Review

Research Trends in Artificial Intelligence Applications in Human Factors Health Care: Mapping Review

Onur Asan, PhD; Avishek Choudhury, MSc
School of Systems and Enterprises, Stevens Institute of Technology, Hoboken, NJ, United States

Corresponding Author:
Onur Asan, PhD
School of Systems and Enterprises
Stevens Institute of Technology
1 Castle Point Terrace
Hoboken, NJ, 07030
United States
Phone: 1 4145264330
Email: oasan@stevens.edu

Abstract

Background: Despite advancements in artificial intelligence (AI) to develop prediction and classification models, little research has been devoted to real-world translations with a user-centered design approach. AI development studies in the health care context have often ignored two critical factors of ecological validity and human cognition, creating challenges at the interface with clinicians and the clinical environment.

Objective: The aim of this literature review was to investigate the contributions made by major human factors communities in health care AI applications. This review also discusses emerging research gaps, and provides future research directions to facilitate a safer and user-centered integration of AI into the clinical workflow.

Methods: We performed an extensive mapping review to capture all relevant articles published within the last 10 years in the major human factors journals and conference proceedings listed in the “Human Factors and Ergonomics” category of the Scopus Master List. In each published volume, we searched for studies reporting qualitative or quantitative findings in the context of AI in health care. Studies are discussed based on the key principles such as evaluating workload, usability, trust in technology, perception, and user-centered design.

Results: Forty-eight articles were included in the final review. Most of the studies emphasized user perception, the usability of AI-based devices or technologies, cognitive workload, and user’s trust in AI. The review revealed a nascent but growing body of literature focusing on augmenting health care AI; however, little effort has been made to ensure ecological validity with user-centered design approaches. Moreover, few studies (n=5 against clinical/baseline standards, n=5 against clinicians) compared their AI models against a standard measure.

Conclusions: Human factors researchers should actively be part of efforts in AI design and implementation, as well as dynamic assessments of AI systems’ effects on interaction, workflow, and patient outcomes. An AI system is part of a greater sociotechnical system. Investigators with human factors and ergonomics expertise are essential when defining the dynamic interaction of AI within each element, process, and result of the work system.

(JMIR Hum Factors 2021;8(2):e28236) doi: 10.2196/28236

KEYWORDS
artificial intelligence; human factors; health care systems; ecological validity; usability; trust; perception; workload

Introduction

Influx of Artificial Intelligence in Health Care

The influx of artificial intelligence (AI) has been shifting paradigms for the last decade. The term “AI” has been often used and interpreted with different meanings [1], and there is a lack of consensus regarding AI’s definition [2]. In general, AI can be defined as a computer program or intelligent system capable of mimicking human cognitive function [3]. Over the years, the capabilities and scope of AI have substantially increased. AI now ranges from algorithms that operate with predefined rules and those that rely on if-then statements (decision tree classifiers) [4] to more sophisticated deep-learning...
algorithms that have the capabilities to automatically learn and improve through statistical analyses of large datasets [5,6]. There have been many studies and advancements with AI as it continues to evolve in numerous domains, including health care. AI applications such as MelaFind, a virtual assistant software, and IBM Watson have been introduced to improve health care systems, foster patient care, and augment patient safety [7]. AI applications have been developed and studied for every stakeholder in health care, including providers, administrators, patients, families, and insurers. In some specific areas such as radiology and pathology, there are strong arguments that AI systems may supersede doctors as a result of studies showing that AI algorithms outperformed doctors in accurately detecting cancer cells [8-10].

Further, developments in AI-enabled health information technologies (eg, AI-enabled electronic health records [EHRs] or clinical decision support systems) have benefitted from the availability of big data to predict clinical outcomes and assist providers in parsing through their EHRs to find individual pieces of medical information [11]. Despite AI having great potential, it is still in its infancy. The existing clinical AI systems are far from perfect for several well-known reasons, including (a) discriminatory biases coming from the input data; (b) lack of transparency in AI decisions, particularly neural networks, due to the black-box nature; and (c) sensitivity of the resulting decisions to the input data [6,12].

**Typical AI-User Interactions**

AI systems are complex in the sense of being a black box for the users who might not have adequate expertise in statistics or computer science to be able to comprehend the functioning of AI. Thus, AI can undesirably complicate the relationships between users and computer systems if not well designed. Unlike other health care technologies, AI can interact (eg, through chatbots, automated recommender systems, health apps) with clinicians and patients based on the inputs (feedback) that it receives from the user, thus creating what we refer to as “the interaction loop.” Unlike non-AI technologies, AI’s output (result generated by the AI) largely depends on the information fed into it; for instance, in AI-based reinforcement learning [13], the system may learn and adapt itself based on user input. Therefore, the human-AI interaction may influence the human as well as the AI system: the user feeds AI with some information; the AI learns from this information, performs analyses, and sends an output to the user; the user receives the output, comprehends it, and acts accordingly; and the new data generated by the user’s action goes back to the AI. Figure 1 illustrates three fundamental and typical interaction loops highlighting fundamental plausible transactions among clinicians, patients, and the AI system, in which the AI technology (such as Apple Watch) continuously measures the user’s health information (heart rate, oxygen level) and sends the data to the user’s health care provider. The care provider can then make treatment plans or clinical recommendations based on the AI results, which will then influence user health or health-related behavior (Loop 1). Other common user-AI interactions can be observed in online health services in which the user interacts with an AI-enabled chatbot for preliminary diagnoses (Loop 2). The third, but less common, user-AI interaction is when a doctor and patient together leverage an AI system for obtaining a better diagnosis in a clinical environment (Loop 3). For all of these applications, it is essential for the users to make a correct interpretation of AI outcomes, and to have a basic understanding of AI requirements and limitations. The optimum and successful user-AI interaction depends on several factors, including the physical (eg, timely access to technology, and visual and hearing ability, particularly of patients), cognitive (eg, ability to comprehend AI functioning, ability to reason and use AI-enabled devices), and emotional (eg, current state of mind, willingness to use AI, prior experience with AI technology) resources of people (eg, health professionals and caregivers).
Efforts to Improve AI and the Essential Role of Human Factors

The developers of health care AI apps have primarily focused on AI’s analytical capabilities, accuracy, speed, and data handling (see Figure 2) and have neglected human factors perspectives, which lead to poorly designed apps [14]. Although recent studies have reported the impact of biased data [15], as well as interpretability, interoperability, and lack of standardization [7,16] on AI outcomes, very few have acknowledged the need to assess the interactions among AI, clinicians, and care recipients.

