HFES Policy Statement:
Artificial Intelligence in Health Care

Artificial intelligence (AI) is being proposed as a tool to aid health care professionals in detecting, diagnosing, treating and monitoring illnesses, as well as for examining large datasets of biological data for medical research at the cellular and genetic level. Further there is a confluence of AI with wearable and sensing technologies that can generate very large datasets from the general population, allowing for new forms of bio-monitoring and direct delivery of AI-generated diagnoses or health alerts to both clinicians and patients.

While potentially promising, there are a number of socio-technical challenges and limitations for these goals as outlined in a recent report from the National Academy of Medicine[1], including: a) the accessibility, quality, and standardization of data for training AI algorithms, b) ensuring that data sets are ethically collected(e.g. informed consent) and include representative populations to avoid inequities in health care outcomes, c) a need for AI transparency and support for establishing user trust, d) allowing for human over-ride of AI, e) creating training and educational support for health care professionals who use AI systems, f) establishing best practices, and g) establishing a flexible approach to regulation.

Human Performance Challenges Associated with Health Care AI Technologies

In addition to these technical issues, there are a number of significant challenges faced by the health care professionals, researchers, and general population who receive recommendations from AI systems that must be addressed in order to support accurate human decision making based on AI applications, thus avoiding human error and enhancing safety.

AI Limitations and Software Biases – A key challenge associated with AI systems is that while many people may view them as omniscient, in fact they are subject to significant limitations that can affect their reliability. Today's AI systems apply machine learning techniques to large data-sets to find hidden patterns that are used to create data classifications. There are several key limitations of this approach.

a) Datasets used for learning/training AI algorithms may be generated based on noisy or limited data. For example, heart rate monitors on consumer wearable devices have been shown to produce 20-70% of readings with errors greater than 20 BPM, and in many cases errors of over 50 BPM[2].

b) AI algorithms may contain their own decision biases[3; 4]. For example, if the training data used to develop the AI algorithm contains an under-representation of individuals from different racial groups, genders, ages, body types, or other factor, the AI algorithm will be similarly limited and may provide incorrect analyses with other groups[1]. This type of bias tends to be deeply encoded and not readily apparent to either its developers or users.

c) The ability of this type of AI is fundamentally limited as it has no model of the underlying system that is needed for projecting future outcomes[5]. AI created from non-supervised learning systems can only detect patterns. It cannot understand diagnoses or prognoses, or the complex factors affecting how the body functions.

Best Applications for AI **–** AI may be potentially applied to health care systems in a number of ways to: a) improve people's understanding of the information by monitoring, filtering or integrating data for presentation to users, b) make recommendations or decisions, or c) carry out actions.[6-8] Over 30 years of research on how these various options affect human performance shows that the higher the level of automation, the more likely people are to be out-of-loop and unaware of when human intervention is needed.[8; 9] Therefore, unless the AI is highly reliable, full automation that allows the AI to detect, decide and carry-out actions should be avoided.

Further, AI that provides recommendations can lead to *decision biasing*. If the system is correct in its recommendations people are more likely to make a correct decision, but if it is incorrect then people are more likely to make the erroneous choice than if they had not been provided with the AI recommendation at all. [10-13]

This shows that human performance is not independent of the AI tool, but rather is significantly affected by it, with inaccurate AI potentially decreasing the overall reliability of the combined human-AI system.[14]

The best application of AI is to aid in integration of information across multiple large data sets to improve monitoring of key events and people's understanding of data and ability to project based on it. Significant benefits can be found from systems that aid people by reducing unnecessary searching, sorting and transforming data to support their decision needs[14].

