Designing for Control in Nurse-AI Collaboration During Emergency Medical Calls

Arngeir Berge
arngeir.berge@norceresearch.no
NORCE Norwegian Research Centre
Bergen, Norway

Frode Guribye
frode.guribye@uib.no
University of Bergen
Bergen, Norway

Siri-Linn Schmidt Fotland
sifo@norceresearch.no
NORCE Norwegian Research Centre
Bergen, Norway

Gro Fonnes
grfo@norceresearch.no
NORCE Norwegian Research Centre
Bergen, Norway

Ingrid Hjulstad Johansen
ijoh@norceresearch.no
NORCE Norwegian Research Centre
Bergen, Norway

Christoph Trattner
christoph.trattner@uib.no
University of Bergen
Bergen, Norway

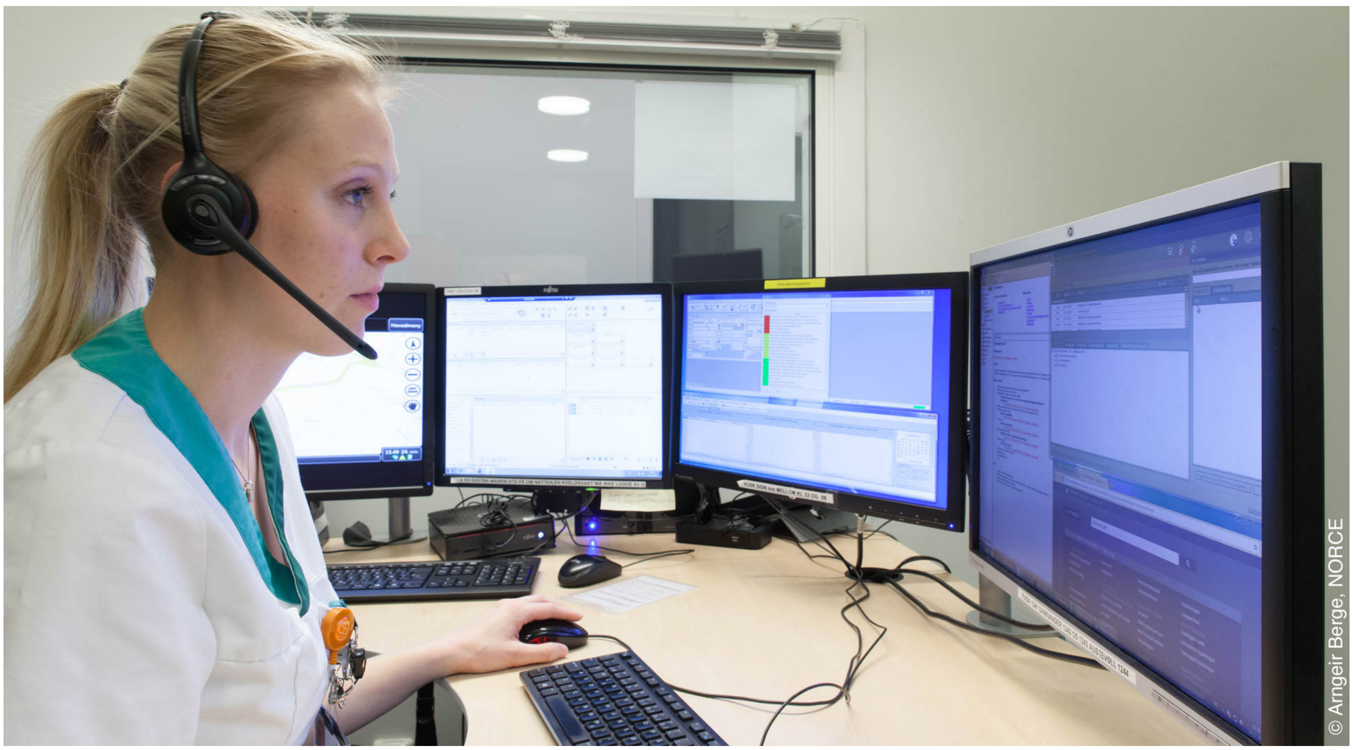

Figure 1: A telenurse at her multi-display workstation in a local emergency medical communication center.

ABSTRACT

AI-powered symptom checkers are automating the work of telephone triage nurses in assessing patient urgency. Yet, these systems exclude several vulnerable patient groups and overlook telenurses' competent interaction with their patients. This study, conducted in collaboration with telenurses, examines how AI can support their clinical assessment and was carried out in four phases: 1) interviews that revealed telenurses' challenge of juggling decision-support and documentation interfaces, 2) a co-design workshop that conceptualized continuous nurse-AI interaction, 3) development of a prototype that suggested questions for nurses to ask callers, and 4) a role-play workshop that demonstrated nurse-AI interaction in practice. The study addresses how we can design for control in human-AI collaboration in order to enhance, rather than replace, human decision-making processes.

DIS '23, July 10–14, 2023, Pittsburgh, PA, USA

ACM ISBN 978-1-4503-9893-0/23/07.
https://doi.org/10.1145/3563657.3596110

CCS CONCEPTS

• **Human-centered computing** → **User interface design**; *User centered design*; *Empirical studies in interaction design.*

KEYWORDS

human–AI interaction, artificial intelligence, telephone triage, telenursing, clinical decision-support systems, intelligent documentation-support systems

ACM Reference Format:

Arngeir Berge, Frode Guribye, Siri-Linn Schmidt Fotland, Gro Fonnes, Ingrid Hjulstad Johansen, and Christoph Trattner. 2023. Designing for Control in Nurse-AI Collaboration During Emergency Medical Calls. In *Designing Interactive Systems Conference (DIS '23), July 10–14, 2023, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3563657.3596110

1 INTRODUCTION

When introducing artificial intelligence into work practice, there is a choice between strategies. One strategy can be to automate the process and let the AI tool have control of tasks and decisions; another can be to aim to support the autonomy of users [7], thereby augmenting their competence. In human-centered AI (HCAI), the latter is emphasized. [75] The two strategies are also present in the telephone triage of patients in emergency medical services. The first strategy is found in AI-driven systems known as symptom checkers. These are automated systems—often in a quasi-conversational form—that perform triage by asking patients a range of questions to assist them by advising on the need for further care [24]. Symptom checkers are becoming widespread in healthcare as a first contact point [53] and can help reduce the call volume to emergency medical communication centers. The information collected by a symptom checker can be made available to telephone triage operators—a term encompassing both registered nurses and non-medical personnel—and be a starting point for further triage. Such hybrid services, however, are not for everyone. McCartney [47] argues that GP at Hand in London, a service that combines a symptom checker with occasional video consultations, enacts exclusion by design; The National Health Service suggests that the service is less accessible for vulnerable groups such as people with learning difficulties, dementia, complex mental health conditions, drug dependence, terminal illness, or even for pregnant women or frail elderly.

In some places, symptom checkers are used by operators during telephone triage, where an operator asks the caller system-generated questions. This workflow requires the operator to adhere strictly to the system's protocols, reduces the operator's autonomy in the process, and undermines the need for clinical competence and judgment. Clinical assessment is left to the formalized and automated process of the system. This threat to professional autonomy has led many healthcare professionals to be negative about the use of symptom checkers in their workflow [40]. Further, such system-led interactions infringe on patients' ability to explain complex matters freely, as system protocols do not always fit the complexities of troubles and needs in real life. Front-line practice in healthcare can be messy and flexibility needs to be built into the design and delivery of tools and services [45].

For AI to reach its potential in healthcare delivery, a sociotechnical approach is recommended where the clinical workflow is systematically considered throughout the design process [66]. Over time, full automation of clinical assessment can lead to skill degradation in operators [56], which may lead to less resilient healthcare systems. Following the second strategy in line with an HCAI approach and, arguably, contrary to that of symptom checkers, AI can be built into tools that are deliberately designed to support the autonomy of telenurses and be a resource for their conversations with patients without imposing a strict script to follow. Call centers with telenurses can serve a broader patient group than symptom checkers alone, providing a service that machines cannot easily replace. Patients value telenurses' service when they meet expectations for active listening, clarity, collaboration, and competence [54].

This study examines ways in which an AI tool can support the operators' and callers' control and autonomy during telephone triage. We followed a call in the human-computer interaction (HCI) community to examine productive forms of AI-human collaboration [62, 75, 76] by deploying a study in four phases. In Phase 1, we interviewed seven triage nurses from five local emergency communication centers in Norway—hereafter referred to as "call centers". We inquired about their workflow and how they used their current interfaces. In addition, and most notably, we asked them to speculate how an AI tool could enhance their workflow in the future. A thematic analysis of the transcribed interviews identified two central interface-related tasks, namely, using decision support and documentation. The participants pointed to the opportunity to merge these two tasks into one interface, using nurse-AI interaction to automate more of the decision support and documentation. In Phase 2, we facilitated a co-design workshop with six telenurses from a single call center. We used different assignments to foster speculation on automated decision support and documentation. The insights were used in Phase 3 to make a prototype to support a future telephone triage workflow. The prototype combined AI-driven decision support and documentation support, functioning as a recommender system that suggested next questions to ask the caller. It organized the documentation thematically, as well as highlighted and pursued potential urgent symptom combinations. The interaction with the prototype was studied in Phase 4, a role-play workshop with six telenurses from four call centers.

This research addresses how we can engage with designing for human-AI collaboration, integrating AI support in high-risk decision-making. We contribute to the discussion on how we can design human-centered AI systems by introducing a design process that focuses on supporting autonomy, control, and competent human actors in the interaction with the systems.

2 BACKGROUND

Telephone triage is increasingly used as a first contact with the emergency healthcare service to prioritize and guide patients to the right level of care at the right time. To assist with clinical assessment, telenurses commonly rely on clinical decision-support systems (CDSSs) and documentation support systems. Symptom checkers are a third class that substitutes some of the telenurses' work, but they cannot provide adequate service for all patients.

