HTA10

HTA Considerations for Large Language Models in Healthcare

Caoimhe Leonard,[1] Harriet Unsworth,[2] Liesl Gildea,[1] Sheryl Warttig,[1] Margaret Mordin,[3] Caroline Ling[1]

[1] RTI Health Solutions, Manchester, United Kingdom; [2] Digital Cancer Research, CRUK National Biomarker Centre, University of Manchester, Manchester, United Kingdom; [3] RTI Health Solutions, Research Triangle Park, NC, United States

Objectives

- To consider the complexity and requirements for evaluating LLMs
- To suggest updates to the ESF to help HTA bodies and developers of DHTs meet standards to successfully approach HTA for LLMs in healthcare

Background

The large language model (LLM) healthcare landscape is rapidly evolving, which poses challenges for regulators and health technology assessment (HTA) bodies to keep up. There are no current healthcare LLMs with regulatory approval.

The Evidence Standards Framework (ESF), published by the National Institute for Health and Care Excellence (NICE), provides a set of 21 standards that should be met in order for the National Health Service (NHS) to be able to adopt a digital health technology (DHT). The ESF aims to inform purchasing decisions in the NHS and to guide DHT developers in generating evidence for their technologies.

The ESF does not include standards designed to capture risks and benefits specific to healthcare LLMs because healthcare LLMs were not available in 2022.

Healthcare Large Language Models, Generative AI, and Current Capabilities

LLMs are generative artifical intelligence (AI) models trained on extremely large data sets for processing natural language for a range of different purposes. Figure 1 presents LLMs' current capabilities.

However, risks exist in many commercial LLMs. These include hallucination, where the LLM creates realistic sounding but untrue text if it cannot find a correct response, and data security risks, whereby any confidential data or personal information entered into commercial LLMs (such as OpenAI's ChatGPT) is used as training data and may be provided as a response to another query.[1]

Examples of healthcare LLMs[2]:

- **Med-PaLM** provides a range of functions, such as diagnostic assistance, synthesising and communicating information from images and other medical data, and discussing results with clinicians through natural language dialogue.[2]
- **Bloom** is a multilingual LLM with the core functionality being text generation. Bloom is prompted with an initial context and then can complete the text. This type of LLM helps with text, such as clinical notes, patient histories, and scientific articles.[2]
- **LLaMA 3** is a transformer-based LLM that is efficient at understanding and producing human-like texts, creating possibilities for supporting healthcare professionals in the diagnostic process.[2]

Evidence Standards

NICE first published the ESF for DHT evaluations in 2019 and updated it in 2022 to cover the types of AI that were most frequently used in the NHS at that time.[3]

The update in 2022 included:

- Aligning the standards to DHT regulatory requirements
- Specifying evidence requirements relevant for AI and data-driven DHTs that have fixed or adaptive machine learning algorithms
- Making the framework easier to use by removing the evidence requirements of describing 'minimum' and 'gold standard' for each standard
- Outlining a subset of early deployment standards that can be used as a minimum entry point for evidence generation programmes

The standards are designed to capture the following dimensions:

- Design factors (health and care inequalities and bias mitigations)
- Performance (DHTs' performance measuring overtime or real-world evidence claiming the benefits)
- Deployment considerations (end-user consent and their understanding)
- Describing and delivering value, the intended purpose of DHT, and the economic analysis