Recently, as acknowledged in the Annual Meetings of the Human Factors Ergonomics Society [17,18], increasing autonomous activities in health care can pose risks and concerns regarding AI. Therefore, there is a need to integrate human factors and ergonomics (HFE) principles and methods into developing AI-enabled technologies for better use, workflow integration, and interaction. In health care AI research, two factors have not been sufficiently addressed by researchers, namely ecological validity and human cognition, which may create challenges at the interface with clinicians as well as the clinical environment and lead to errors. Moreover, there is insufficient research focusing on improving the human factors, mainly (a) how to ensure whether clinicians are implementing the AI correctly, (b) the cognitive workload it imposes on clinicians working in stressful environments, and (c) its impact on clinical decision-making and patient outcome. The inconvenient truth is that most of the AI showing prominent ability in research and the literature is not currently executable in a clinical environment [19,20]. Therefore, to better identify the current state of HFE involvement in health care AI, we performed a mapping review of studies published in major human factors journals and proceedings related to AI systems in health care. The aim of the mapping review was to highlight what has been accomplished and reported in HFE journals and discuss the roles of HFE in health care AI research in the near future, which can facilitate smoother human-system interactions.
Methods

Design and Data Source

We performed a mapping review to explore the trending and initial areas regarding health care AI research in HFE publications. Our protocol was registered with the Open Science Framework on October 2, 2020 [21]. Mapping reviews are well-developed approaches to cover the representative literature (not exhaustive) for exploring and demonstrating trends in a given topic and time duration [22]. In this study, we selected major human factors journals and conferences that potentially publish health care–related work as our data source. Our selection of journals and conferences was guided by the “Human Factors and Ergonomics” category of the Scopus Master List and Scimago Journal & Country Rank. We also added two journals that potentially publish patient safety-related human factors work: *Journal of Patient Safety* and *BMJ Quality and Safety*. In total, we explored 24 journals and 9 conference proceedings (see Multimedia Appendix 1). All the authors approved the final list of journals and conferences with consensus.

Inclusion and Exclusion Criteria

We performed an extensive manual search to capture all relevant articles published in English within the last 10 years (January 2010 to December 2020) in the journals and conference proceedings listed in Multimedia Appendix 1. In each published volume, we searched for studies reporting qualitative or quantitative findings in the context of AI in health care. The selected studies needed to (1) be framed in the context of health care; (2) cover an AI algorithm or AI-enabled technology such as machine learning, natural language processing, or robotics; and (3) report either qualitative or quantitative findings/outcomes. We only included journal papers and full conference proceeding papers. Other materials such as conference abstracts, editorials, book chapters, poster presentations, perspectives, study protocols, review papers, and gray literature (eg, government reports and policy statement papers) were excluded.

Paper Selection and Screening

Articles in the journal and conference list were manually screened by two reviewers (AC and a research assistant) based on titles and abstracts using one of the inclusion criteria (ie, to be framed in the context of health care). We exported all of the retrieved publications to Sysrev software. In the second step, we excluded all ineligible publications (eg, reviews, short abstracts, and posters), as explained in the preceding section. In the last step, two reviewers (AC and a research assistant) independently screened all of the selected full papers based on the remaining two inclusion criteria: (1) covering an AI algorithm or AI-enabled technology such as machine learning, natural language processing, or robotics; and (2) reporting either qualitative or quantitative findings/outcomes. The reviewers also confirmed that the studies were framed in a health care context. The reviewers achieved 82% agreement. The lead researcher (OA) then resolved all conflicts, screened all of the shortlisted full-text articles, and finalized the article selection.

Data Extraction and Analysis

We followed a similar data extraction approach and analysis as reported by Holden et al [23]. Metadata (author names, the title of the paper, abstract) for each of the included articles were recorded in a standard Excel sheet. In our analysis, both authors (AC and OA) coded each included paper on different dimensions such as (1) sample/participant type, (2) AI system used, (3) source of data collection, and (4) objective and outcomes. Studies were also discussed based on the HFE principles such as evaluating workload, usability, trust in technology,
perception, and user-centered design. These HFE principles and subcategories for the dimensions were derived from the final selected papers and were checked for face validity by the researchers. We iteratively worked on the data extraction process and revised the categories to achieve a final consensus.

**Results**

**Summary of Included Studies**

Figure 3 illustrates the screening and selection process. As a result of screening 24 selected journals and 9 conference proceedings (Multimedia Appendix 1), we finalized 48 articles matching our inclusion criteria, which were included in the scoping review with consensus from all reviewers. These 48 articles were published in 10 journals and 3 conference proceedings, as illustrated in Figure 4.

Figure 3. Selection and exclusion process. AI: artificial intelligence.
Table 1 shows the following dimensions: (1) objective of the study; (2) overall methods used, including the ethnographic/quantitative analysis methods adopted, and the type of data (“Methods and Data” column); (3) study participants (user of the AI system); and (4) primary outcome/findings of the study. Most studies involved human participants such as clinicians and patients (n=33) as shown in the “Study Participants” column in Table 1. However, some studies used data from online sources such as Reddit, Twitter, and clinical databases. Approximately 26 studies conducted surveys and interviews to gain insight from study participants, as shown in the “Methods and Data” column. Some studies emphasized algorithms to analyze video, text, and sensor data. Overall, we observed that most studies evaluated AI from the user perspective and others leveraged AI to augment user performance.
Table 1. Evidentiary table of selected publications, summarizing their objectives, methods, participants, and outcomes (N=48).