2.1 AI Support for Clinical Decisions

CDSSs for telephone triage are traditionally built around a limited number of medical topics like fever, pregnancy, or breathing problems. Telenurses will choose the chapter that corresponds most to the caller's description and then read in it to obtain decision support in mainly two ways: suggested inquiries to clarify the condition and suggested urgency for different combinations of symptoms that patients may experience. A number of initiatives for AI-enhanced CDSSs are geared towards hands-on triage in the hospital emergency rooms [22, 36, 69] for various aims, like predicting urgency level, treatment outcomes, or chance of re-hospitalization. There are fewer AI initiatives for telephone triage. Proof-of-concept studies exist for AI-based urgency prediction based on selected symptoms that operators tick off in a schema [78, 82]. Telephone triage operators' free-form notes have been used successfully to improve predictions of whether patients need transportation [71]. AI initiatives for telephone triage are often limited to identifying single conditions, notably cardiac arrest, myocardial infarction, and sepsis [3, 8, 13, 25]. Also, AI technologies have been proposed to enhance operators' situational awareness, supporting their assessments in various ways. These technologies can include identifying the kind of room the caller is in [57], as well as detecting the caller's emotions, such as fear [79], stress [70] and similar [5, 15, 18]. Other examples are gender classification of callers [2, 68] and detection of deceptive calls [12].

2.2 AI-Support for Clinical Documentation

Written records from patient conversations are used in communication between healthcare providers and can even become part of legal processes. Documentation is an essential task for healthcare professionals, but it is time-consuming and can take away focus from the conversation. Therefore, AI initiatives have sought to ease the documentation burden. AI-enabled contextual autocomplete can increase the quality and reduce the amount of time required for data entry [26]. In some countries, human scribes are still employed to transcribe doctors' spoken notes [17]. However, the use of digital scribes is increasing, leveraging advancements in speech recognition, natural language processing, and AI to automatically document spoken elements of the clinical encounter [16]. Improvements in speech summarization can potentially create a new generation of digital scribes that auto-generate clinical records from speech. However, the evidence for the effectiveness of such systems remains limited [60].

Interviews and co-design workshops have been used in a primary care study to understand the potential roles of future AI-based documentation support in general practitioners' consultations [37]. The study showed how general practitioners wanted to be in control of the clinical assessment and be able to overrule AI suggestions in case of disagreement. The clinicians were worried that automation bias would cause them not properly to evaluate the AI output and highlighted their own writing process as a way to structure thinking. The researchers argue that automated documentation systems should be designed for human-AI collaboration rather than aiming for full automation of the process where an AI system substitutes the human.

2.3 Support Versus Substitution

The two classes of systems described above have been followed by a third class called symptom checkers. "Symptom checkers are artificial intelligence-supported software tools that use a conversational 'chatbot' format to support rapid diagnostics and consistent triage." [53] These are mainly patient-facing [24], and the clinical assessment is formalized and built into the system. The bots acquire an overview of the patient's situation, and a documentation summary is provided along with an urgency prediction and advice for how to seek further help. However, these automated assessments do not give adequate service to all patient groups [47]. In addition to patient groups mentioned in the introduction, other groups can be ill-served by symptom checkers, such as individuals with reading difficulties, those with comorbidities that affect their urgency, and diverse or generally underrepresented groups [80], as well as people in distress or need of immediate support. Barriers that young people perceive to using the tools are lack of internet access, low health literacy, and low technology literacy [1]. Such barriers disproportionately affect minorities and people with lower income [29]. The accuracy of symptom checkers is questioned [17, 81], and the sensitivity to emergency cases appears to be particularly low [81]. When the tools recommend seeking emergency assistance, users more often choose too low of an urgency level and opt to wait for an appointment at the primary care physician's office [19].

In some healthcare systems, symptom checkers are used by either medical and non-medical telephone triage operators to guide the triage process [32, 33]. First, the operator checks whether the caller is in a patient group that can be served. After the operator has asked about the main symptoms, the system dynamically selects tailored questions to be asked to the patient by the operator. This results in a system-led conversation [32]. In cases where operators are nurses, their professional competence and clinical assessments may not be sufficiently used as a resource by symptom checkers, as they are supposed to ask the questions proposed by the system. There is little research on operator-facing symptom checkers, but research on CDSSs in call centers shows that the systems can be too rigid and steer the conversations unduly [51]. There is a need to examine how competent nurses can remain in control of the clinical assessment and yet receive AI support both in determining relevant questions and documenting the conversation. Our study addresses this research gap by looking at how we can design a system that gives telenurses flexibility and control in their practice.

Evaluation of AI-enhanced triage tools often focuses on the algorithmic accuracy of predicted triage outcomes [22, 36, 69]. However, in the HCI community, there is a call for developing AI tools that provide high utility and enhance people's work [75]. Giving advice to novices is different from delivering it to experts. For domain experts, decision support should provide information that is useful throughout the decision process instead of just recommending an outcome [41, 61]. All Norwegian telephone triage operators are registered nurses, which creates an opportunity to examine how highly skilled users can use AI in their workflow. Rogers [62] calls for exploring how to develop new AI tools that people can work, create, or solve problems with, while also emphasizing the role of agents that can make proactive suggestions.

There is a growing interest in the HCI community for examining various kinds of interaction between humans and intelligent systems [10, 62]. Xu and Gao [75] propose HCI design goals of developing accountable AI systems that guarantee human-driven decision-making, taking complementary advantages of human intelligence and machine intelligence. We aim to examine together with telenurses what a fruitful human-machine collaboration can look like in their context.

3 METHODS

This study was organized into four phases consisting of both empirical inquiries and design activities. The participants consisted of 17 nurses from 7 call centers across 5 of Norway's 11 counties. Among the participants, 15 were women and only 2 were men, reflecting that men make up only 9% of the Norwegian telenurse population [49]. Two of the telenurses participated in both Phase 1 and Phase 2 and were the only ones involved in more than one phase.

3.1 Phase 1: Interviews

We performed five semi-structured interviews. The interviews included several questions in the form of thought experiments, a speculative approach [6]. According to Auger [4], speculative design does not only enable people to think about the future but to critique current practice. The nurses were first presented with examples of current uses of AI to offer a "perceptual bridge" [4] so they could more easily speculate about AI in their context. In total, seven nurses from five different call centers participated. There were four telenurses and three call center leaders. Two of the leaders were occasionally practicing as telenurses. We did the interviews through a video meeting solution, recording the audio and transcribing it verbatim. We then performed a thematic analysis [11] of the transcriptions. The text was coded with inductive codes like "decision support", "clinical pathways", or "recommendations", often with multiple codes on the same passages. After coding all the text, we revisited the material. Some codes were selected and merged into themes as suggested by Braun and Clarke [11]. We then chose three themes for analysis based on their explanatory value [11] for understanding how the telephone triage workflow is now, how the nurses foresee it in the future, and how an AI tool may bridge this gap. Preliminary findings from the interviews were further pursued in a co-design workshop.

3.2 Phase 2: Co-Design Workshop

Our understanding of AI support from the interviews was a system that suggested relevant inquiry items to ask callers, adjusting the suggestions during the course of the call. Documenting answers to these inquiry items could be as easy as ticking off "yes" or "no". We were curious about how more telenurses would receive these ideas and were also eager to speculate further with them on how an AI tool should be designed to answer specific requirements in telephone triage. Since it was early summer—a busy time of year for call center leaders—we opted to make arrangements with only one leader. The leader was free to organize who would participate and assigned six telenurses. Two of the authors, information science researchers, facilitated the workshop, which lasted four hours and was held in person at the call center. Earlier in the day, we scheduled to see a telenurse workstation on site (similar to that in Figure 1) with the physical layout and software used so that we could draw a schematic overview of the nurses' tools.

Based on initial findings from the interviews, we prepared assignments to aid speculation about the future role of AI in the triage workflow. The participants were asked to document a patient call based on an anonymized call transcript. We discussed similarities between their approaches and then documentation in general. After that, participant groups used sticky notes to envision how facts, symptoms, and affected body parts could be organized visually in an AI tool. They were also asked to draw how a tool could display an urgent symptom pattern. We made cards for potential interface elements, and they sorted them to suit a future workflow. Notes were taken from the workshop, and the resulting artifacts from the participant assignments were photographed for later analysis. The photos, written artifacts, and field notes were structured, and for each assignment, the results were compared between groups or participants to see patterns in the data. The co-design workshop was set up to complement the analysis of interviews by conceptualizing how an AI tool could contribute to a future workflow.

3.3 Phase 3: Designing a Working Prototype

Insights were drawn from the interviews and the co-design workshop to make a working prototype for decision and documentation support. To enable easy confirmation of symptoms and facts, it is recommended to use an interface terminology [63], which can be defined as "a systematic collection of clinically oriented phrases (terms), whose purpose is to support clinicians' entry of patient information into computer programs, such as clinical note capture and decision support tools." [64] Techniques from recommender systems have previously been used for patient conversations [28, 65], and we used such techniques to provide telenurses with continuous suggestions for what they could ask next. Prototyping human-AI interactions is difficult since the exact behavior of AI may be hard to predict or understand for both users and designers [76]. A suggested approach is sketching [76], which we employed in the co-design workshop. We wanted to go a step further and developed the working prototype to study nurse-AI interaction.

3.4 Phase 4: Role-Play Workshop

A role-play workshop was arranged for six telenurses, assigned from three call centers to try out the prototype in person in a lab environment. Since the prototype was at an early stage, we did not want to test it during calls with actual patients. Instead, we used role-play, a method that has been used in HCI to simulate conversations between healthcare professionals and patients [46]. The restricted access to patients and the sensitive nature of calls make role-play a natural choice. Telenurses are particularly adept at playing patients due to their many conversations with patients and because role-play is often part of nurse education. Role-playing is suitable for evaluating prototypes and anticipating the user experience if a prototype were realized [48].

We had asked the participants in advance to make anonymous clinical cases that could be used to simulate calls to the call center. The participants were paired up to impersonate a caller and an operator who used the prototype. We asked the callers to improvise and elaborate on the clinical cases to answer whatever questions the operators may ask. After a pair had played out a clinical case, they would debrief the conversation and the user experience of the prototype. They were asked to point out any observations or problems and whether they could document what they intended.