Scan this QR code to view the full ESF from NICE

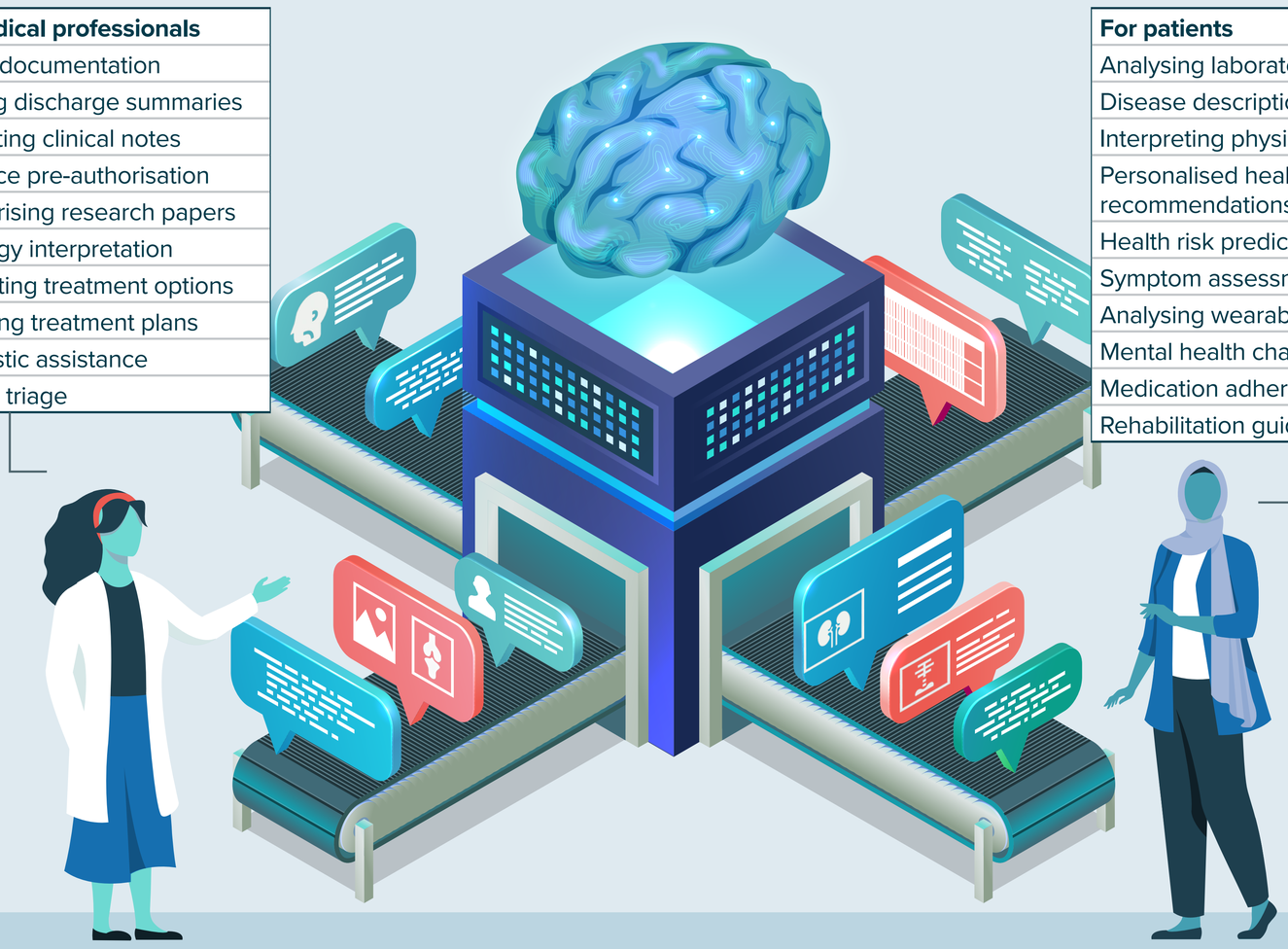
Considerations in Evaluating LLMs

Key aspects have been identified to evaluate healthcare LLMs[4,5]:

- **Measuring effectiveness:** The accuracy, safety, and (for LLMs with a clinical function) clinical effectiveness will need to be evaluated. This may require new approaches to determine the most relevant outcome measures and study designs for LLMs, especially those that report a range of different healthcare functions.
- **Transparency:** Limited transparency and explainability around how a healthcare LLM was trained and how it reaches a decision can reduce clinician and patient trust and weaken clinical evaluation or the decision-making process.
- **Bias and errors in decision-making:** Healthcare LLMs' ability to generate human-like text raises ethical challenges. LLMs are known to be at risk of producing incorrect or biased output, and it can be difficult for end-users to identify these errors. Such errors might be reproduced and amplified as the LLM outputs are used to train future model iterations. If unchecked, this will lead to biases and errors in healthcare decision-making.
- **Continuous validation:** Some healthcare LLMs could use a continuous learning model, whereby all new data are used to continuously retrain and update the LLM, changing its effectiveness/accuracy. This is a challenge to traditional methods of healthcare evaluation, which assume that the healthcare product is static.