<table>
<thead>
<tr>
<th>Study</th>
<th>Objective</th>
<th>Methods and Data</th>
<th>Study participants</th>
<th>Immediate outcome observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aldape-Pérez et al [24]</td>
<td>To promote collaborative learning among less experienced physicians</td>
<td>Mathematical/numerical data</td>
<td>NA&lt;sup&gt;a&lt;/sup&gt; (online database)</td>
<td>Delta Associative Memory was effective in pattern recognition in the medical field and helped physicians learn</td>
</tr>
<tr>
<td>Azari et al [25]</td>
<td>To predict surgical maneuvers from a continuous video record of surgical benchtop simulations</td>
<td>Mathematical/video data</td>
<td>37 surgeons</td>
<td>Machine learning’s prediction of surgical maneuvers was comparable to the prediction of robotic platforms</td>
</tr>
<tr>
<td>Balani and De Choudhury [26]</td>
<td>To detect levels of self-disclosure manifested in posts shared on different mental health forums on Reddit</td>
<td>Mathematical/text data</td>
<td>NA (Reddit posts from 7248 users)</td>
<td>Mental health subreddits can allow individuals to express or engage in greater self-disclosure</td>
</tr>
<tr>
<td>Cai et al [27]</td>
<td>To identify the needs of pathologists when searching for similar images, retrieved using a deep-learning algorithm</td>
<td>Survey study: Mayer’s trust model, NASA-TLX, questions for mental support for decision-making, diagnostic utility, workload, future use, and preference</td>
<td>12 pathologists</td>
<td>Users indicated having greater trust in SMILY; it offered better mental support, and providers were more likely to use it in clinical practice</td>
</tr>
<tr>
<td>Cvetković and Cvetković [28]</td>
<td>To analyze the influence of age, occupation, education, marital status, and economic condition on depression in breast cancer patients</td>
<td>Interview study using the Beck Depression Inventory guide</td>
<td>84 patients</td>
<td>Patient age and occupation had the most substantial influence on depression in breast cancer patients</td>
</tr>
<tr>
<td>Ding et al [29]</td>
<td>To learn about one’s health in everyday settings with the help of face-reading technology</td>
<td>Interview study: specific questions about time and location of usage, users’ perceptions and interpretations of the results, and intentions to use it in the future</td>
<td>10 users</td>
<td>Technology acceptance was hindered due to low technical literacy, low trust, lack of adaptability, infeasible advice, and usability issues</td>
</tr>
<tr>
<td>Erebak and Turgut [30]</td>
<td>To study human-robot interaction in elder care facilities</td>
<td>Survey study: Godspeed anthropomorphism scale, trust checklist [31], scales from [32], and automated functions of [33].</td>
<td>102 caregivers</td>
<td>No influence of anthropomorphism was detected on trust in robots; providers who trusted robots had more intention to work with them and preferred a higher automation level</td>
</tr>
<tr>
<td>Gao et al [34]</td>
<td>To detect motor impairment in Parkinson disease via implicitly sensing and analyzing users’ everyday interactions with their smartphones</td>
<td>Mathematical; sensor data</td>
<td>42 users</td>
<td>Parkinson disease was detected with significantly higher accuracy when compared to a clinical reference</td>
</tr>
<tr>
<td>Hawkins et al [35]</td>
<td>To measure the patient-perceived quality of care in US hospitals</td>
<td>Survey study: hospitals were asked to provide feedback regarding their use of Twitter for patient relations</td>
<td>NA (Tweets)</td>
<td>Patients use Twitter to provide input on the quality of hospital care they receive; almost half of the sentiment toward hospitals was, on average, favorable</td>
</tr>
<tr>
<td>Hu et al [36]</td>
<td>To detect lower back pain from body balance and sway performance</td>
<td>Mathematical; sensor data</td>
<td>44 patients and healthy participants</td>
<td>The machine-learning model was successful in identifying patients with back pain and responsible factors</td>
</tr>
<tr>
<td>Jin et al [37]</td>
<td>To identify, extract, and minimize medical error factors in the medication administration process</td>
<td>Mathematical/text data</td>
<td>NA (data from 4 hospitals)</td>
<td>The proposed machine-learning model identified 12 potential error factors</td>
</tr>
<tr>
<td>Kandaswamy et al [38]</td>
<td>To predict the accuracy of an order placed in the EHR by emergency medicine physicians</td>
<td>Mathematical/text and numerical data</td>
<td>53 clinicians</td>
<td>Machine-learning algorithms identified error rates in imaging, lab, and medication orders</td>
</tr>
<tr>
<td>Komogortsev and Holland [39]</td>
<td>To detect mild traumatic brain injury (mTBI) via the application of eye movement biometrics</td>
<td>Mathematical/video data</td>
<td>32 patients and healthy participants</td>
<td>Supervised and unsupervised machine learning classified participants with detection scores ≤ –0.870 and ≥ 0.79 as having mTBI, respectively</td>
</tr>
<tr>
<td>Krause et al [40]</td>
<td>To support the development of understandable predictive models</td>
<td>Mathematical/numerical data</td>
<td>5 data scientists</td>
<td>Interactive visual analytic systems helped data scientists to interpret predictive models clinically</td>
</tr>
<tr>
<td>Study</td>
<td>Objective</td>
<td>Methods and Data</td>
<td>Study participants</td>
<td>Immediate outcome observed</td>
</tr>
<tr>
<td>-------</td>
<td>-----------</td>
<td>------------------</td>
<td>--------------------</td>
<td>---------------------------</td>
</tr>
<tr>
<td>Ladstatter et al [41]</td>
<td>To measure the feasibility of artificial neural networks in analyzing nurses' burnout process</td>
<td>Survey study: Nursing Burnout Scale Short Form</td>
<td>465 nurses</td>
<td>The artificial neural network identified personality factors as the reason for burnout in Chinese nurses</td>
</tr>
<tr>
<td>Ladstatter et al [42]</td>
<td>To assess whether artificial neural networks offer better predictive accuracy in identifying nursing burnouts than traditional statistical techniques</td>
<td>Survey study: Nursing Burnout Scale Short Form</td>
<td>462 nurses</td>
<td>Artificial neural networks identified a strong personality as one of the leading causes of nursing burnout; it produced a 15% better result than traditional statistical instruments</td>
</tr>
<tr>
<td>Lee et al [43]</td>
<td>To determine how wearable devices can help people manage their itching conditions</td>
<td>Interview study: user experience and acceptance of the device</td>
<td>40 patients and 2 dermatologists</td>
<td>Machine learning–based itchector algorithm detected scratch movement more accurately when patients wore it for a longer duration</td>
</tr>
<tr>
<td>Marella et al [44]</td>
<td>To develop a semiautomated approach to screening cases that describe hazards associated with EHRs from a mandated, population-based reporting framework for patient safety</td>
<td>Mathematical/text and numerical data</td>
<td>NA</td>
<td>Naive Bayes Kernel resulted in the highest classification accuracy; it identified a higher proportion of medication errors and a lower proportion of procedural error than manual screening</td>
</tr>
<tr>
<td>Maziul et al [45]</td>
<td>To evaluate the impact of a wearable device on gait assist among patients with Parkinson disease</td>
<td>Mathematical/text data</td>
<td>18 patients and 5 healthy participants</td>
<td>AF-based Gait Assist was perceived as useful by the patients. Patients reported a reduction in freezing of gait duration and increased confidence during walking</td>
</tr>
<tr>
<td>McKnight [46]</td>
<td>To analyze patient safety reports.</td>
<td>Mathematical/text data</td>
<td>NA</td>
<td>Natural language processing improved the classification of safety reports as Fall and Assault; it also identified unlabeled reports</td>
</tr>
<tr>
<td>Moore et al [47]</td>
<td>To evaluate natural language processing’s performance for extracting abnormal results from free-text mammography and Pap smear reports.</td>
<td>Mathematical/text data</td>
<td>NA</td>
<td>The performance of natural language processing was comparable to a physician’s manual screening</td>
</tr>
<tr>
<td>Morrison et al [48]</td>
<td>To evaluate the usability and acceptability of ASSESS MS.</td>
<td>Interview study: feedback questionnaires, usability scales</td>
<td>51 patients, 6 neurologists, and 6 nurses</td>
<td>ASSESS MS was perceived as simple, understandable, effective, and efficient; both patients and doctors agreed to use it in the future</td>
</tr>
<tr>
<td>Muñoz et al [49]</td>
<td>To augment the relationship between physical therapists and their patients recovering from a knee injury, using a wearable sensing device</td>
<td>Interview study to understand how physical therapists work with their patients; user interface design considering usability and comfort</td>
<td>2 physical therapists</td>
<td>Machine learning–based wearable device correctly identified exercises such as leg lifts (100% accuracy) but also incorrectly identified three nonleg lifts as successfully performed leg lifts (3/18 false positives)</td>
</tr>
<tr>
<td>Nobles et al [50]</td>
<td>To identify periods of suicidality</td>
<td>Survey study: evaluating psychology students’ communication habits using electronic services</td>
<td>26 patients</td>
<td>The machine-learning model accurately identified 70% of suicidality when compared to the default accuracy (56%) of a classifier that predicts the most prevalent class</td>
</tr>
<tr>
<td>Ong et al [51]</td>
<td>To automatically categorize clinical incident reports</td>
<td>Mathematical/text and numerical data</td>
<td>NA</td>
<td>Naive Bayes and support vector machine correctly identified handover and patient identification incidents with an accuracy of 86.29%-91.53% and 97.98%, respectively</td>
</tr>
<tr>
<td>Park et al [52]</td>
<td>To compare discussion topics in publicly accessible online mental health communities for anxiety, depression, and posttraumatic stress disorder</td>