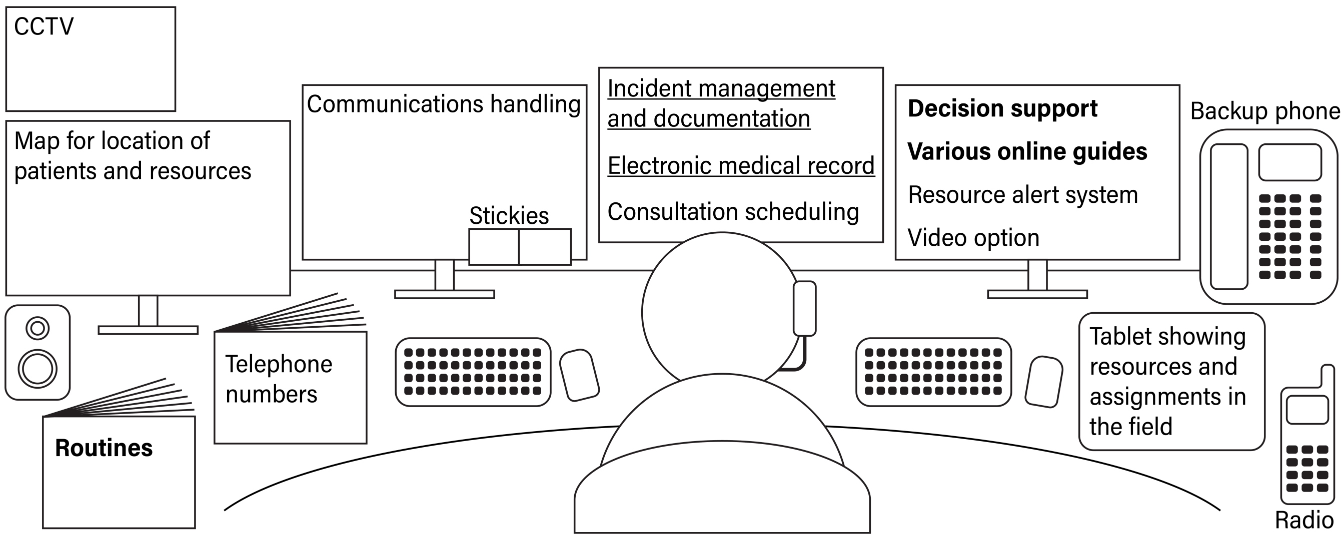

Figure 2: Call triage interfaces arranged after the workstation setup in Figure 1. Interfaces with a clear documenting purpose have underlined names. Interfaces with a clear decision-support purpose have names in bold typeface.

Their voices and the computer screens were recorded to preserve the simulated conversations, the interaction with the prototype, and the debriefings. The pairs changed roles internally to alternate who enacted caller and operator. After a few clinical cases, the participants were rotated and matched with someone new. This way, the experience the previous pairs had gained with the prototype was transmitted to the new pairs. The full-day workshop ended with a half-hour group interview.

A video editing program was used to annotate how the participants interacted with the prototype. We used markers on the timeline to code and describe observations. The codes were reviewed and thematically organized. The recording of the group interview was transcribed and coded for thematic analysis.

4 RESULTS

Based on the interviews and the visit to the telenurse workstation, we drew Figure 2 schematically as an example of interfaces that currently support the telephone triage workflow. The drawing highlights interfaces associated with documentation (underlined names) and the use of decision support (names with bold typeface). Other tasks have their specialized interfaces, forming a complex work environment. The following results detail the current workflow and show how an AI tool could assist during different phases.

4.1 Phase 1: Findings from the Interviews

The interviews gave us background information about the current workflow. Other research has depicted the telephone triage process as iterative [27] and found that callers do not always formulate their problem in an orderly manner [31, 44]. Nonetheless, we describe how we understand the workflow's main phases from the nurses' descriptions as analytical categories, shown in Figure 3. After the telenurses answer a call, they will typically determine who is calling and from where (Personal Data). Then the caller will tell why they are calling (Problem Formulation). Next, the telenurses will lead a discussion about what happened, the symptoms, and the caller's condition (Exploring Symptoms). The nurses will decide on the urgency level of the call and a corresponding response (Response), which is explained to the caller, and any necessary arrangements with the caller are made. If the nurses assess the urgency as low, they will advise the caller about self-care (Nurse Advice). After the call, they will finish the documentation and schedule a consultation if they have asked the caller to visit the emergency primary care center (Follow-Up Work). The interviews revealed that the workflow has two main interface-related tasks: documenting and using CDSSs. The horizontal lines in Figure 3 indicate how these tasks are attended to more (solid lines) or less (dotted lines) during each phase of a call.

Through a thematic analysis, three analytical themes were selected from the interviews: clinical assessment and decision support, documenting the call, and AI interactive suggestions. The themes are presented in the following.

4.1.1 *Clinical Assessment and Decision Support.*

The call centers used any of the three most common CDSSs [52] to aid decisions: NIMN (Norwegian Index for Medical Emergency Assistance) [73], LVI (Index for telephone triage and advice in emergency primary healthcare) [38], MTS TTA (Manchester Triage System—Telephone Triage and Advice) [23]. These systems are not AI-enhanced and are organized with a limited number of chapters for the most important or common problems.

The participants explained how they were supposed to use the CDSS. In the Problem Formulation phase of the workflow, many telenurses would listen to the patient's problem and try to identify the main symptom or condition. They would then open a corresponding chapter in the CDSS, for example *Breathing Problems*. Each chapter contains a list of symptom combinations for choosing

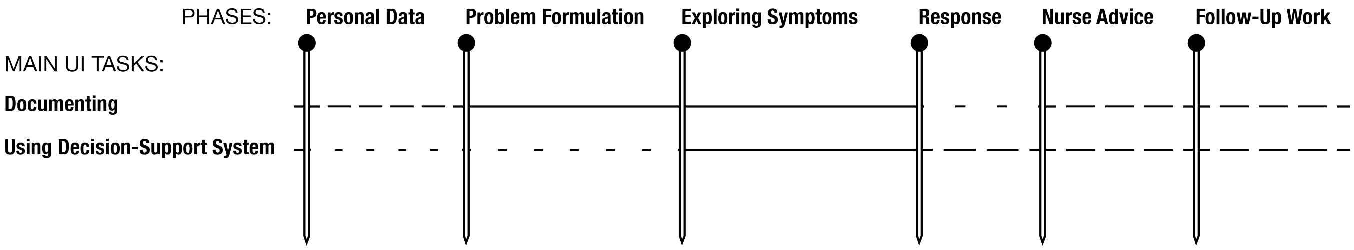

Figure 3: Current triage workflow with phases and main user interface-related tasks. Line segments with smaller dots mean the task is less present in those phases.

a response, like sending an ambulance, offering a doctor's consultation, or offering medical advice. The symptom combinations are listed with the highly urgent at the top and the least urgent at the bottom. In the Exploring Symptoms phase, nurses would ask the patients about symptoms and work their way from the top towards the bottom; they hoped to rule out the highly urgent symptom combinations on top and find a less urgent matching symptom combination further down. This was called to work "reductively", meaning that they asked the patients questions until confident that they could rule out the most urgent symptom combinations in the chosen chapter. For example, a particular symptom combination in the Breathing Problems chapter consisted of the following: Suspected foreign bodies in the airways + Problems breathing. This symptom combination had the corresponding response of sending an ambulance.

Clinical assessment involves any qualified decision about the patient's condition and what the appropriate response is. Below, one of the participants uses the peculiar term of making decisions "in the Index" (meaning the CDSS) and points out a problem in this approach:

> It might be like the decisions you make in the Index, then, that they don't fit perfectly with what the patient has been calling about. That it becomes a bit too vague, perhaps. That you don't get it narrowed down enough.

The participant describes that their assessments must relate what the caller is saying to the symptom combinations listed in a CDSS and that the categories and formal descriptions found there are not necessarily optimal for describing the current case.

4.1.2 Documenting the Call. Telenurses are required by law to document the patient call [43]. The problem formulation, symptoms, response, and advice are typically all documented in the same free-form text input field. The participants found the shifting focus challenging: They must type the documentation at the same time as traversing the CDSS and talking to the patient. During Follow-Up Work they complete unfinished parts of the documentation. The participants said there is a lot of variation in how telenurses document. Some write "long dissertations" so that it is hard for readers to get an overview, others write very little and may omit important information. The participants noted that the documentation was done outside of the CDSS and suggested that the CDSS and the documentation system should be integrated.

4.1.3 AI Interactive Suggestions. During the Exploring Symptoms phase of the workflow, the current CDSSs contain recommended questions for different conditions. Based on the chosen symptom combination the tools also recommend a response for the Response phase. Telenurses first have to make a clinical assessment of the most fitting category in a CDSS and must also navigate to the right chapter to see the recommendations. This can be challenging to the extent that some telenurses avoid using the CDSS during these phases. Also, the task of documenting is currently done manually. AI tools have the capacity to give more proactive support than static CDSSs, and several participants speculated that this could be taken advantage of. AI suggestions was a recurring theme in the interviews. In the following quote, a participant speculates how AI could offer support during a conversation:

> Yes, if the artificial intelligence hears "fainted", right, "fainting" could appear. Then questions could appear: Have you been fainting, or do you feel that you are going to faint? Are you clammy? [...] You tick off and move on, then it goes straight into the journal, right? That grandmother has said she is dizzy and clammy. You tick it off, and then it will conclude that she needs an ambulance, she needs to be seen by a doctor, and then it goes automatically.

Here, the AI tool can identify words in the call, like the keyword "fainted". The tool responds with questions that the telenurse may ask to clarify and get more information about the incident. In the described scenario the telenurse confirms correct information like "dizzy" and "clammy", and the tool will find the warranted healthcare service response and initiate it automatically. When asked about the telenurses' part in the decision or the dispatching of an ambulance, the participant said that the professional nurse would have to check and tick off the output along the way. Such a system is not without risk, though, as worded by another participant:

> If you get too much direction or directions, it is like: "No, this— this is THAT," then one will quickly be caught in the trap of maybe not looking broadly enough. A little bit of both.

This participant sees a risk of being led down the wrong clinical pathway by AI and thereby getting a wrong impression of the caller's condition. Getting "too much" direction seems to counteract that telenurses keep a broad perspective. The participant seems afraid of quickly settling on what the caller's condition is.

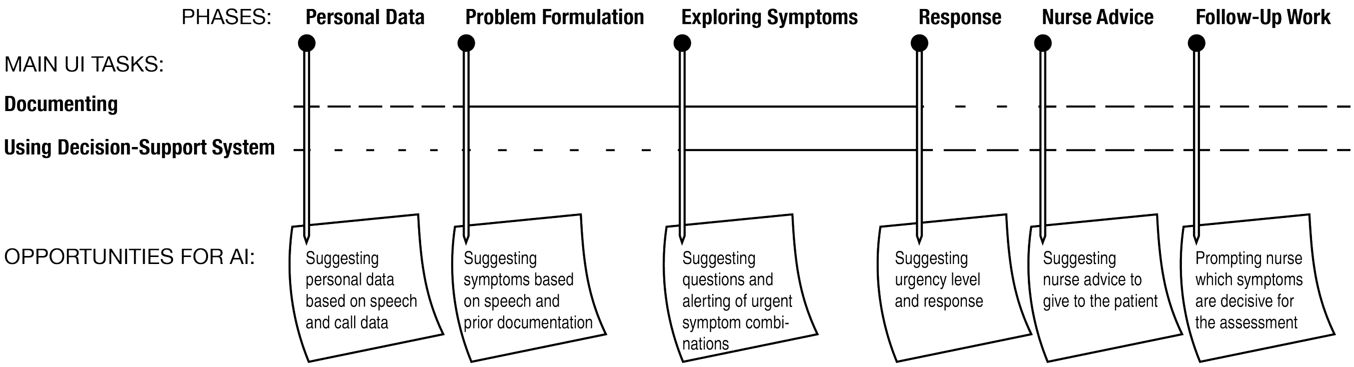

Figure 4: Opportunities for nurse–AI interaction and support in the triage workflow.