Trust and AI Transparency – In order to be useful, people must trust the output of the AI when it is correct, and must know when to reject that output when it is incorrect or inappropriate for the situation.[15] They must also be aware of when the AI system is not functioning properly so that they avoid reliance on unreliable data and can take corrective actions. This requires *AI transparency* which involves presenting information to users on the level of reliability of the AI for the situation at hand. Transparency means that users should be provided with information on how well the AI is working, the reliability of the underlying data or sensors that feed the AI, and its level of confidence in any assessments or recommendations that it makes.[14]

Further, users need to understand how the automation works in terms of what its capabilities are and its limitations for addressing different types of situations and classes of data.[14] AI transparency is important for not just understanding the overall reliability and robustness of the system in general, but for allowing people to properly calibrate their trust in real-time. AI that provides just-in-time information with the intention of serving as a decision support tool (in line with Appropriate Use Criteria[16]), and specifically for diagnostic testing, must be transparent about the capabilities, confidence and variables considered within the AI model.

Under/Over Diagnosis – A critical factor for any system is the trade-off between under-diagnosis of some condition (type I error) and over-diagnosis of that condition (type II error). The sensitivity of any detection technology must be finely tuned between these two types of error. If the system is too conservative in its criteria, it may miss many cases where disease is present. If it is too liberal in its criteria, it may generate many false alarms, leading to alarm fatigue, high workload, and loss of trust in the system. Currently many systems opt for avoiding a missed problem and therefore generate a high level of alarm fatigue.

For example, a watch that collects EKG data may provide alerts to the wearer if it detects an unsafe heart condition, and recommend that a physician be consulted.   The utility of these alerts is significantly degraded, however, if there are a high number of false alarms. When AI generates a high number of false alerts (e.g. due to sensor problems or limitations of the AI algorithm), a cry wolf effect can occur and people ignore the alerts in general (i.e. lose trust), thereby missing legitimate problems when they occur.[17] In addition, many people will actively disable alarm systems that activate too frequently and have too high of a false alarm rate.[18]

To avoid these problems, the reliability of the system should be increased as much as possible to avoid false alerts, alarm messages should be clear and unambiguous, and support for alarm verification should be provided.[14] Further, users should also be made aware of the course of actions that will occur should an alert occur (e.g., who will be contacted; alert protocols).

Socio-Technical Systems – AI technologies need to support the cognitive work performed by complex, distributed teams composed of clinicians, patients and caregivers. AI technologies are part of a larger socio-technical (work) system and need to be designed so that they fit the entire socio-technical (work) system.[19; 20] In particular, it is critical to pay attention to the challenges of integrating AI technologies into the clinical workflow to make sure it fits with their temporal demands and order of tasks. The way in which home healthcare technologies (e.g. AI-based symptom checkers used by patients) will support the workflow of physicians and other members of the care team also must be addressed.[21] In addition, there are concerns about how AI technologies will change the interactions between patients (who may not understand their limitations) and health care professionals[22], potentially affecting quality of care and the expense of unneeded medical tests.

Usability & Workload – To address these problems and enhance the reliability and efficacy of AI applications, the usability and workload of AI systems applied to healthcare must be addressed during system development. The users of AI systems must be carefully integrated into AI development programs to aid in identifying the appropriate data schedules and formats, and an understanding of the context and constraints associated with their use. For example, many AI systems for healthcare will be used in home and community environments, which can vary considerably. Some rural settings may have limited or slow internet access that can affect how well the AI system will work in some locations. In addition, information generated from AI must be readily available, at appropriate time intervals, and in a form that is readable and useful for the workflow of healthcare providers.

The usability of AI technology includes addressing its learnability, efficiency, memorability, likelihood of errors, and user satisfaction. AI applications must be tested with representative user groups (including healthcare providers, consumers, informal caregivers, and/or emergency responders) to identify usability problems and correct them in advance. Products or systems that are not sufficiently usable can lead to significant increases in user workload and errors, and may frequently be abandoned.

Training – Because AI is often intended to support decision making, user training within the context of medical care is necessary. As caregivers are making complex decisions for a patient, there are many criteria that influence *how* and *when* a tool should best be used, and how to avoid hidden problems, such as decision bias.