<td>Mathematical/text data</td>
<td>NA</td>
<td>Depression clusters focused on self-expressed contextual aspects of depression, whereas the anxiety disorders and posttraumatic stress disorder clusters addressed more treatment- and medication-related issues</td>
</tr>
<tr>
<td>Patterson et al [53]</td>
<td>To understand how transparent complex algorithms can be used for predictions, particularly concerning imminent mortality in a hospital environment</td>
<td>Interview study: group discussion</td>
<td>3 researchers</td>
<td>All participants gave contradicting responses</td>
</tr>
<tr>
<td>Study</td>
<td>Objective</td>
<td>Methods and Data</td>
<td>Study participants</td>
<td>Immediate outcome observed</td>
</tr>
<tr>
<td>-----------------------</td>
<td>---------------------------------------------------------------------------</td>
<td>------------------------------------------------------</td>
<td>--------------------</td>
<td>-----------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Pryor et al [54]</td>
<td>To analyze the use of a software medical decision aid by physicians and nonphysicians</td>
<td>Observation study; the study indirectly tested the usability and users’ trust in the device</td>
<td>34 clinicians and 32 nonclinical individuals</td>
<td>Physicians did not follow tool recommendations, whereas nonphysicians used diagnostic support to make medical decisions</td>
</tr>
<tr>
<td>Putnam et al [55]</td>
<td>To describe a work-in-progress that involves therapists who use motion-based video games for brain injury rehabilitation</td>
<td>Interview study to understand therapists’ experiences, opinions, and expectations from motion-based gaming for brain injury rehabilitation</td>
<td>11 therapists and 34 patients</td>
<td>Identifying games that were a good match for the patient’s therapeutic objectives was important; traditional therapists’ goals were concentration, sequencing, coordination, agility, partially paralyzed limb utilization, reaction time, verbal reasoning, and turn-taking</td>
</tr>
<tr>
<td>Sbernini et al [56]</td>
<td>To track surgeons’ hand movements during simulated open surgery tasks and to evaluate their manual expertise</td>
<td>Mathematical/sensor data</td>
<td>18 surgeons</td>
<td>Strategies to reduce sensory glove complexity and increase its comfort did not affect system performance substantially</td>
</tr>
<tr>
<td>Shiner et al [57]</td>
<td>To identify inpatient progress notes describing falls</td>
<td>Mathematical/text data</td>
<td>NA</td>
<td>Natural language processing was highly specific (0.97) but had low sensitivity (0.44) in identifying fall risk compared to manual records review</td>
</tr>
<tr>
<td>Songur and Top [58]</td>
<td>To analyze clusters from 12 regions in Turkey in terms of medical imaging technologies’ capacity and use</td>
<td>Mathematical/text and numerical data</td>
<td>NA</td>
<td>The study identified inequities in medical imaging technologies according to regions in Turkey and hospital ownership</td>
</tr>
<tr>
<td>Swangnetr and Kaber [59]</td>
<td>To develop an efficient patient-emotional classification computational algorithm in interaction with nursing robots in medical care</td>
<td>Survey study: self-assessment manikin questionnaire to measure emotional response to the robot</td>
<td>24 residents</td>
<td>Wavelet-based denoising of galvanic skin response signals led to an increase in the percentage of correct classifications of emotional states, and more transparent relationships among physiological responses and arousal and valence</td>
</tr>
<tr>
<td>Wagland et al [60]</td>
<td>To analyze the patient experience of care and its effect on health-related quality of life</td>
<td>Survey study regarding treatment, disease status, physical activity, functional assessment of cancer therapy, and social difficulties inventory</td>
<td>NA</td>
<td>Nearly half of the total comments analyzed described positive care experiences. Most negative experiences concerned a lack of posttreatment care and insufficient information concerning self-management strategies or treatment side effects</td>
</tr>
<tr>
<td>Wang et al [61]</td>
<td>To evaluate a population health intervention to increase anticoagulation use in high-risk patients with atrial fibrillation</td>
<td>Mathematical/text and numerical data</td>
<td>NA (data from 14 primary care clinics)</td>
<td>After pharmacist review, only 17% of algorithm-identified patients were considered potentially undertreated</td>
</tr>
<tr>
<td>Waqar et al [62]</td>
<td>To analyze patients’ interest in selecting a doctor</td>
<td>Survey study: systems evaluation from patients’ and doctors’ perspectives</td>
<td>NA (data from 3 hospitals)</td>
<td>The proposed system solved the problem of doctor recommendations to a good effect when evaluated by domain experts</td>
</tr>
<tr>
<td>Xiao et al [63]</td>
<td>To achieve personalized identification of cruciate ligament and soft tissue insertions and, consequently, capture the relationship between the spatial arrangement of soft tissue insertions and patient-specific features extracted from the tibia outlines</td>
<td>Mathematical/image data</td>
<td>20 patients</td>
<td>The supervised learning and prediction method developed in this study provided accurate information on soft tissue insertion sites using the tibia outlines</td>
</tr>
<tr>
<td>Valik et al [64]</td>
<td>To develop and validate an automated Sepsis-3–based surveillance system in a nonintensive care unit</td>
<td>Mathematical/text and numerical data</td>
<td>NA</td>
<td>The Sepsis-3 clinical criteria determined by physician review were met in 343 of 1000 instances</td>
</tr>
<tr>
<td>Bailey et al [65]</td>
<td>To study the implementation of a clinical decision support system (CDSS) for acute kidney injury</td>
<td>Interview and observation study: organizational work of technology adoption</td>
<td>49 clinicians</td>
<td>Hospitals faced difficulties in translating the CDSS’s recommendations into routine proactive output</td>
</tr>
<tr>
<td>Carayon et al [66]</td>
<td>To improve the usability of a CDSS</td>
<td>Experimental study: simulation and observation to evaluate the usability</td>
<td>32 clinicians</td>
<td>Emergency physicians faced lower workload and higher satisfaction with the human factors–based CDSS compared to the traditional web-based CDSS</td>
</tr>
<tr>
<td>Study</td>
<td>Objective</td>
<td>Methods and Data</td>
<td>Study participants</td>
<td>Immediate outcome observed</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>---------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------</td>
<td>--------------------</td>
<td>-----------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Parekh et al [67]</td>
<td>To develop and validate a risk prediction tool for medication-related harm in older adults</td>
<td>Mathematical/numerical data</td>
<td>1280 elderly patients</td>
<td>The tool used eight variables (age, gender, antiplatelet drug, sodium level, antidiabetic drug, past adverse drug reaction, number of medicines, living alone) to predict harm with a C-statistic of 0.69</td>
</tr>
<tr>
<td>Gilbank et al [68]</td>
<td>To understand the needs of the user and design requirements for a risk prediction tool</td>
<td>Survey and interview study: informal, semistructured meetings</td>
<td>15 stakeholders from hospitals, academia, industry, and nonprofit organizations</td>
<td>Nine physicians emphasized the need for a prerequisite for trusting the tool. Many participants preferred the technology to have roles complementary to their expertise rather than to perform tasks the physicians had been trained for. Having a tailored recommendation for a local context was deemed critical</td>
</tr>
<tr>
<td>Miller et al [69]</td>
<td>To understand the usability, acceptability, and utility of AI-based symptom assessment and advice technology</td>
<td>Survey study to measure ease of use</td>
<td>523 patients</td>
<td>425 patients reported that using the Ada symptom checker would not have made a difference in their care-seeking behavior. Most patients found the system easy to use and would recommend it to others</td>
</tr>
<tr>
<td>ter Stal et al [70]</td>
<td>To analyze the impact of an embodied conversational agent’s appearance on user perception</td>
<td>Interview study: Acosta and Ward Scale [71]</td>
<td>20 patients</td>
<td>The older male conversational agent was perceived as more authoritative than the young female agent ($P = .03$). Participants did not see an added value of the agent to the health app</td>
</tr>
<tr>
<td>Gabrielli et al [72]</td>
<td>To evaluate an online chatbot and promote the mental well-being of adolescents</td>
<td>Experimental, participatory design, and survey study to measure satisfaction</td>
<td>20 children</td>
<td>Sixteen children found the chatbot useful and 19 found it easy to use</td>
</tr>
<tr>
<td>Liang et al [73]</td>
<td>To develop a smartphone camera for self-diagnosing oral health</td>
<td>Interview Study to measure usability (NASA-TLX)</td>
<td>500 volunteers</td>
<td>Two experts agreed that OralCam could give acceptable results. The app also increased oral health knowledge among users</td>
</tr>
<tr>
<td>Chatterjee et al [74]</td>
<td>To access the feasibility of a mobile sensor-based system that can measure the severity of pulmonary obstruction</td>
<td>Mathematical/numerical data</td>
<td>91 patients, 40 healthy participants</td>
<td>Most patients liked using a smartphone as the assessment tool; they found it comfortable (mean rating 4.63 out of 5 with $\sigma = 0.73$)</td>
</tr>
<tr>
<td>Beede et al [75]</td>
<td>To evaluate a deep learning–based eye-screening system from a human-centered perspective</td>
<td>Observation and interview study: unstructured</td>
<td>13 clinicians, 50 patients</td>
<td>Nurses faced challenges using the deep-learning system within clinical care as it would add to their workload. Low image quality and internet speed hindered the performance of the AI system</td>
</tr>
</tbody>
</table>