From the themes in the analysis, we learned that the participants have two main interface-related tasks (using CDSS and documenting) that run throughout all the phases of the workflow. Figure 4 expands Figure 3 with boxes where we present opportunities for AI interaction for each phase of the triage workflow. In short, the tasks of using decision support and documenting constitute dual duties in the workflow and the participants speculated how these duties can be managed by one interface—one that makes recommendations to support clinical assessment and that automates documentation based on the same assessment. We facilitated a workshop to co-design aspects of such an interface with telenurses, and the results are presented in the following.

4.2 Phase 2: Concepts from the Co-design Workshop

The co-design workshop had activities related to both documentation and the use of decision support.

4.2.1 Documentation. To investigate the current documentation characteristics, we asked the participants to write documentation from a use case the way they normally would in the free-form text input field in their incident management interface. The use case was an anonymized transcript of a call. Since the transcript was the same for all, we could easily compare the resulting documents and discuss further. Several participants said they simplify patient accounts into what may have a clinical significance. Still, they kept some of the nuances that set cases apart. A participant underscored this by saying: "All of my 24,000 notes are different". The participants' documentation showed a sensitivity to the patient's perspective. They said that if they felt the need to make an assessment different from a caller's account, a strategy would be to group statements to drive the point through. For example, a patient might answer ten according to the numeric rating scale (NRS) pain screening tool, which means the patient assesses the pain as *the worst pain imaginable*. If the participants had another impression, they might accompany the pain score with the information: "The patient talks in complete sentences" (which is hard to do when in excruciating pain). The participants said they avoid making explicit statements differing from the patient's account, partly because patients have the legal right to read documentation about themselves.

4.2.2 Decision support. The participants expressed support for an AI tool suggesting relevant questions during the call, saying this could make documentation more efficient. However, they were concerned that the suggestions might lead them down a clinical path and make them miss critical alternative paths. They were skeptical that a tool could suggest all relevant questions. As one participant said: "We are the Whitepages directory of the healthcare services", meaning that people call them with all kinds of requests, including the highly unlikely. In a group assignment to organize interface elements for AI support, one of the groups placed a sticky note "suggestions for matching symptom combinations" early in the workflow, saying that they used the symptom combinations of their current CDSS not only to look for a corresponding response but to get ideas for relevant questions. This was possible because symptom combinations are made up of symptoms that must be clarified to determine whether a combination matches. The participants discussed whether symptom combinations would have a different role in an AI tool. Today telenurses scour symptom combinations during the call. In contrast, the participants asked if not a function of an AI tool would be to identify and suggest a symptom combination that matches the patient's problem and calls for a specific response.

We pursued this further with an activity focusing on pathways of symptom inquiries; see the example in Figure 5. Based on the use case from an actual patient call, we had written down information bits on sticky notes and asked the participants to supply any missing information. The assignment was to organize the information visually on the wall according to how the participants would like it to appear in a future AI tool. They assessed the use case as having medium urgency. Then we made a single change to the use case by telling them to imagine that a symptom in the call was more serious and asked them to visualize how an AI tool should signify an urgent symptom combination. The participants modified the layout with new system-suggested questions and actions.

The workshop assignments encouraged speculation about clinical pathways, where participants envisioned an AI tool that assists telenurses to choose the most fitting pathway by unobtrusive signals. Both interview and workshop participants were concerned with the risk of being led down a misguided pathway by an AI tool. They expressed support for their choices being automatically documented but remained skeptical that a system could provide

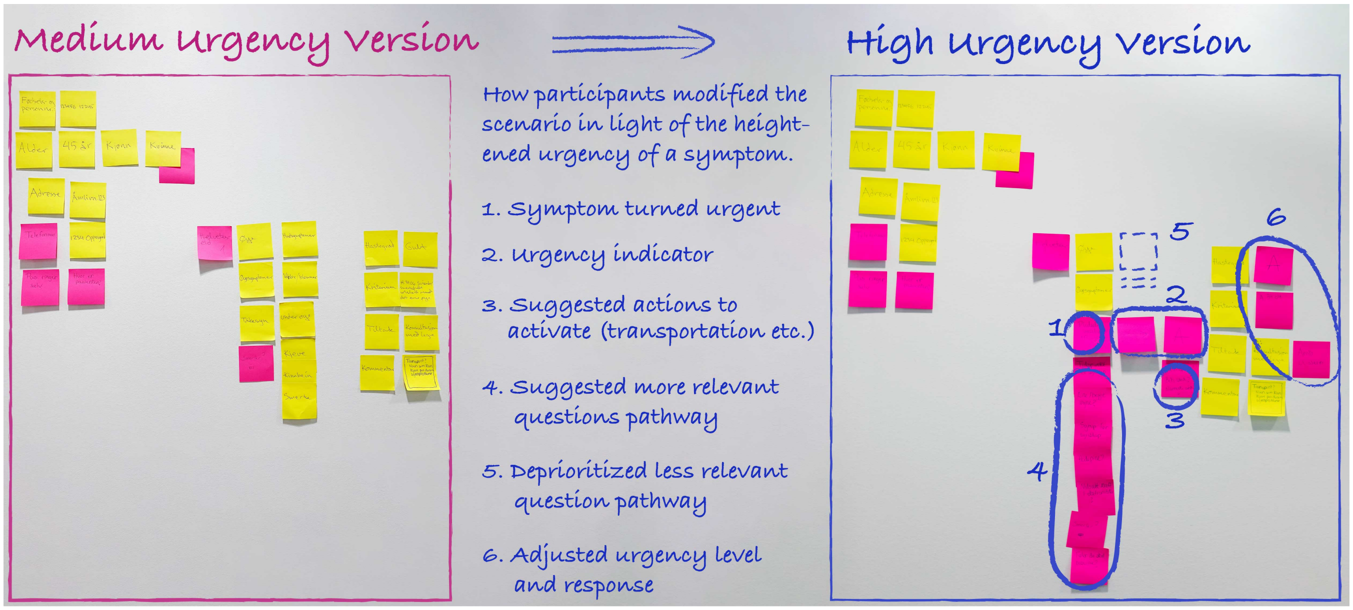

Figure 5: Two human-AI interaction scenarios that were identical up to a certain point in an imagined call. The right-hand scenario took a pivotal turn when an alternative description of a symptom indicated high urgency. Labeled by the authors.

standardized documentation for the full breadth of cases handled by the call center.

4.3 Phase 3: Prototyping for Nurse-AI Interaction

The interviews and co-design workshop results informed the design of a working prototype. The problem of dual duties was tackled by combining decision support and documentation support in the same interface. The prototype was designed to suggest relevant next inquiry items that could easily be confirmed or denied to become part of the structured documentation. Below is a list of some of the prototype's features, corresponding to letters in Figure 6:

a) The left column suggested relevant inquiry items
b) The suggestions could be confirmed or denied
c) Inquiry items that were not in the suggestion list could be searched for
d) The right column contained the progressing documentation
e) Details about an inquiry item were ordered below it
f) Denied inquiry items were marked as such
g) Red squares appeared in front of inquiry items that were part of highly urgent symptom combinations
h) Telenurses could select an urgency level from a drop-down list and overrule the tool if it predicted the highest urgency level, communicated by red squares
i) Response type could be documented from a drop-down list

The participants were concerned that individual chapters in their current CDSS might not cover the patient's presented problem. In contrast, the prototype was designed with the flexibility to tackle complex health problems. It would suggest context-sensitive inquiry items, but all inquiry items were made available to the nurses via a search function. Additional decision support was provided by highlighting combinations of documented symptoms that may raise concerns. The highly urgent symptom patterns (critical combinations) from NIMN, the most used Norwegian CDSS, were coded in a classification and regression tree (CART). This enabled the prototype to effectively highlight known critical combinations close to being present. The documented patient data in nearly present critical combinations were marked with a pink square in front, and the inquiry item suggestions belonging to the identified combinations were also marked with a pink square. A rule parser would verify fully present critical combinations in the documentation and mark the associated patient data with red squares.

The participants reported that callers present a wide range of issues and were worried that not all could be documented in a standardized manner. Therefore, we decided to make a high number of standardized inquiry items. Two of the authors with telehealth experience lead the development of an interface terminology of 1200 inquiry items consisting of questions, symptoms, and conditions, with their corresponding answer options. Inquiry items were written in layperson terms similar to those the nurses typically use in conversations with callers. Most of the inquiry items could simply be confirmed or denied. A group of inquiry items was made to describe symptoms: when they had occurred, if they had become more pronounced over time, etc. As recommended in the literature [63], the terminology included synonyms for each inquiry item so that different search terms would lead to the intended item. The interface terminology was organized hierarchically, with general inquiry items at the top and the more special ones in the branches further down. Any inquiry item appearing as a suggestion or a search result could be traversed in the hierarchy for easier access to what the nurses may look for. A visual body representation enabled symptoms to be localized by clicking a body part.