Reliance on highly technical user manuals is not sufficient for supporting the training needs of users who must develop detailed cognitive models of how the AI works. These mental models are needed to direct user attention, properly interpret AI recommendations, and to form accurate expectations regarding its capabilities and limitations. To develop these mental models, experiential learning with simulated cases is needed, as well as focused instruction. Training protocols (based on the best human factors training research and guidelines) must be developed for new AI systems for healthcare. These protocols must be carefully tested with user groups and then made readily accessible for supporting new users of this technology.

Policy Recommendations:

(1) AI Applications — AI applications should be user-centered***.***

a. AI systems should be best employed to integrate multiple large datasets to identify potential problems or healthcare issues, and to support physician decision making. Actual/Final diagnoses of conditions should be left to physicians who are trained in differential diagnosis and have more information on variables that can affect proper diagnosis.

b. Efforts to design and develop AI systems for healthcare applications should be team-based and include expertise from medicine and human factors researchers and designers.

(2) *AI Transparency* —AI systems for Healthcare must provide a high degree of transparency to support user understanding of the reliability of its recommendations and its limitations.

a. User interface designs should make the underlying algorithms and their behavior interpretable so that its capabilities and limits are clear to users and healthcare providers.

b. Any limitations of under-lying datasets (e.g. lack of inclusion of certain populations) should be clearly stated.

c. System confidence in recommendations, diagnoses or classifications provided should be provided to the user.

d. The basis for system recommendations, diagnoses or data classifications (e.g. relevant features, characteristics, parameters, rules or other criteria) should be clearly identified for human review.

e. AI outputs should be presented to the user in a way that is valuable for 'just-in-time' use. It must be usable within the context of normal clinical work, not requiring additional deep interrogation of the AI behind it.

f. The AI must operate in line with evidence based best practice. In any case where an AI recommendation deviates from established best practice, this should be clearly alerted.

(3) User Training — Detailed training should be provided to users of AI systems.

a. Users of AI systems should be provided with sufficient training on the capabilities, limitations, and behaviors of these systems (including the range of healthcare conditions they can handle) so that users obtain an accurate mental model required for effective oversight and interaction with them.

b. New user training should be provided on any software updates that are made over the course of the system's lifetime so that the AI's behavior remains understandable to the user.

(4) *User Testing* – AI systems applied to healthcare should be carefully user tested to ensure high levels of human-system performance.

a. AI systems for health care should require FDA approval for use, including appropriate human factors design and testing (e.g. 21 CFR 820.30, Design Controls).

b. User testing should be applied to ensure the systems provide a high level of usability and safety, avoid user errors (e.g. user understanding of information, decision biasing, user interaction), fit well with existing workflows, and do not create extra workload for system users.

c. A wide range of users and use environments should be included in these tests, ensuring equally effective deployment and interpretation of the AI system by users with different levels of clinical experience, interaction with the system, and responsibility for patient care.

(5) *Support for Research on AI in Healthcare* — Funding from Congress (via NSF, NIH, NIST, etc.) is needed to address key challenges in the effective use of AI in healthcare, including:

a. Research on education and training when developing and deploying AI for healthcare applications, including methods for creating good understanding of AI capabilities and limitations, and methods for overcoming decision biasing,

b. Research on effective user interface design approaches for creating AI transparency to support accurate user understanding of AI and appropriate levels of trust across different situations,

c. Research on technical, social, and ethical issues in AI-driven health treatment plans, and

d. Research to support the update of FDA human factors design guidance and standards to address the design and testing of AI systems for safety and effectiveness.

