting our work at the Envisioning the Future of Interactive Health workshop at CHI, we hope to engage with researchers developing AI-based healthcare systems, gain insights into best practices for designing AI-mediated clinical documentation, and inform the next steps of our study based on the feedback. We hope for those discussions to shape the next phase of our research and ensure that our findings contribute to the HCI and Health community.

## REFERENCES


[1] Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, Mercy Ranjit, Shaury Srivastav, Julia Gong, Noel C. F. Codella, Fabian Falck, Ozan Oktay, Matthew P. Lungren, Maria Teodora Wetscherek, Javier Alvarez-Valle, and Stephanie L. Hyland. 2024. MAIRA-2: Grounded Radiology Report Generation. https://doi.org/10.48550/arXiv.2406.04449 arXiv:2406.04449 [cs].

[2] Advait Bhat, Saaket Agashe, Parth Oberoi, Niharika Mohile, Ravi Jangir, and Anirudha Joshi. 2023. Interacting with Next-Phrase Suggestions: How Suggestion Systems Aid and Influence the Cognitive Processes of Writing. In *Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23)*. Association for Computing Machinery, New York, NY, USA, 436–452. https://doi.org/10.1145/3581641.3584060

[3] Hai Dang, Karim Benharrak, Florian Lehmann, and Daniel Buschek. 2022. Beyond Text Generation: Supporting Writers with Continuous Automatic Text Summaries. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. ACM, Bend OR USA, 1–13. https://doi.org/10.1145/3526113.3545672

[4] Fiona Draxler, Anna Werner, Florian Lehmann, Matthias Hoppe, Albrecht Schmidt, Daniel Buschek, and Robin Welsch. 2024. The AI Ghostwriter Effect: When Users do not Perceive Ownership of AI-Generated Text but Self-Declare as Authors. *ACM Transactions on Computer-Human Interaction* 31, 2 (April 2024), 1–40. https://doi.org/10.1145/3637875

[5] Jonas Frich, Michael Mose Biskjaer, and Peter Dalsgaard. 2018. Twenty Years of Creativity Research in Human-Computer Interaction: Current State and Future Directions. In *Proceedings of the 2018 Designing Interactive Systems Conference*. ACM, Hong Kong China, 1235–1257. https://doi.org/10.1145/3196709.3196732

[6] Katherine E. Goodman, Paul H. Yi, and Daniel J. Morgan. 2024. AI-Generated Clinical Summaries Require More Than Accuracy. *JAMA* 331, 8 (Feb. 2024), 637–638. https://doi.org/10.1001/jama.2024.0555

[7] Jeffrey T Hancock, Mor Naaman, and Karen Levy. 2020. AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations. *Journal of Computer-Mediated Communication* 25, 1 (March 2020), 89–100. https://doi.org/10.1093/jcmc/zmz022

[8] Florian Lehmann and Daniel Buschek. 2024. Functional Flexibility in Generative AI Interfaces: Text Editing with LLMs through Conversations, Toolbars, and Prompts. https://doi.org/10.48550/arXiv.2410.10644 arXiv:2410.10644 [cs].

[9] Florian Lehmann, Niklas Markert, Hai Dang, and Daniel Buschek. 2022. Suggestion Lists vs. Continuous Generation: Interaction Design for Writing with Generative Models on Mobile Devices Affect Text Length, Wording and Perceived Authorship. In *Mensch und Computer 2022*. ACM, Darmstadt Germany, 192–208. https://doi.org/10.1145/3543758.3543947

[10] Florian Leiser, Sven Eckhardt, Valentin Leuthe, Merlin Knaeble, Alexander Mädche, Gerhard Schwabe, and Ali Sunyaev. 2024. HILL: A Hallucination Identifier for Large Language Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3613904.3642428

[11] Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J. Topol. 2022. AI in health and medicine. *Nature Medicine* 28, 1 (Jan. 2022), 31–38. https://doi.org/10.1038/s41591-021-01614-0 Publisher: Nature Publishing Group.