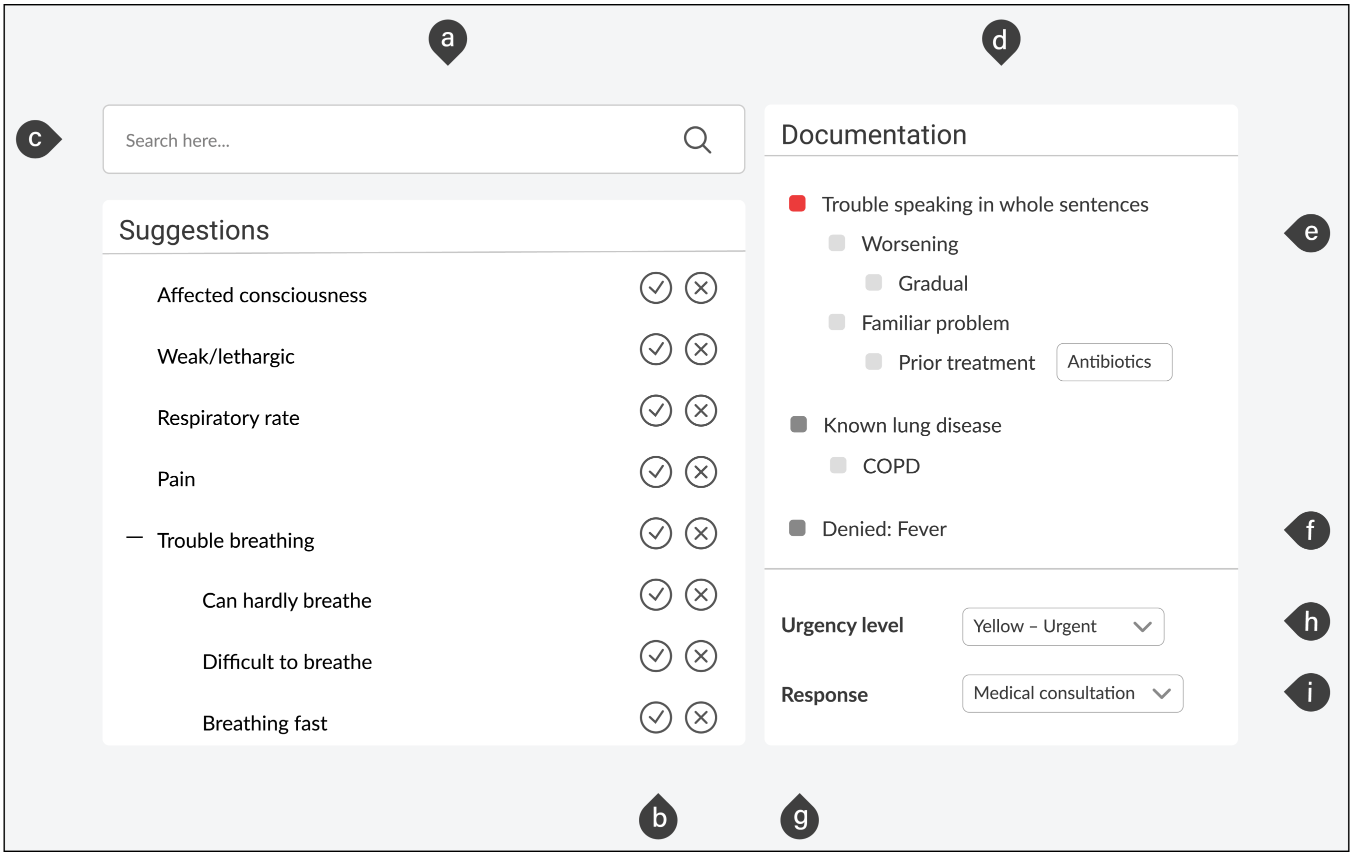

Figure 6: A prototype to facilitate nurse-AI interaction during triage.

We used different techniques known from recommender systems to show nurses the list of suggested inquiries. The list was populated by eight items: two items for identifying highly urgent cases, provided by a CART machine learning model that picked unanswered items from the nearest critical combination; two items from similar calls, derived using cosine similarity with one-hot encoded vectors; two items from the most frequent questions; and two items from already documented topics according to the topical hierarchy of the terminology. The algorithms were implemented in an ensemble to ensure that no item occupied two slots in the list simultaneously. The rationale behind using the different algorithms was as follows: items related to high urgency were presented because the participants were concerned about being led down a low-urgency pathway too early; items related to similar calls reflected how many calls follow a similar pattern; frequent items reflected how some questions, like questions of pain, appear in calls for a variety of problems; and items with topical similarity reflected how humans tend to stay on a topic for a while.

Access to medical records and telephone recordings in the Norwegian public health system is highly restricted. For this reason, we enlisted telenurses to document anonymous cases representative of calls they experienced in their work. Through a web interface, they provided 257 cases with an average of 12 documented inquiry items each. Our approach of using an ensemble of algorithms was chosen to mitigate the sparsity of the case data and still allow telenurses to interact with a system that suggested relevant questions. The findings from the investigation of the interaction between telenurses and the prototype are presented in the following.

4.4 Phase 4: Observations from the Role-Play Workshop

The analysis of the role-play recordings showed lively conversations where the simulated callers improvised and elaborated on the clinical cases. They used their own experience to simulate calls that were not straightforward and could challenge the operator (the participant currently answering the telephone in the role play), reflecting the messy nature of frontline service. The prototype could suggest inquiry items from a wide variety of topics and narrow the suggestions as the conversation progressed. In role-players' debriefings after each simulated call, the pairs often discussed which inquiry items they could not find in the terminology. They also noted which inquiry-item suggestions they would have liked to see during the conversation.

From the list of suggested inquiry items, the operators could expand a question to see more detailed inquiry items below. They actively used this functionality, drilling down in the hierarchy when they wanted to go further into the details of a topic. They typically asked about several of these detailed inquiry items, and when presented together, the operators could confirm or deny them quickly. The participants spoke favorably of how the system allowed them to deny symptoms and document that the denied symptoms at least had been asked about. They noted how part of their job is ruling out serious symptoms and symptom combinations. During debriefing, they expressed that in their regular documentation system, they sometimes overlook recording the inquiry items that were denied by the caller, despite the potential significance of this information for a doctor reviewing the documentation. If there were inquiry items that they wanted to ask about but could not find through search, they pointed this out. And despite all the inquiry items in the terminology, they wanted it to be easier to provide free-form supplementary notes for all inquiry items if needed. Mental health problems were seen as particularly challenging to document and requiring free-form text supplementation. Moreover, they would like a future system to suggest nurse advice because much of the advice is routinely given.

Regarding the presentation of the documentation, the participants valued that the patient information was grouped thematically. They would also like it to be ordered according to urgency, with the most serious patient data on top. If they disagreed with the system's order, they wanted to be able to change it. For calls where pink or red squares appeared in the documentation, the operators sometimes interrupted their simulation to comment on the markings and whether they could make sense of them. They would typically click the squares to have the prototype provide a list of inquiry items for clarifying the critical combinations at issue. After role-play conversations, both operator and caller often looked through the latest list of suggested inquiry items. Several times they noticed items they wished the operator had asked about. The participants said that interacting with the AI tool was a new way of working and that they would need practice to keep the flow of the conversation.

The thematic analysis of the group interview at the end of the role-play workshop resulted in four main themes emerging. These aligned with the observations from the role-play and debriefing recordings as described above. The themes were as follows: missing inquiry items that they would like to add to the terminology, the missing ability to reorganize the documentation according to what they thought was the most important, the missing flexibility to add notes to all inquiry items, and the missing suggestions of nurse advice. Regarding the latter, a participant said in the interview that they give the same advice "maybe 20 times per duty, especially for ill children". Even standard advice is sometimes elaborate, and the participants said they spend a considerable amount of time documenting the advice they give.

5 DISCUSSION

5.1 Designing for Control in Nurse-AI Collaboration

Emergency calls are inclusive and welcoming in their form but can also be complex and unpredictable for the telenurses. They talk to patients that may be vulnerable. As such, the conversation is not only about screening and identifying symptoms but also about understanding, compassion, and rapport. To illustrate, the telenurses often say: "We will help you."

In Norway, the telenurses in the local emergency medical communication centers have the autonomy to exert clinical assessment. They have a flexible and practical approach to triage and do not always follow the current decision-support systems' guidance [34, 35]. As pointed out in the introduction, one strategy for introducing AI into telephone triage is using symptom checkers for system-led patient conversations. This strategy challenges the autonomy of healthcare professionals and can lead to a lack of support when implementing such tools in the workflow [40]. Designing for a fruitful integration of AI tools and human decision-making can increase healthcare professionals' trust in the tools [14].

Automation of central work tasks can challenge the nurses' clinical competence and their ability to perform clinical assessments [30, 56, 67]. When designing automation tools for decision-making, it is essential to consider and explicitly address how to retain "meaningful human agency" in the work processes [72]. This would be particularly valid in complex clinical practices. Taking an HCAI approach [75], this study has focused on how we can design an AI tool that preserves the autonomy of telenurses, aiming to enhance their work rather than completely automate clinical assessments and patient interactions.

Clinical assessment of patient calls is complex, and to design AI-enhanced decision-support, the algorithmic outputs must have a meaningful entryway into decision-making [14]. The AI support needs to be an integrated experience [77]. Shaw et al. [67] state that: "From a clinical perspective, algorithms that perform isolated risk prediction may be less useful. Clinical decision-making is a complex process involving the integration of a variety of data sources. To inform this decision-making process more intuitively, attention is increasingly being devoted to communication tools".

In line with how Xu et al. [75] point to the importance of taking complementary advantages of machine intelligence and human intelligence, we have designed a tool to be used during the entire length of a decision process. The HCAI approach includes three aspects of design work: technology, human factors, and ethics, promoting the idea that we should keep humans in a central position when developing AI systems [74]. Our study addresses how we can devise a participatory design process when designing AI systems, including how we can expand and include a sociotechnical understanding of the practice. Further, our study emphasizes that it is fundamental to explicitly focus on retaining the autonomy of the human actors in the design of AI systems.

5.2 Designing AI-Enhanced Decision Support

In our research, we have followed the call in the HCI community to examine meaningful human-AI interaction [62, 75] by encouraging speculation about future collaboration between telenurses and an AI tool. We have identified an opportunity to incorporate telenurses' dual duties of using a CDSS and documenting the call. A prototype was designed and was different from symptom checkers and other CDSSs in several ways. Telenurses received suggestions about inquiry items without being forced to use them or follow predefined pathways. They could always search for another inquiry item or use free-form text to document. Simple interactions lead to automatic documentation forming during the entire conversation to structure the assessment work.

Richard et al. [61] point to the risk that CDSSs can infringe upon the expertise and responsibility of the decision-maker. A problem with operator-facing symptom checkers or CDSSs is that they often have rigid pathways for the different conditions the operators encounter. When operators have to follow such pathways, there is a risk of interactional misalignment, where operator and caller are talking at cross purposes [51]. The operator's ability to deviate from the scripts can be reduced, and it may be hard for the patient to break through with important information [21, 51, 55]. In primary care, communication failures, as well as diagnosis and assessment problems, have been linked to incidents of patient harm [58]. Algorithmic questioning can be inappropriate for patients who have speech difficulties or have limited capacity to express themselves verbally, and professional telenurses may accommodate by asking fewer and more apt questions [20]. Schematic questioning may also be less accessible to children or others who are more capable of expressing themselves in the *voice of life world* than in the *voice of medicine* [50]. A mode of conversation where callers more readily can influence the course of conversation would make the service more accessible.