aNA: not applicable; these studies have only used data for their respective analyses without involving any human participant (user).
bEHR: electronic health record.
cAI: artificial intelligence.

We observed various algorithms in the final selection, with machine learning being the most common (n=18). Some studies also compared different algorithms based on analytical performance. However, few studies (n=5 against clinical/baseline standards, n=5 against clinicians) compared their AI models against a standard measure. The general flaws and trends of such studies have been addressed in our prior work [7]. Overall, our review indicates that the dimensions of usability, user’s perception, workload, and trust in AI have been the most common interest of research in this field.
Table 2. Artificial intelligence (AI) studies that primarily focused on machine learning (ML) algorithm development (n=18).

<table>
<thead>
<tr>
<th>Reference</th>
<th>AI/ML recommended by the study</th>
<th>Other AI/ML/non-AI used in the study</th>
<th>Proposed AI model(s) for comparison (1=compared; 0=not compared)</th>
<th>Other AI systems</th>
<th>Existing system (not AI)</th>
<th>Clinical or gold standard</th>
<th>Clinicians or user</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aldape-Pérez et al [24]</td>
<td>Delta Associative Memory</td>
<td>AdaBoostM1; bagging; Bayes Net; Dagging; decision table naïve approach; functional tree; logistic model trees; logistic regression; naïve Bayes; random committee; random forest random subspace; Gaussian radial basis function network; rotation forest; simple logistic; support vector machine</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Azari et al [25]</td>
<td>Random forest and hidden Markov model</td>
<td>Not applicable</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Balani and De Choudhury [26]</td>
<td>Perceptron</td>
<td>Naïve Bayes; k-nearest neighbor; decision tree</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cvetković and Cvetković [28]</td>
<td>Neural network and fuzzy logic</td>
<td>Not applicable</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Gao et al [34]</td>
<td>AdaBoost</td>
<td>k-nearest neighbor, support vector machine, decision tree, random forest, naïve Bayes</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Hu et al [36]</td>
<td>Deep neural network</td>
<td>Deep neural network with different inputs</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Kandaswamy et al [38]</td>
<td>Random forest</td>
<td>Naïve Bayes; logistic regression; support vector machine</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Komogortsev and Holland [39]</td>
<td>Supervised support vector machine</td>
<td>Unsupervised support vector machine and unsupervised heuristic algorithm developed by the authors</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Marella et al [44]</td>
<td>Naïve Bayes kernel</td>
<td>Naïve Bayes; k-nearest neighbor; rule induction</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Nobles et al [50]</td>
<td>Deep neural network</td>
<td>Support vector machine</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Ong et al [51]</td>
<td>Naïve Bayes; support vector machine with radial-bias function</td>
<td>Support vector machine with a linear function</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Shiner et al [57]</td>
<td>Natural language processing</td>
<td>Incident reporting system; manual record review</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Wagland et al [60]</td>
<td>Did not recommend any particular algorithm</td>
<td>Support vector machine; random forest; decision trees; generalized linear models network; bagging; max-entropy; logi-boost</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Waqar et al [62]</td>
<td>Hybrid algorithm developed by the authors</td>
<td>Not applicable</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Xiao et al [63]</td>
<td>The authors developed a new algorithm</td>
<td>Linear regression with regularization; LASSO; k-nearest neighbor; population mean</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Valik et al [64]</td>
<td>The authors developed a new algorithm</td>
<td>Not applicable</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Parekh et al [67]</td>
<td>The authors developed an algorithm based on multi-variable logistic regression</td>
<td>Not applicable</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Chatterjee et al [74]</td>
<td>Gradient boosted tree</td>
<td>Random forest, adaptive boosting</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