References

1. Matheny, M., Israni, S. T., Ahmed, M., & Whicher, D. (Eds.). (2020). Artificial Intelligence in healthcare: The hope, the hype, the promise, the peril. Washington, DC: National Academy of Medicine.
2. Burke, M., & Whelan, M. (1987). The accuracy and reliability of commercial heart rate monitors. British journal of sports medicine, 21(1), 29-32.
3. Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., & Tsaneva-Atanasova, K. (2019). Artificial intelligence, bias and clinical safety. BMJ Qual Saf, 28(3), 231-237.
4. Osoba, O. A., & Welser IV, W. (2017). An intelligence in our image: The risks of bias and errors in artificial intelligence: Rand Corporation.
5. Pearl, J., & Mackenzie, D. (2018). The book of why: The new science of cause and effect. New York: Basic Books.
6. Endsley, M. R., & Kaber, D. B. (1997). The use of level of automation as a means of alleviating out-of-the-loop performance problems: A taxonomy and empirical analysis. In P. Seppala, T. Luopajarvi, C. H. Nygard & M. Mattila (Eds.), 13th Triennial Congress of the International Ergonomics Association (Vol. 1, pp. 168-170). Helsinki: Finnish Institute of Occupational Health.
7. Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model of types and levels of human interaction with automation. IEEE Transactions on Systems, Man and Cybernetics, 30(3), 286-297.
8. Endsley, M. R. (2017). From here to autonomy: Lessons learned from human-automation research. Human Factors, 59(1), 5-27.
9. Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2014). Human performance consequences of stages and levels of automation: An integrated meta-analysis. Human Factors, 56(3), 476-488.
10. Layton, C., Smith, P. J., & McCoy, C. E. (1994). Design of a cooperative problem-solving system for en-route flight planning: An empirical evaluation. Human Factors, 36(1), 94-119.
11. Olson, W. A., & Sarter, N. B. (1999). Supporting informed consent in human machine collaboration: The role of conflict type, time pressure, and display design. Proceedings of the Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting (pp. 189-193). Santa Monica, CA: Human Factors and Ergonomics Society.
12. Sarter, N. B., & Schroeder, B. (2001). Supporting decision making and action selection under time pressure and uncertainty: The case of in-flight icing. Human Factors, 43(4), 573-583.
13. Smith, P. J., McCoy, E., Orasanu, J., Denning, R., Van Horn, A., & Billings, C. (1995). Cooperative problem solving in the interactions of airline operations control centers with the national aviation system Columbus, OH: Cognitive Systems Engineering Laboratory, Ohio State University.
14. Endsley, M. R., & Jones, D. G. (2012). Designing for situation awareness: An approach to human-centered design (2nd ed.). London: Taylor & Francis.
15. Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50-80.
16. Centers for Medicare and Medicaid Services. (2020). Appropriate Use Program, from https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Appropriate-Use-Criteria-Program
17. Breznitz, S. (1983). Cry wolf: The psychology of false alarms. Hillsdale, NJ: Lawrence Erlbaum.
18. Sorkin, R. (1989). Why are people turning off our alarms? Human Factors Bulletin, 32(4), 3-4.
19. Carayon, P., Hundt, A. S., Karsh, B.-T., Gurses, A. P., Alvarado, C. J., Smith, M., & Brennan, P. F. (2006). Work system design for patient safety: The SEIPS model. . Quality & Safety in Health Care, 15, i50-i58.
20. Carayon, P., Wetterneck, T. B., Rivera-Rodriguez, A. J., Hundt, A. S., Hoonakker, P., Holden, R., & Gurses, A. P. (2014). Human factors systems approach to healthcare quality and patient safety. Applied Ergonomics, 45(1), 14-25.
21. Meyer, A. N., Giardina, T. D., Spitzmueller, C., Shahid, U., Scott, T. M., & Singh, H. (2020). Patient Perspectives on the Usefulness of an Artificial Intelligence–Assisted Symptom Checker: Cross-Sectional Survey Study. Journal of Medical Internet Research, 22(1), e14679.
22. Sujan, M., White, S., Furniss, D., Habli, I., Grundy, K., Grundy, H., . . . Reynolds, N. (2019). Human factors challenges for the safe use of artificial intelligence in patient care. BMJ Health and Care Informatics.