The telenurses in our study were concerned with an AI tool giving *too much direction*, making them focus on less serious symptoms, perhaps failing to unveil the true urgency of the patient's condition. When introducing an AI tool into the triage process, a concern is how to balance, on the one hand, the risk of steering the call down the wrong clinical pathway and, on the other, giving confident advice on what medical assessments and measures can be made based on the information provided during the call. To avoid steering the conversation down the wrong pathway, an AI tool could present alternative pathway turns to support the operator in exploring the matter holistically. Kudina and de Boer [39] claim that even if a CDSS suggests another pathway than the healthcare professional's initial choice, the decision process will benefit since the patient's condition will be explored more thoroughly with the combined approaches of human and machine intelligence. If the telenurse misinterprets the patient's situation, an AI tool's suggestions may help turn the conversation in the right direction again. Experimental evidence supports the effectiveness of human-AI collaboration in medical decision-making [59], and the experiment suggests that collaboration is most effective when the AI output's confidence is intuitively readable to the user. In telephone triage, where the direction of the call has to be decided continuously, an AI tool could indicate its confidence in another direction by persistently suggesting inquiry items that point in that direction.

The interview participants said that the available symptom combinations in the current CDSSs did not always fit, and they struggled to find a symptom combination that was specific enough. When the patient's condition is complex, one must rely more on telenurses' clinical assessment. In the prototype, we made a design choice that telenurses could get inquiry-item suggestions from the entire terminology instead of just from predetermined pathways. This way, all patient information could be documented, even that which did not align with conventional pathways. The participants expected an AI tool to identify critical symptom combinations that would call for urgent responses. The prototype would mark not only one but all critical combinations that were present or nearly present. Research suggests clinicians are more likely to embrace AI-enhanced decision-support that contributes to the process yet remains unobtrusive [77]. The marks were small squares that telenurses could choose to click. They used the functionality actively to see which inquiry items would be needed to satisfy a critical pattern and how the different items in play were related. They also explored their agreement with the tool when critical symptom combinations were marked. Enabling exploration and hypothesis-testing for healthcare professionals have been found valuable. [14]

Our study demonstrates consideration of how an AI system can provide support in clinical decisions without forcing a direction. In line with this, we explore how we can enhance the clinical competence of telenurses and support their nuanced clinical assessments. Such design considerations can also be relevant in other high-risk domains that rely on human expertise.

5.3 Standardization and Flexibility in Documentation Support

From the co-design workshop, it was clear that the nurses simplified patient accounts in their documentation. This indicated that the documentation would be suitable for more standardization. Furthermore, the nurses noted that the lack of current standardization posed a problem, especially since the patients called in with a wide range of problems. Therefore, we made an extensive terminology to facilitate the documentation of complex and varied patient problems in a standardized way. Machine learning models can reduce the number of documentation items necessary to predict urgency [82]. Nonetheless, our understanding from the telenurses is that more detailed documentation is valuable for them and the patients who may claim access to it.

We learned in interviews and the co-design workshop that the participants wanted flexibility. For the nurses not to be limited by the terminology, the prototype allowed them to supplement selected inquiry items with free-form text. However, the participants in the role-play workshop wanted to be able to add free-form supplementary notes to many more symptoms. They wanted this function in case they did not find a fitting inquiry item or for patients with psychiatric problems. They said describing this vulnerable group's health problem is especially important. Flexibility is vital for how humans and AI tools may collaborate and learn from each other in the long run. Bowker and Star [9] point out that if healthcare professionals must adhere to strict categories, these can become self-reliant and constrain how patient information is documented. Categories that do not quite fit, may be reinforced in machine learning. With flexibility for humans, however, machine learning can identify new, more fitting categories.

It is advisable to build flexibility into improvement interventions in healthcare [45]. The participants in the role-play workshop had not been using the prototype before. We had built in some flexibility to document with free-form text. Building in even more flexibility would have made it voluntary for novices to use the automatic suggestions. This could lower the threshold for using such systems, and users could take their time to familiarize themselves with the standardized inquiry items. The prototype was designed to help telenurses to remember questions to ask the callers. After a call, participants sometimes noticed questions they wished had been asked. A system that suggests often overlooked inquiry items at the right time might help avoid incomplete understanding and documentation of the patient's situation. This would especially be useful for novice nurses.

Our human-AI collaboration study illustrates consideration of how standardization and flexibility can be balanced to enable effective documentation of dynamic real-world problems.

6 CONCLUSION

Front-line healthcare services, with their close relation to their patients, are particularly suitable for AI initiatives [42]. We believe that designing AI tools together with healthcare professionals is a path to resilient AI tools that can align with healthcare professionals' and patients' needs in the long run. This will lay the grounds for data resilience where the AI tools can adapt to changing clinical practices and epidemiological transition [9, 42].

This study addresses how we can design for control in human-AI collaboration. The focus is on designing AI support that integrates with the workflow to enhance, rather than replace, the capabilities of healthcare professionals. When designing AI-enhanced clinical decision support systems, designers should balance the need to steer the process in a particular direction with encouraging the exploration of alternative pathways. Documentation support can be standardized to the extent that healthcare professionals cannot nuance the documentation of clinical assessments satisfactorily [9], and a key design challenge is for automatic documentation processes to balance standardization and flexibility.

This research answers a call in the HCI community to design HCAI by addressing how we can design for meaningful human-AI collaboration in the case of telephone triage in local emergency medical communication centers. The study has used in-depth interviews with telenurses to provide a rich picture of their work practice and workflows, how their practice is currently supported by a number of tools and interfaces, and given a portrayal of how these tools figure in their work. Further, the interviews contained a speculative element, inviting telenurses to discuss how AI could automate and support parts of their workflow. These insights were followed up in a co-design workshop where professional telenurses were invited to co-create interface elements and understand how these elements and features could be used in a future workflow. The co-design workshop was followed by the design of an AI-driven working prototype, which was later employed in a role-play workshop. Based on these empirical inquiries and design activities, we have discussed how the challenge of dual duties of using decision support in clinical assessments and creating documentation can be addressed. The study's main aim has been identifying opportunities and challenges for integrating human-AI collaborations into the telephone triage workflow with an explicit focus on retaining the telenurses' control over the process and supporting rather than replacing their clinical assessments.

ACKNOWLEDGMENTS

We want to give our heartfelt thanks to leaders and telenurses from the participating call centers Arendal, Bergen, Bjørnafjorden og Samnanger, Engerdal og Trysil, Sandnes, Sotra, and Trondheim. Their commitment and wholehearted engagement have been inspiring and vital. A big thank you goes to Bjørnar Jensen for his foundational work on the prototype and all the enjoyable exchanges. We want to express our gratitude to Helsetjenestens driftsorganisasjon for nødnett HF (HDO) and its dedicated team, including Jannike Andresen and Emil Perry, for opening doors for us in Phase 1 by allowing us to ask their nurse contacts about AI and for sharing valuable insights from HDO's own requirement-elicitation project. Our sincere appreciation goes to Professor Saba Mylvaganam for his insightful input during the final stage of the paper. This research is part of the RE-AIMED project, funded by the Research Council of Norway (grant number 310468).

REFERENCES

[1] Stephanie Aboueid, Samantha Meyer, James R Wallace, Shreya Mahajan, and Ashok Chaurasia. 2021. Young Adults' Perspectives on the Use of Symptom Checkers for Self-Triage and Self-Diagnosis: Qualitative Study. *JMIR Public Health and Surveillance* 7, 1 (Jan 2021), e22637. https://doi.org/10.2196/22637
[2] Jamil Ahmad, Muhammad Sajjad, Seungmin Rho, Soon-il Kwon, Mi Young Lee, and Sung Wook Baik. 2018. Determining speaker attributes from stress-affected speech in emergency situations with hybrid SVM-DNN architecture. *Multimedia Tools and Applications* 77, 4 (Feb 2018), 4883–4907. https://doi.org/10.1007/s11042-016-4041-7
[3] Tayla Anthony, Amit Kumar Mishra, Willem Stassen, and Jarryd Son. 2021. The Feasibility of Using Machine Learning to Classify Calls to South African Emergency Dispatch Centres According to Prehospital Diagnosis, by Utilising Caller Descriptions of the Incident. *Healthcare (Basel, Switzerland)* 9, 9 (Aug 2021), 1107. https://doi.org/10.3390/healthcare9091107
[4] James Auger. 2013. Speculative design: crafting the speculation. *Digital Creativity* 24, 1 (Mar 2013), 11–35. https://doi.org/10.1080/14626268.2013.767276
[5] Abdul Malik Badshah, Nasir Rahim, Noor Ullah, Jamil Ahmad, Khan Muhammad, Mi Young Lee, Soonil Kwon, and Sung Wook Baik. 2019. Deep features-based speech emotion recognition for smart affective services. *Multimedia Tools and Applications* 78, 5 (Mar 2019), 5571–5589. https://doi.org/10.1007/s11042-017-5292-7
[6] Laura Barendregt and Nora Sørensen Vaage. 2021. Speculative design as thought experiment. *She Ji: The Journal of Design, Economics, and Innovation* 7, 3 (2021), 374–402. https://doi.org/10.1016/j.sheji.2021.06.001
[7] Clara Berridge, Yuanjin Zhou, Amanda Lazar, Anupreet Porwal, Nora Mattek, Sarah Gothard, and Jeffrey Kaye. 2022. Control Matters in Elder Care Technology: Evidence and Direction for Designing It In. In *Designing Interactive Systems Conference (DIS '22)*. Association for Computing Machinery, New York, NY, USA, 1831–1848. https://doi.org/10.1145/3532106.3533471
[8] Stig Nikolaj Blomberg, Helle Collatz Christensen, Freddy Lippert, Annette Kjær Ersbøll, Christian Torp-Petersen, Michael R. Sayre, Peter J. Kudenchuk, and Fredrik Folke. 2021. Effect of Machine Learning on Dispatcher Recognition of Out-of-Hospital Cardiac Arrest During Calls to Emergency Medical Services: A Randomized Clinical Trial. *JAMA Network Open* 4, 1 (Jan 2021), e2032320. https://doi.org/10.1001/jamanetworkopen.2020.32320
[9] Geoffrey C. Bowker and Susan Leigh Star. 2000. *Sorting Things Out: Classification and Its Consequences*. The MIT Press, Cambridge, Massachusetts. https://doi.org/10.7551/mitpress/6352.001.0001
[10] Tone Bratteteig and Guri Verne. 2018. Does AI make PD obsolete? Exploring challenges from artificial intelligence to participatory design. In *Proceedings of the 15th Participatory Design Conference: Short Papers, Situated Actions, Workshops and Tutorial - Volume 2 (PDC '18)*. Association for Computing Machinery, New York, NY, USA, 1–5. https://doi.org/10.1145/3210604.3210646
[11] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (Jan 2006), 77–101. https://doi.org/10.1191/1478088706qp063oa
[12] Mary B. Burns and Kevin C. Moffitt. 2014. Automated deception detection of 911 call transcripts. *Security Informatics* 3, 1 (Aug 2014), 8. https://doi.org/10.1186/s13388-014-0008-2
[13] Fredrik Byrsell, Andreas Claesson, Mattias Ringh, Leif Svensson, Martin Jonsson, Per Nordberg, Sune Forsberg, Jacob Hollenberg, and Anette Nord. 2021. Machine learning can support dispatchers to better and faster recognize out-of-hospital cardiac arrest during emergency calls: A retrospective study. *Resuscitation* 162, March 6 (2021), 218–226. https://doi.org/10.1016/j.resuscitation.2021.02.041
[14] Carrie J. Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S. Corrado, Martin C. Stumpe, and Michael Terry. 2019. Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300234