*LASSO: least absolute shrinkage and selection operator.

**Perception, Usability, Workload, and Trust**

**Perception**

The perception of users was analyzed by several studies to adequately assess the quality of the proposed AI-based recommender system. Some studies incorporated perceptions of both patients and doctors [62,73] in developing their AI systems. Another study interviewed providers (therapists) about their experiences, opinions, expectations, and perceptions of a motion-based game for brain injury rehabilitation to guide the
design of the proposed AI-based recommender system, which was a case-based reasoning (CBR) system [55]. The AI system ASSESS MS was also developed and evaluated based on users’ perceptions [48]. Studies included in our review that developed AI-based apps [27,29], AI robots [30], and wearable AI devices such as Gait Assist [45] and Itchtector [43] also accounted for users’ perceptions. From a psychological perspective, emotions might facilitate perception [76]. One study in our review measured users’ perception of an AI-based conversational agent [70]; and another study developed an AI algorithm for real-time detection of patient emotional states and behavior adaptation to encourage positive health care experiences [59].

Usability

Some studies in our review performed usability testing of AI systems. For example, one study used AI to develop an adaptable CBR to help therapists ensure proper usability and functioning of CBR [55]. Guided by users’ needs, one study [27] developed an AI application (SMILY) to ensure good usability. Users found the clinical information to have higher diagnostic utility while using SMILY (mean 4.7) than while using the conventional interface (mean 3.7). They also experienced less effort (mean 2.8) and expressed higher trust (mean 6) in SMILY than with the conventional interface (mean 4.7; P<0.01), as well as higher benevolence (mean 5.8 vs 2.6; P<0.001). Another study included in our review noted the literacy gap as a significant hurdle in the usability of an AI-based face-reading app, and identified the impact of adaptability and cultural sensitivity as a limiting factor for usability [29]. Another study codesigned an AI chatbot with 20 students and performed a formative evaluation to better understand their experience of using the tool [72]. Two recent studies measured the perceived usability of AI-based decision-making tools: Ada, an AI tool that helps patients navigate the right type of care [69], and PE-Dx CDS, a tool for diagnosing pulmonary embolism [66]. However, in another study, the researchers primarily focused on developing the algorithm for assessing the severity of pulmonary obstruction and obtained users’ feedback on the end product [74]. Poor usability often leads to an increased workload, particularly when the user (provider or patient) is not trained in using the AI system, device, or app.

Workload

Caregivers are subject to workplace stress and cognitive workload, mostly due to the complexities and uncertainty of patient health and related treatment [77-79], and AI promises to minimize the health care workload through the automation of various levels. Nevertheless, if an AI system or program is poorly designed, the workload may possibly be elevated. Two studies in our review used a radial basis function network to assess burnout among nurses, and consequently captured the nonlinear relationship of the burnout process with the workload, work experience, conflicting interaction, role ambiguity, and other stressors [41,42]. The demand-control theory of work stress implies that workload abnormalities and job intensity can aggravate user fatigue by excessive workloads and trigger anxiety [80]. According to Maslach and Leiter [81], a mismatch between one’s skill sets (ability to perform a task) and responsibility (skills required to complete a task) intensifies users’ workload. Three studies in our review were invested in minimizing users’ workload by assessing the usability of AI systems such as ASSESS MS [48], Gait Assist [45], and SMILY [27].

Trust

Trust shapes clinicians’ and patients’ use, adoption, and acceptance of AI [6]. Trust is a psychological phenomenon that supports the inconsistency between the known (clinicians’ awareness, patient experience) and the unknown (deep-learning algorithms). Three studies included in our review measured user trust in health care AI systems. One study reported that the anthropomorphism of AI-based care robots has no influence on providers’ trust but was significantly related to the level of automation and intention to work with the robot [30]. This study proposed that providers who trusted robots more intended to work with them and preferred a higher automation level [30]. A recent perspective discusses the risk of overreliance or maximum trust in AI (automation) and instead suggests optimal trust between the user and AI system [6]. Besides experience, expertise, and prior knowledge, the performance of the AI technology also determines users’ trust. A study included in our review, using a poststudy questionnaire, found that doctors (pathologists) expressed higher trust in SMILY, an AI-based application, due to its better performance, interface, and higher benevolence compared with the conventional app [27]. By contrast, another study reported lower trust of experienced physicians in an AI-based recommendation tool due to its inefficient performance [54]. Based on patient data, expert physicians were able to identify the alternative and better explanation for patient health compared to the AI-based tool [54]. A recent study identified the impact of the AI interface on user’s trust [68]. Physicians in this study considered AI’s transparency and performance as facilitators of engendering trust.