[12] Mohi Reza, Nathan M Laundry, Ilya Musabirov, Peter Dushniku, Zhi Yuan "Michael" Yu, Kashish Mittal, Tovi Grossman, Michael Liut, Anastasia Kuzminykh, and Joseph Jay Williams. 2024. ABScribe: Rapid Exploration & Organization of Multiple Writing Variations in Human-AI Co-Writing Tasks using Large Language Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3613904.3641899

[13] Nikhil Singh, Guillermo Bernal, Daria Savchenko, and Elena L. Glassman. 2023. Where to Hide a Stolen Elephant: Leaps in Creative Writing with Multimodal Machine Intelligence. *ACM Trans. Comput.-Hum. Interact.* 30, 5 (Sept. 2023), 68:1–68:57. https://doi.org/10.1145/3511599

[14] Shaury Srivastav, Mercy Ranjit, Fernando Pérez-García, Kenza Bouzid, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Harshita Sharma, Maximilian Ilse, Valentina Salvatelli, Sam Bond-Taylor, Fabian Falck, Anja Thieme, Hannah Richardson, Matthew P. Lungren, Stephanie L. Hyland, and Javier Alvarez-Valle. 2024. MAIRA at RRG24: A specialised large multimodal model for radiology report generation. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, Dina Demner-Fushman, Sophia Ananiadou, Makoto Miwa, Kirk Roberts, and Junichi Tsujii (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 597–602. https://doi.org/10.18653/v1/2024.bionlp-1.50

[15] Melissa Suran. 2023. Overworked and Understaffed, More Than 1 in 4 US Nurses Say They Plan to Leave the Profession. *JAMA* 330, 16 (Oct. 2023), 1512–1514. https://doi.org/10.1001/jama.2023.10055

[16] Jamie Thompson. 2022. A Guide to Abductive Thematic Analysis. *The Qualitative Report* (May 2022). https://doi.org/10.46743/2160-3715/2022.5340

[17] Nur Yildirim, Hannah Richardson, Maria Teodora Wetscherek, Junaid Bajwa, Joseph Jacob, Mark Ames Pinnock, Stephen Harris, Daniel Coelho De Castro,

Shruthi Bannur, Stephanie Hyland, Pratik Ghosh, Mercy Ranjit, Kenza Bouzid, Anton Schwaighofer, Fernando Pérez-García, Harshita Sharma, Ozan Oktay, Matthew Lungren, Javier Alvarez-Valle, Aditya Nori, and Anja Thieme. 2024. Multimodal Healthcare AI: Identifying and Designing Clinically Relevant Vision-Language Applications for Radiology. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–22. https://doi.org/10.1145/3613904.3642013

[18] Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y. Ng, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. 2023. Evaluating progress in automatic chest X-ray radiology report generation. *Patterns* 4, 9 (Sept. 2023). https://doi.org/10.1016/j.patter.2023.100802 Publisher: Elsevier.

[19] Hubert D. Zając. 2024. *It takes a Village to Raise Clinical AI: Towards Clinical Usefulness of AI in Healthcare*. Department of Computer Science, Faculty of Science, University of Copenhagen. https://di.ku.dk/english/research/phd/phd-theses/2024/Thesis_Hubert_Dariusz_Zajac.pdf

[20] 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 Transactions on Computer-Human Interaction* 30, 2 (April 2023), 1–39. https://doi.org/10.1145/3582430

[21] Hubert Dariusz Zając and Niels Kvorning Ternov. 2025. Reframing Automation Benefits: A Human-Centered Approach to Expanding the Value of AI-Generated Reports in Healthcare. *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)* (2025). https://doi.org/10.1145/3706599.3719979

[22] Jonah Zaretsky, Jeong Min Kim, Samuel Baskharoun, Yunan Zhao, Jonathan Austrian, Yindalon Aphinyanaphongs, Ravi Gupta, Saul B. Blecker, and Jonah Feldman. 2024. Generative Artificial Intelligence to Transform Inpatient Discharge Summaries to Patient-Friendly Language and Format. *JAMA Network Open* 7, 3 (March 2024), e240357. https://doi.org/10.1001/jamanetworkopen.2024.0357

[23] David Zhou and Sarah Sterman. 2024. Ai.llude: Investigating Rewriting AI-Generated Text to Support Creative Expression. In *Creativity and Cognition*. ACM, Chicago IL USA, 241–254. https://doi.org/10.1145/3635636.3656187