[15] Kuan-Chen Chin, Tzu-Chun Hsieh, Wen-Chu Chiang, Yu-Chun Chien, Jen-Tung Sun, Hao-Yang Lin, Ming-Ju Hsieh, Chi-Wei Yang, Albert Y. Chen, and Matthew Huei-Ming Ma. 2021. Early recognition of a caller's emotion in out-of-hospital cardiac arrest dispatching: An artificial intelligence approach. *Resuscitation* 167 (Oct 2021), 144–150. https://doi.org/10.1016/j.resuscitation.2021.08.032

[16] Enrico Coiera, Baki Kocaballi, John Halamka, and Liliana Laranjo. 2018. The digital scribe. *npj Digital Medicine* 1, 11 (Oct 2018), 1–5. https://doi.org/10.1038/s41746-018-0066-9

[17] Enrico Coiera and Sidong Liu. 2022. Evidence synthesis, digital scribes, and translational challenges for artificial intelligence in healthcare. *Cell Reports Medicine* 3, 12 (2022), 100860.

[18] Theo Deschamps-Berger, Lori Lamel, and Laurence Devillers. 2022. Investigating Transformer Encoders and Fusion Strategies for Speech Emotion Recognition in Emergency Call Center Conversations.. In *Companion Publication of the 2022 International Conference on Multimodal Interaction (ICMI '22 Companion)*. Association for Computing Machinery, New York, NY, USA, 144–153. https://doi.org/10.1145/3536220.3558038

[19] Jiatao Ding, Michael Freeman, and Sameer Hasija. 2022. Can Predictive Technology Help Improve Acute Care Operations? Investigating the Impact of Virtual Triage Adoption. *INSEAD Working Paper* 2022, 47 (Oct 2022), 56. https://doi.org/10.2139/ssrn.3806478

[20] Irene Eriksson, Kristina Ek, Sofie Jansson, Ulrika Sjöström, and Margaretha Larsson. 2019. To feel emotional concern: A qualitative interview study to explore telephone nurses' experiences of difficult calls. *Nursing Open* 6, 3 (2019), 842–848. https://doi.org/10.1002/nop2.264

[21] Annica Ernesäter, Inger Holmström, and Maria Engström. 2009. Telenurses' experiences of working with computerized decision support: supporting, inhibiting and quality improving. *Journal of Advanced Nursing* 65, 5 (2009), 1074–1083. https://doi.org/10.1111/j.1365-2648.2009.04966.x

[22] Marta Fernandes, Susana M. Vieira, Francisca Leite, Carlos Palos, Stan Finkelstein, and João M. C. Sousa. 2020. Clinical Decision Support Systems for Triage in the Emergency Department using Intelligent Systems: A Review. *Artificial Intelligence in Medicine* 102 (Jan 2020), 101762. https://doi.org/10.1016/j.artmed.2019.101762

[23] ALSG Centre for Training & Development. 2023. Manchester Triage System (MTS). https://www.triagenet.net/classroom

[24] Hamish Fraser, Enrico Coiera, and David Wong. 2018. Safety of patient-facing digital symptom checkers. *The Lancet* 392, 10161 (Nov 2018), 2263–2264. https://doi.org/10.1016/S0140-6736(18)32819-8

[25] Kevin Gormley, Katy Lockhart, and Jolly Isaac. 2022. Using natural language processing in facilitating pre-hospital telephone triage of emergency calls. *British Paramedic Journal* 7, 2 (Sep 2022), 31–37. https://doi.org/10.29045/14784726.2022.09.7.2.31

[26] Nathaniel R. Greenbaum, Yacine Jernite, Yoni Halpern, Shelley Calder, Larry A. Nathanson, David A. Sontag, and Steven Horng. 2019. Improving documentation of presenting problems in the emergency department using a domain-specific ontology and machine learning-driven user interfaces. *International Journal of Medical Informatics* 132 (Dec 2019), 103981. https://doi.org/10.1016/j.ijmedinf.2019.103981

[27] Mary Elizabeth Greenberg. 2009. A comprehensive model of the process of telephone nursing. *Journal of Advanced Nursing* 65, 12 (2009), 2621–2629. https://doi.org/10.1111/j.1365-2648.2009.05132.x

[28] Felix Gräßer, Stefanie Beckert, Denise Küster, Jochen Schmitt, Susanne Abraham, Hagen Malberg, and Sebastian Zaunseder. 2017. Therapy Decision Support Based on Recommender System Methods. *Journal of Healthcare Engineering* 2017 (Mar 2017), e8659460. https://doi.org/10.1155/2017/8659460

[29] Christina Harrington and Tawanna R Dillahunt. 2021. Eliciting Tech Futures Among Black Young Adults: A Case Study of Remote Speculative Co-Design. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3411764.3445723

[30] Jennifer Y Hong, Catherine H Ivory, Courtney B VanHouten, Christopher L Simpson, and Laurie Lovett Novak. 2021. Disappearing expertise in clinical automation: Barcode medication administration and nurse autonomy. *Journal of the American Medical Informatics Association* 28, 2 (Feb 2021), 232–238. https://doi.org/10.1093/jamia/ocaa135

[31] Alison Imbens-Bailey and Allyssa McCabe. 2000. The discourse of distress: a narrative analysis of emergency calls to 911. *Language & Communication* 20, 3 (Jul 2000), 275–296. https://doi.org/10.1016/S0271-5309(99)00025-7

[32] Infermedica. 2023. Médis - Case study - Infermedica. https://infermedica.com/case-studies/medis

[33] Infermedica.com. 2023. PZU Zdrowie - Case study - Infermedica. https://infermedica.com/case-studies/pzu-zdrowie

[34] Lars E. F. Johannessen. 2017. Beyond guidelines: discretionary practice in face-to-face triage nursing. *Sociology of Health & Illness* 39, 7 (2017), 1180–1194. https://doi.org/10.1111/1467-9566.12578

[35] Lars E. F. Johannessen. 2018. Workplace assimilation and professional jurisdiction: How nurses learn to blur the nursing-medical boundary. *Social Science & Medicine* 201 (Mar 2018), 51–58. https://doi.org/10.1016/j.socscimed.2018.02.004

[36] Abirami Kirubarajan, Ahmed Taher, Shawn Khan, and Sameer Masood. 2020. Artificial intelligence in emergency medicine: A scoping review. *Journal of the American College of Emergency Physicians Open* 1, 6 (2020), 1691–1702. https://doi.org/10.1002/emp2.12277

[37] A Baki Kocaballi, Kiran Ijaz, Liliana Laranjo, Juan C Quiroz, Dana Rezazadegan, Huong Ly Tong, Simon Willcock, Shlomo Berkovsky, and Enrico Coiera. 2020. Envisioning an artificial intelligence documentation assistant for future primary care consultations: A co-design study with general practitioners. *Journal of the American Medical Informatics Association* 27, 11 (Nov 2020), 1695–1704. https://doi.org/10.1093/jamia/ocaa131

[38] Nasjonalt kompetansesenter for legevaktmedisin. 2023. Legevaktindeks – Beslutningsstøtte for legevakthenvendelser. https://legevaktindeks.no/

[39] Olya Kudina and Bas de Boer. 2021. Co-designing diagnosis: Towards a responsible integration of Machine Learning decision-support systems in medical diagnostics. *Journal of Evaluation in Clinical Practice* 27, 3 (2021), 529–536. https://doi.org/10.1111/jep.13535

[40] Sari Kujala and Iiris Hörhammer. 2022. Health Care Professionals' Experiences of Web-Based Symptom Checkers for Triage: Cross-sectional Survey Study. *Journal of Medical Internet Research* 24, 5 (May 2022), e33505. https://doi.org/10.2196/33505 Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research publisher: JMIR Publications Inc., Toronto, Canada.

[41] Min Hun Lee, Daniel P. P. Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez Bermúdez i Badia. 2021. A Human-AI Collaborative Approach for Clinical Decision Making on Rehabilitation Assessment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–14. https://doi.org/10.1145/3411764.3445472

[42] Steven Lin. 2022. A Clinician's Guide to Artificial Intelligence (AI): Why and How Primary Care Should Lead the Health Care AI Revolution. *The Journal of the American Board of Family Medicine* 35, 1 (Jan 2022), 175–184. https://doi.org/10.3122/jabfm.2022.01.210226

[43] Lovdata. 2001. *Lov om helsepersonell m.v. (helsepersonelloven)*. https://lovdata.no/dokument/NL/lov/1999-07-02-64/#KAPITTEL_8

[44] Megh Marathe, Jacki O'Neill, Paromita Pain, and William Thies. 2016. ICT-enabled grievance redressal in Central India: A comparative analysis. In *Proceedings of the Eighth International Conference on Information and Communication Technologies and Development*. ACM, Ann Arbor, Michigan, USA, 1–11.

[45] Martin Marshall, Debra de Silva, Lesley Cruickshank, Jenny Shand, Li Wei, and James Anderson. 2017. What we know about designing an effective improvement intervention (but too often fail to put into practice). *BMJ Quality & Safety* 26, 7 (Jul 2017), 578–582. https://doi.org/10.1136/bmjqs-2016-006143

[46] Mark Matthews, Geri Gay, and Gavin Doherty. 2014. Taking part: role-play in the design of therapeutic systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. Association for Computing Machinery, New York, NY, USA, 643–652. https://doi.org/10.1145/2556288.2557103

[47] Margaret McCartney. 2017. Margaret McCartney: General practice can't just exclude sick people. *BMJ* 359 (2017), 1–2.