User-Centered Design

A user-centric design requires multidisciplinary cooperation between HFE experts, technologists, and end users. The inadequacy of a user-centered design also hinders user perception, usability, and trust, and increases the possibility of errors. The majority of the health care AI literature focuses on quantitative constraints, including performance metrics and precision, and is less focused on the user-centric development of AI technologies. Due to the lack of standard guidelines [7,16], not much research has invested in incorporating a user-centered design in AI-based technologies within the health care industry. In this review, we identified studies that performed experiments involving clinicians and patients, and consecutively evaluated their AI system’s (eg, app, wearable device) interface [27], applicability [27,29], and appearance (anthropomorphism) [30] to ensure user-centeredness. Other studies [43,45,48,49,55,62] also addressed user requirements such as wearability and privacy concerns. A recent study further acknowledged the importance of a user-centered clinical field study, and identified external factors such as low lighting, expensive image annotation, and internet speed that can deter the effectiveness of AI systems for diagnosing diabetic retinopathy [75].
Discussion

Main Findings

Research concerning AI in health care has shown promise for augmenting the quality of health care. However, there is a need for more theoretical advances and interventions that cover all levels and operations across the health care system. We need a systematic approach to safely and effectively bring AI into use, providing human factors, user-centered design, and delivery and implementation science. Many current AI models focus on engineering technology (informatics concepts) and do not sufficiently discuss the relevance of HFE in health care [82]. In this review, we explored and portrayed the involvement of HFE journals and conferences in health care AI research. We identified 48 studies, trending as more publications in recent years, which shows increased attention of the HFE community in this field.

Although advancement and focus have been made in the use of machine learning/AI to develop prediction and classification models, little research has been devoted to real-world translations with a user-centered design approach. To determine the diverse relationships between individuals and technology within a work environment, it is necessary to provide a better explanation as to how AI can be part of the overall health care system through a variety of HFE methods such as the Systems Engineering Initiative for Patient Safety (SEIPS) [83]. The SEIPS provides a framework that helps in comprehending the work system (people, tools and technologies, tasks, working environment, and organization), process (clinical process and process assisting the same), and outcomes (patient outcome, organizational outcome) in the health care domain [83]. This framework also helps to assess and understand the complex interaction between elements of the work system, and shows the impact of any technology-based intervention on the overall system [83].

This review also highlights the need for a systematic approach that evaluates AI’s impact (effectiveness) on patient care based on its computational capabilities and compatibility with clinical workflow and usability. Although some studies have acknowledged AI’s challenges from both humans factors (biases and usability) [84] and technical (quality of training data and standardization of AI) [7] standpoints, less emphasis has been given so far to the impact of AI integration into clinical processes [16] and services as well as to the user-centered design of AI systems for better human-AI interaction [84,85]. At this stage, where human beings and AI come together, challenges to human factors will likely arise.

Next Steps

The next push for researchers should be to move AI research beyond solely model development into sociotechnical systems research and effectively use human factors principles. HFE researchers should consider users’ needs, capabilities, and interactions with other elements of the work system to ensure the positive impact of AI in transforming health care. Clinical systems are not inherently equivalent to predictable mechanical systems and need a systematic approach. One of the pivotal myths of automation is the assumption that AI can replace clinicians [33]. In fact, the use of AI can shape the activities and duties of clinicians, and might help them in their decision-making. In the domain of medical imaging, AI has shown great promise and is increasing rapidly. For instance, on January 18, 2021, an image analysis platform named AI Metrics received US Food and Drug Administration (FDA) 510(k) clearance [86]. Likewise, in the last 5 years, approximately 222 AI-based medical devices have been approved in the United States [87]. As AI continuous to grow, the associated risks also increase. Many health care AI systems are poorly designed and not evaluated thoroughly [14], and have neglected clinicians’ limited absorptive and cognitive capacities and their ability to use AI in clinical settings under a high cognitive workload [88-90]. Incorrect usage or misinterpretation of AI, similar to that of EHRs [91], may also result in patient harm. Therefore, more HFE research should focus on cognitive factors (biases, perceptions, trust), usability, situation awareness, and methodological aspects of AI systems.

Usability

A user-centered design is essential for health care technologies, where the user is centrally involved in all phases of the design process [92]. However, when the user environment and activities are varied, designing standardized protocols for health care devices and software is complicated. As stated in this study, the problem further increases due to the heterogeneity of applications and AI variants. The human-computer interaction community has developed different user-centered design techniques. However, these methods are often understudied by software development teams and organizations [93].

Usually, AI algorithms are complex, opaque, and thus difficult to understand. Therefore, it might be difficult for clinicians/end users to understand and interpret AI outcomes effectively without adequate instruction. Cognitive ergonomics is a fundamental principle dealing with usability issues [94]. Necessary procedural information stored in long-term memory is required to use a technical device [95]. Kieras and Polson [95] suggested the cognitive complexity theory (CCT) explicitly addressing the cognitive complexity of the user-to-device/interface interaction by explaining the user’s goals on the one hand and the computer system reaction on the other hand using production rules. The laws of production can be viewed as a series of rules in the form of IF conditions (display status) and THEN actions (input or action taken by the user). According to CCT, cognitive complexity is defined as the number of production rules segregated and learned in a specific action sequence. The definition of cognitive complexity in an AI-based health app can be as helpful as the definition of production rules (ie, the specification of what the system says and how users react) and factors that may contribute concurrently to the app’s complexity (ie, interface, menu structure, the language of communication, transparency of functions’ naming). It is, however, debatable whether the mere counting of production rules will reasonably assess the troubles perceived by users, considering that various factors contribute equitably to cognitive complexity. Cognitive computing systems [96], which are computing systems that can incorporate human-like cognitive abilities, can also augment and safeguard health care AI by making AI adaptive (learning from a changing context and predicting human needs and behaviors). This approach is essential for ensuring that AI systems are not only effective but also safe and reliable in clinical settings.
environment, changing patient health, changing clinician’s requirements), interactive (easier human-AI interaction, better usability, easy to understand), iterative and stateful (narrowing down on the problem, considering past decisions/consequence while making current recommendations/tasks), and contextual (consider contextual elements) [96].

Moreover, challenges and hardships perceived by users might be a function of several factors not limited to the user’s experience, knowledge, intention of use, and working environment [97]. Therefore, an adaptable usability scale that encompasses the complexity of AI and the common usability factors applicable to that particular system or software should be created by HFE researchers. Perception of an AI system or its perceived ease of use can potentially be a function of users’ cognitive and physical abilities. Additionally, the obvious question is, where should user-centered design techniques and knowledge be considered in the life cycle of AI’s development?

**Trust and Biases**

Human factors research on “automation surprises” primarily began with large-scale industrialization that involved autonomous technologies [84,98,99]. The automation surprise arises when an automated machine acts counterintuitively [100]. In health care, automation surprises might lead to confusion, higher workload, distrust, and inefficient operations [101]. In the health care environment, inadequate mental models and insufficient information about AI-based technology might lead to automation surprises and negatively influence trust [6]. Trust can also be hindered if an automated system tends to deter clinicians’ performance [6]. Research evaluating the performance of radiologists observed their deterring performance when aided by a decision support system [102]. Therefore, more HFE studies are needed that explore the factors and design requirements influencing users’ and clinicians’ optimal trust in AI. Future studies should also focus on patient trust in AI-generated recommendations.