Figure 1: Examples of Use Cases of LLMs[6]

For medical professionals
Clinical documentation
Creating discharge summaries
Generating clinical notes
Insurance pre-authorisation
Summarising research papers
Radiology interpretation
Suggesting treatment options
Designing treatment plans
Diagnostic assistance
Medical triage



For patients
Analysing laboratory results
Disease descriptions
Interpreting physician notes
Personalised health recommendations
Health risk prediction
Symptom assessment
Analysing wearables data
Mental health chatbot
Medication adherence
Rehabilitation guidance

Evidence Standards Selected for Update

Standard 1
DHT should comply with relevant safety and quality standards

Standard 4
Consider health and care inequalities and bias mitigation

Standard 15
Show real-world evidence that the claimed benefits can be realised in practice

Standard 16
The company and evaluator should agree on a plan for measuring usage and changes in the DHT's performance over time

Standard 19
Ensure transparency about requirements for deployment

Standard 20
Describe strategies for communication, consent, and training processes to allow the DHT to be understood

https://www.nice.org.uk/corporate/ecd7/chapter/section-c-evidence-standards-tables

Recommendations for Successful LLM Implementation in Healthcare Systems

The ESF was designed before the recent deluge of healthcare LLMs reached the market, and so it was not designed to capture the challenges specific to LLMs. Although the ESF is not the basis of formal HTA, it can complement NICE evaluations and promote consistency.

Future updates to the ESF will be informed by the EU Horizons projects EDiHTA and ASSESS DHT. The following updates should be considered to successfully implement LLMs into healthcare systems:

- **Ethical concerns,** such as bias, should be addressed in standard 4 of the ESF, which covers inclusion and inequality but lacks ethical challenges around generation of human-like text. Standard 1 of the ESF covers safety and quality, including the data protection act and other quality standards; however, the inclusion of specificity to LLMs would help generate a strong evaluation of healthcare LLMs.
- The current standard 15 includes a recommendation to show **clinical validation** results of in-practice evaluations. This could be expanded to include an assessment of the LLMs' outputs to measure risk of a hallucination or incorrect information.
- We recommend updating standard 16 for evaluating LLMs in a **continuous validation** to include an assessment of how the end user's relationship with technology changes and whether trust in the LLM is sufficient to maximise the potential benefits of these tools.
- Update standards 19 and 20, detailing **transparency and explainability** to introduce information on handling DHTs tasked with natural language processing. This will reduce uncertainties, increase trustworthiness, and help evaluate LLMs and end-user understanding.

References

1. Clusmann J, et al. Comm Med. 2023,3:141.
2. Nassiri K, Akhloufi MA. BioMedInformatics 2024;4:1097-143.
3. NICE. 2022. https://www.nice.org.uk/corporate/ecd7
4. Reddy S. Inform Med Unlocked. 2023;41:2352-9148.
5. Ong JCL, et al. NEJM AI. 2024;1(7):AIra2400038.
6. Meskó B, Topol EJ. NPJ Digit Med. 2023;6(1):120.

Contact Information

Caoimhe Leonard, MSc
Research Associate, Value and Access
RTI Health Solutions
Phone: +44 (0) 161 447 6037
Email: cleonard@rti.org

Scan this QR code to download poster

The power of **knowledge.**
The value of **understanding.**

Presented at: ISPOR Europe 2024; 17-20 November 2024; Barcelona, Spain