[48] Ben Medler and Brian Magerko. 2010. The implications of improvisational acting and role-playing on design methodologies. In *CHI '10: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Vol. 1. ACM, Atlanta, Georgia, USA, 483–492. https://doi.org/10.1145/1753326.1753398

[49] Vivian Midtbø, Jens Leirvåg, Steinar Hunskaar, and Tone Morken. 2017. *Kompetanse i legevakt og legevaktsentral – implementering av akuttmedisinforskriften*. Number 5 in NKLM-rapport. Bergen. 37 pages. https://norceresearch.brage.unit.no/norceresearch-xmlui/bitstream/handle/1956/17307/Nklm-rapport-5-2017.pdf

[50] Elliot George Mishler. 1984. *The discourse of medicine: Dialectics of medical interviews*. Ablex, New Jersey.

[51] James I. Morgan and Thomas Muskett. 2020. Interactional misalignment in the UK NHS 111 healthcare telephone triage service. *International Journal of Medical Informatics* 134 (Feb 2020), 104030. https://doi.org/10.1016/j.ijmedinf.2019.104030

[52] Tone Morken, Line Remme Solberg, and Merete Allertsen. 2018. *Legevaktorganisering i Norge*. Number 4 in NKLM-rapport. Bergen. 108 pages.

[53] Keith E. Morse, Nicolai P. Ostberg, Veena G. Jones, and Albert S. Chan. 2020. Use Characteristics and Triage Acuity of a Digital Symptom Checker in a Large Integrated Health System: Population-Based Descriptive Study. *Journal of Medical Internet Research* 22, 11 (Nov 2020), e20549. https://doi.org/10.2196/20549

[54] Susan Randles Moscato, Barbara Valanis, Christina M Gullion, Christine Tanner, Susan E Shapiro, and Shigeko Izumi. 2007. Predictors of Patient Satisfaction With Telephone Nursing Services. *Clinical Nursing Research* 16, 2 (May 2007), 119–137. https://doi.org/10.1177/1054773806298507

[55] Jamie Murdoch, Rebecca Barnes, Jillian Pooler, Valerie Lattimer, Emily Fletcher, and John L. Campbell. 2015. The impact of using computer decision-support software in primary care nurse-led telephone triage: Interactional dilemmas and conversational consequences. *Social Science and Medicine* 126 (2015), 36–47. https://doi.org/10.1016/j.socscimed.2014.12.013

[56] Linda Onnasch, Christopher D. Wickens, Huiyang Li, and Dietrich Manzey. 2014. Human Performance Consequences of Stages and Levels of Automation: An Integrated Meta-Analysis. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 56, 3 (May 2014), 476–488. https://doi.org/10.1177/0018720813501549

[57] Nils Peters, Howard Lei, and Gerald Friedland. 2012. Name that room: room identification using acoustic features in a recording. In *Proceedings of the 20th ACM international conference on Multimedia (MM '12)*. Association for Computing Machinery, New York, NY, USA, 841–844. https://doi.org/10.1145/2393347.2396326

[58] Philippa Rees, Adrian Edwards, Colin Powell, Peter Hibbert, Huw Williams, Meredith Makeham, Ben Carter, Donna Luff, Gareth Parry, Anthony Avery, Aziz Sheikh, Liam Donaldson, and Andrew Carson-Stevens. 2017. Patient Safety Incidents Involving Sick Children in Primary Care in England and Wales: A Mixed Methods Analysis. *PLOS Medicine* 14, 1 (Jan 2017), 1–23. https://doi.org/10.1371/journal.pmed.1002217

[59] Carlo Reverberi, Tommaso Rigon, Aldo Solari, Cesare Hassan, Paolo Cherubini, and Andrea Cherubini. 2022. Experimental evidence of effective human–AI collaboration in medical decision-making. *Scientific reports* 12, 1 (2022), 14952.

[60] Dana Rezazadegan, Shlomo Berkovsky, Juan C. Quiroz, A. Baki Kocaballi, Ying Wang, Liliana Laranjo, and Enrico Coiera. 2020. Automatic speech summarisation: A scoping review. 2008.11897 (2020).

[61] Antoine Richard, Brice Mayag, François Talbot, Alexis Tsoukias, and Yves Meinard. 2020. What does it mean to provide decision support to a responsible and competent expert? *EURO Journal on Decision Processes* 8, 3 (Nov 2020), 205–236. https://doi.org/10.1007/s40070-020-00116-7

[62] Yvonne Rogers. 2022. Commentary: human-centred AI: the new zeitgeist. *Human–Computer Interaction* 37, 3 (May 2022), 254–255. https://doi.org/10.1080/07370024.2021.1976643

[63] S. Trent Rosenbloom. 2012. *Interface Terminologies*. Springer Science & Business Media, London, 95–106.

[64] S. T. Rosenbloom, R. A. Miller, K. B. Johnson, P. L. Elkin, and S. H. Brown. 2006. Interface Terminologies: Facilitating Direct Entry of Clinical Data into Electronic Health Record Systems. *Journal of the American Medical Informatics Association* 13, 3 (May 2006), 277–288. https://doi.org/10.1197/jamia.M1957

[65] Salim Salmi, Saskia Mérelle, Renske Gilissen, and Willem-Paul Brinkman. 2021. Content-Based Recommender Support System for Counselors in a Suicide Prevention Chat Helpline: Design and Evaluation Study. *Journal of Medical Internet Research* 23, 1 (Jan 2021), e21690. https://doi.org/10.2196/21690 Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research publisher: JMIR Publications Inc., Toronto, Canada.

[66] Megan E. Salwei and Pascale Carayon. 2022. A Sociotechnical Systems Framework for the Application of Artificial Intelligence in Health Care Delivery. *Journal of Cognitive Engineering and Decision Making* 16, 4 (Dec 2022), 194–206. https://doi.org/10.1177/15553434221097357

[67] James Shaw, Frank Rudzicz, Trevor Jamieson, and Avi Goldfarb. 2019. Artificial Intelligence and the Implementation Challenge. *Journal of Medical Internet Research* 21, 7 (Jul 2019), e13659. https://doi.org/10.2196/13659 Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research publisher: JMIR Publications Inc., Toronto, Canada.

[68] Guiyoung Son, Soonil Kwon, and Neungsoo Park. 2019. Gender Classification Based on the Non-Lexical Cues of Emergency Calls with Recurrent Neural Networks (RNN). *Symmetry-Basel* 11, 4 (Apr 2019), 525. https://doi.org/10.3390/sym11040525

[69] Rocío Sánchez-Salmerón, José L. Gómez-Urquiza, Luis Albendín-García, María Correa-Rodríguez, María Begoña Martos-Cabrera, Almudena Velando-Soriano, and Nora Suleiman-Martos. 2022. Machine learning methods applied to triage in emergency services: A systematic review. *International Emergency Nursing* 60 (Jan 2022), 101109. https://doi.org/10.1016/j.ienj.2021.101109

[70] Lauri Tavi. 2019. Classifying females' stressed and neutral voices using acoustic–phonetic analysis of vowels: an exploratory investigation with emergency calls. *International Journal of Speech Technology* 22, 3 (Sep 2019), 511–520. https://doi.org/10.1007/s10772-018-09574-6

[71] Liam Tollinton, Alexander M. Metcalf, and Sumithra Velupillai. 2020. Enhancing predictions of patient conveyance using emergency call handler free text notes for unconscious and fainting incidents reported to the London Ambulance Service. *International Journal of Medical Informatics* 141 (Sep 2020), 104179. https://doi.org/10.1016/j.ijmedinf.2020.104179

[72] Ben Wagner. 2019. Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems. *Policy & Internet* 11, 1 (2019), 104–122. https://doi.org/10.1002/poi3.198

[73] Wikipedia. 2020. Norsk indeks for medisinsk nødhjelp. https://no.wikipedia.org/w/index.php?title=Norsk_indeks_for_medisinsk_n%C3%B8dhjelp&oldid=20689709 Page Version ID: 20689709.

[74] Wei Xu. 2019. Toward human-centered AI: a perspective from human-computer interaction. *Interactions* 26, 4 (Jun 2019), 42–46. https://doi.org/10.1145/3328485

[75] Wei Xu, Marvin J. Dainoff, Liezhong Ge, and Zaifeng Gao. 2022. Transitioning to Human Interaction with AI Systems: New Challenges and Opportunities for HCI Professionals to Enable Human-Centered AI. *International Journal of Human–Computer Interaction* 0, 0 (Apr 2022), 1–25. https://doi.org/10.1080/10447318.2022.2041900

[76] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. *Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376301

[77] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable AI: Fitting Intelligent Decision Support into Critical, Clinical Decision-Making Processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300468

[78] Tohta Yasuda, Yuki Yamada, Fumiya Hamatsu, and Tomoki Hamagami. 2017. A call triage support system for emergency medical service using multiple random forests. *IEEJ Transactions on Electrical and Electronic Engineering* 12 (2017), S67–S73. https://doi.org/10.1002/tee.22567

[79] Shin-ae Yoon, Guiyoung Son, and Soonil Kwon. 2019. Fear emotion classification in speech by acoustic and behavioral cues. *Multimedia Tools and Applications* 78, 2 (Jan 2019), 2345–2366. https://doi.org/10.1007/s11042-018-6329-2

[80] Yue You and Xinning Gui. 2021. Self-Diagnosis through AI-enabled Chatbot-based Symptom Checkers: User Experiences and Design Considerations. *AMIA Annual Symposium Proceedings* 2020 (Jan 2021), 1354–1363.

[81] Stephanie Wing Yin Yu, Andre Ma, Vivian Hiu Man Tsang, Lulu Suet Wing Chung, Siu-Chung Leung, and Ling-Pong Leung. 2020. Triage accuracy of online symptom checkers for Accident and Emergency Department patients. *Hong Kong Journal of Emergency Medicine* 27, 4 (Jul 2020), 217–222. https://doi.org/10.1177/1024907919842486

[82] Shota Yunoki, Tomoki Hamagami, Kenji Oshige, Chihiro Kawakami, and Noriyuki Suzuki. 2014. High Accuracy of Call Triage Decision by Bayesian Network. *Electronics and Communications in Japan* 97, 1 (Jan 2014), 62–69. https://doi.org/10.1002/ecj.11439