When automated diagnostic systems are used in real-life clinics, they most likely are in the form of assistant or recommender systems where the AI system provides information to clinicians or patients as a second opinion. However, if the suggestions made by AI are entirely data-driven without accounting for the user’s opinion, as is the case for current designs, users could be biased toward or against the suggestion of the AI system [103]. Optimizing such user-AI trust interplay remains a challenge that HFE experts should consider as their future endeavor.

It should be noted that advocating for trust in automation for a prolonged time can also promote automation bias. Aviation studies have recorded instances of automation biases where pilots could not track vital flight indicators in the event of failure due to overreliance on the autopilot [104,105]. A review of automation bias focusing on the health care literature noted that the complexity of any assignment and the workload increased the likelihood of excessive reliance on automation [106], which can be detrimental to patient safety. Human factors such as cognitive ergonomics and a user-centered design should be utilized efficiently to minimize the health care AI system’s automation biases.

**Situation Awareness**

Situation awareness is defined as “the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future” [107]. “Good” situation awareness is a prerequisite to better performance [84,107]. There might be an ongoing discussion around maximum versus optimum situation awareness. It is critical to understand that the optimum situation awareness is not necessarily the maximum situation awareness [108]. Maximizing the user’s situation awareness does not necessarily yield the best outcome (decisions from a human-AI collaboration) [108]. For example, concentrating on irrelevant details such as radio commercials, talking passengers, or the colors of other cars while driving may unnecessarily consume the driver’s working memory, increase the workload, or even act as a distraction [109].

Similarly, in a clinical setting, it is better to achieve optimal situation awareness rather than maximum situation awareness. Many studies have shown the deterring impact of excessive and unnecessary information on clinical work [110,111]. For example, false or irrelevant clinical alarms may increase the tension of nurses and even distract them. Performing critical health care tasks (such as administering narcotic medication, watching telemetry monitors) demands optimal situation awareness [112]; however, unnecessary or irrelevant situation awareness can disturb clinicians’ attention and working memory.

Exploration of AI’s influence on clinicians’ situation awareness has not been studied extensively. More HFE-based research is needed to further explain the concept of optimal situation awareness in AI design. Both humans and AI each have skepticism regarding the information generated in their surroundings and extract the data that seem vital for clinical decision-making.

**Ecological Validation**

The development, evaluation, and integration of sophisticated AI-based medical devices can be a challenging process requiring multidisciplinary engagement. It may enable a personalized approach to patient care through improved diagnosis and prognosis of individual responses to therapies, along with efficient comprehension of health databases. This solution has the power to reinvigorate clinical practices. Although the advent of personalized patient treatment is provocative, there is a need to evaluate the true potential of AI. The performance of AI depends on the quantity and quality of data available for training, as acknowledged in recent review papers [7,16]. Perhaps one of the most essential facts from the HFE viewpoint is that poor usability causes improper, inaccurate, and inefficient use [113]. Although the importance of usability testing and a user-centered design for medical devices has been substantially stated by the FDA [114] and other HFE experts, both regulatory guidelines and evaluation approaches fail to reflect the challenges faced by clinicians during their routine clinical activity [115]. In other words, most studies identified in our review were performed in a controlled environment, therefore lacking ecological validity. This finding is consistent with most other research in the field of AI and health care. Recent systematic reviews [7,16,116] analyzing AI’s role and performance in health care
acknowledged that AI systems or models were often evaluated under unrealistic conditions that had minimal relevance to routine clinical practice.

Users under stress and discomfort might not be efficient in utilizing AI devices with poor usability. Unlike research or controlled settings, a clinical setting demands multitasking where clinicians (nurses) have to attend to several patients with different ailments. They also have to write clinical notes, monitor health fluctuations, administer critical medications, float to different departments during shortage of staff, educate new nurses, and respond to protocols in cases of emergency. Under such a working environment and cognitive workload, interpreting or learning to use an AI system that is not designed appropriately can be challenging and risky. Therefore, an AI system that perfectly qualifies usability tests in a research setting may fail in a clinical environment. Given these limitations, the few studies in our review that compared their AI model with clinical standards (see Table 2) are less relevant because the comparisons against clinical standards were made in an (ideal) controlled environment or without providing contextual information about the patient and the environment [117]. Moreover, the work system elements also differ substantially from an intensive care unit to an outpatient clinic. Therefore, AI-based medical systems must be evaluated in their respective clinical environment to ensure safer deployment.

Limitations of the Review
This review does not include the complete available literature but was constrained within the selected journals and conferences. Studies investigating human-AI interactions in a health care context or leveraging HFE principles to evaluate health care AI systems published in non-HFE venues such as pure medical or informatics journals have not been included in this review. Notwithstanding these constraints, our analysis identified possible research gaps in the health disciplines that could, if addressed, help mobilize and integrate AI more efficiently and safely.

Conclusion
HFE researchers should actively design and implement AI, and perform dynamical assessments of AI systems’ effects on interaction, workflow, and patient outcomes. An AI system is part of a greater sociotechnical system. Investigators with HFE expertise are essential when defining the dynamic interaction of AI within each element, process, and result of the work system. This means that we ought to adapt our strategy to the situations and contexts in the field; simultaneously, we must also find practical ways of generating more compelling evidence for our research.

Acknowledgments
We thank Mr. Nikhil Shetty and Ms. Safa Elkefi, graduate students at Stevens Institute of Technology, for assisting us with the preliminary literature search. This research received no specific grant from any funding agency in public, commercial, or not-for-profit sectors.

Authors’ Contributions
OA conceived and designed the study; developed the protocol; participated in data collection (literature review), analysis, and interpretation; drafted and revised the manuscript; and approved the final version for submission. AC designed the study; developed the review protocol and graphical illustrations; participated in the literature review, analysis, and interpretation; drafted and revised the manuscript; and approved the final version for submission.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Journal and conference proceedings list for the review.
[DOCX File, 19 KB-Multimedia Appendix 1]

References


71. Asan & ChoudhuryJMIR HUMAN FACTORS


Abbreviations

AI: artificial intelligence
CBR: case-based reasoning
CCT: cognitive complexity theory
EHR: electronic health record
FDA: Food and Drug Administration
HFE: human factors and ergonomics
SEIPS: Systems Engineering Initiative for Patient Safety

Edited by A Kashinik; submitted 25.02.21; peer-reviewed by M Sujan, M Knop; comments to author 28.03.21; revised version received 14.04.21; accepted 03.05.21; published 18.06.21

Please cite as:
Asan O, Choudhury A
Research Trends in Artificial Intelligence Applications in Human Factors Health Care: Mapping Review
JMIR Hum Factors 2021;8(2):e28236
URL: https://humanfactors.jmir.org/2021/2/e28236
doi: 10.2196/28236
PMID: