Generative AI for Health Technology Assessment: Opportunities, Challenges, and Policy Considerations

Rachael L. Fleurence, PhD, MSc
Office of the Director
National Institute of Biomedical Imaging and Bioengineering
National Institutes of Health
Bethesda, MD
United-States

Jiang Bian, PhD
Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida
Biomedical Informatics, Clinical and Translational Science Institute, University of Florida
Office of Data Science and Research Implementation, University of Florida Health
Gainesville, FL
United-States

Xiaoyan Wang, PhD
Tulane University School of Public Health and Tropical Medicine
New Orleans, LA
Intelligent Medical Objects
Rosemont, IL
United-States

Hua Xu, PhD
Department of Biomedical Informatics and Data Science, School of Medicine, Yale University,
New Haven, CT
United-States

Dalia Dawoud, PhD
National Institute for Health and Care Excellence
London
United Kingdom

Cairo University, Faculty of Pharmacy
Cairo
Egypt

Tala Fakhouri, PhD MP
Office of Medical Policy
Center for Drug Evaluation and Research
United States Food and Drug Administration
Silver Spring, MD
United-States

Mitch Higashi, PhD

ISPOR - The Professional Society for Health Economics and Outcomes Research
Lawrenceville, NJ
United-States

Jagpreet Chhatwal, PhD
Institute for Technology Assessment, Massachusetts General Hospital, Harvard Medical School
Center for Health Decision Science, Harvard University
Boston, MA
United-States

Abstract

This review introduces the transformative potential of generative Artificial Intelligence (AI) and foundation models, including large language models (LLMs), for health technology assessment (HTA). We explore their applications in four critical areas, evidence synthesis, evidence generation, clinical trials and economic modeling: (1) Evidence synthesis: Generative AI has the potential to assist in automating literature reviews and meta-analyses by proposing search terms, screening abstracts, and extracting data with notable accuracy; (2) Evidence generation: These models can potentially facilitate automating the process and analyze the increasingly available large collections of real-world data (RWD), including unstructured clinical notes and imaging, enhancing the speed and quality of real-world evidence (RWE) generation; (3) Clinical trials: Generative AI can be used to optimize trial design, improve patient matching, and manage trial data more efficiently; and (4) Economic modeling: Generative AI can also aid in the development of health economic models, from conceptualization to validation, thus streamlining the overall HTA process. Despite their promise, these technologies, while rapidly improving, are still nascent and continued careful evaluation in their applications to HTA is required. To ensure their responsible use and implementation, both developers and users of research incorporating these tools, should familiarize themselves with their current limitations, including the issues related to scientific validity, risk of bias, and consider equity and ethical implications. We also surveyed the current policy landscape and provide suggestions for HTA agencies on responsibly integrating generative AI into their workflows, emphasizing the importance of human oversight and the fast-evolving nature of these tools.

Introduction

Artificial intelligence (AI) has been defined as "the science and engineering of making intelligent machines" [1]. Advancements in AI and especially machine learning (ML) have been rapidly progressing for several decades [2,3]. However, the recent emergence of generative AI technologies such as large language models (LLMs) has ushered in a transformative era across numerous fields, including science, medicine, and other areas of human activity [4-6] (**Box 1**) (**Figure 1**).

Generative AI employs sophisticated ML models, particularly a class of deep neural networks, that can generate language, images, and code in response to free text prompts provided by users. These models are trained using self-supervised learning techniques on vast amounts of existing data, allowing them to identify patterns and relationships in the underlying data autonomously. Their training requires substantial computational power, often relying on specialized hardware, especially graphics processing units (GPUs) [7].

In the field of Health Economics and Outcomes Research (HEOR), generative AI has the potential to transform evidence generation methods used in health technology assessments (HTA). Despite this promise, the exploration of opportunities, limitations, and risks associated with developing, evaluating, and deploying these models for supporting different approaches to evidence generation remains sparse, given the recency of these new tools. This paper seeks to fill this gap by offering a review of the landscape of generative AI's applications in methodological areas relevant to HTA development. The immediate audience is HEOR professionals, but we anticipate that it will also be useful to other stakeholders, particularly HTA evidence developers and users, regulators, and manufacturers.

The manuscript begins with a concise history of AI and overview of generative AI. Next, we examine the opportunities and challenges of utilizing generative AI to support significant aspects of HEOR. This includes (1) evidence synthesis methods, (2) evidence generation methods including Real-World Evidence (RWE), (3) clinical trials and (4) health economic modelling. We continue with a description of the current limitations and risks associated with generative AI tools. Next, we provide a summary overview of the current policy landscape on AI. We conclude with the results of a brief survey of HTA agencies gauging their current use of tools with generative AI and provide suggestions to HTA agencies on preparing for a more widespread adoption of generative AI tools.

A brief history of AI and generative AI

AI has evolved significantly since its conception, tracing a path of remarkable technological advancements. Introduced by Alan Turing in 1950, the Turing Test challenged the ability to distinguish between responses from a machine and a human during a natural language conversation, setting a foundational criterion for machine intelligence [8]. In the following decade, significant progress was made with the development of symbolic methods to process natural language, such as the ELIZA chatbot in the 1960s [9]. In the medical field during the 1970s, expert systems like Stanford University's MYCIN demonstrated capabilities in reasoning under uncertainty and providing decision rationales, using rule-based AI [10].

The 1990s ushered in the era of ML algorithms with the development of techniques like Support Vector Machines (SVM) and Random Forests. These methods excelled in tasks across various domains by effectively handling pattern recognition and decision-making processes. This period marked a shift from rule-based to data-driven AI, setting the stage for the next big leap. The introduction of deep learning in the 2000s further revolutionized AI capabilities [11]. Deep learning

models based on neural networks, enabled a wide range of complex applications from image recognition to natural language processing (NLP). It also led to breakthroughs such as IBM Watson and AlphaGo—systems that famously outperformed humans in Jeopardy and Go, respectively [12,13]. While IBM Watson was not eventually successful in healthcare applications, AlphaGo led to the development of AlphaFold, which revolutionized the field of structural biology by solving the protein folding problem [14,15].

The emergence of generative AI technologies, especially the LLMs from OpenAI's GPT series, and the debut of ChatGPT in November 2022, marked a paradigm shift in the field of AI. Generative AI focuses on generating new content (e.g., text, images, videos) by learning from large existing data and it differs substantially from earlier AI technologies in several key aspects. First, LLMs employ the innovative transformer architecture, which features self-attention mechanisms to evaluate the relevance of each word in a sentence, regardless of their position. This enhances the model's grasp of context and linguistic subtleties [16]. Pre-trained on vast text corpora, LLMs utilize billions of parameters to excel across a variety of tasks, such as summarization, translation, question answering, and code generation. These models are often called "foundation models" as the fundamental basis for a wide variety of downstream tasks. Since the development of ChatGPT, numerous LLMs have been created, consistently setting new benchmarks in performance across many tasks [7]. Additionally, these generative AI technologies offer user-friendly interfaces that allow for direct interaction through natural language text or speech, significantly reducing the need for structured data preparation typical of earlier ML models like SVMs. LLMs can also quickly adapt to new tasks through natural language instructions and in-context learning.

In summary, generative AI technologies like LLMs represent a fundamental shift in how machines can understand and interact with human language, offering a breadth of capabilities and flexibility that were not possible with earlier AI models. Many applications of LLMs to different data modalities (e.g., text, images, and videos) have been investigated across different industrial sectors including healthcare and life science [17,18].

Applications of generative AI in literature reviews and evidence synthesis

This section identifies applications of generative AI, including the use of foundation models such as LLMs, in evidence synthesis studies such as systematic literature reviews (SLRs) and meta-analyses. LLMs can assist in generating a search strategy by proposing Mesh terms and keywords to input in biomedical search engines such as PubMed [19]. Nevertheless, "hallucinations" can occur, where LLM-generated citations can be fabricated, manually verifying the terms proposed by the models is still necessary [19,20]. Search engines are also improving their search returns with the integration of AI supported tools [21]. Researchers have evaluated the capacities of foundation models to automate abstract and full text screening. When screening full text literature using highly reliable prompts, GPT-4's has been shown to have comparable performance to humans [22]. For example, Guo et al. assessed the performance of OpenAI GPT in automating screening and found a high level of agreement between the model-generated results and a human reviewer [23]. They also found that the model was able to explain its reasoning for excluding papers and correct its initial decisions [23]. Robinson et al. have also evaluated LLMs' capacity to provide exclusion reasoning for abstracts and full texts [24]. Hasan et al compared the agreement between GPT-4 and human reviewers in assessing the risk of bias using the Cochrane Collaboration's risk of bias tool. Their case study demonstrated a fair agreement between the LLM and the human reviewer [25]. Several studies have tested the data extraction capacities of

LLMs. Reason et al. found that GPT-4 >99% accuracy in replicating the data extraction in 4 network meta-analyses [26]. Gartlehner et al. assessed the performance of CLAUDE-2 in extracting data elements from published studies compared to human extraction [27] and found that CLAUDE-2 achieved an overall accuracy of 96.3%. Schopow et al. found a high degree of concordance between ChatGPT (GPT-3.5 and 4.0) and human researchers on the data extraction task [28]. Nevertheless, using GPT-4, another study only found moderate performance in assisting with data abstraction [22]. In addition to data extraction, the foundation model can also be directed to generate the code to conduct meta-analyses. Reason et al. found that the LLM used was able to generate error-free code in R for network meta-analyses [26]. However, Qureshi et al. found errors in code generated in an earlier GPT model, GPT-3.5 [19]. Finally, report writing has also been evaluated with researchers finding that the LLM is able to generate reasonable drafts [26].

In summary, these early applications show that there is promise in using foundation models to support a range of tasks required in SLRs but this rapid overview indicates that LLMs still have limitations in automating SLR tasks and human verification is necessary. Comparisons with human conducted tasks show reasonable accuracy in many case studies but not all, emphasizing the need for continued human validation at all steps [20]. It is expected that both user expertise with foundation models, and the actual performance of foundation models themselves, will improve rapidly thereby likely augment the conduct SLRs in the near future.

Applications of generative AI to evidence generation

Increasingly, RWE is being used by a wide range of stakeholders in the HTA process [29]. The potential benefits of using generative AI in RWE are profound, including increased efficiency and speed (e.g., faster data processing and analysis, and reduction in manual efforts), enhanced accuracy and consistency (e.g., minimizing human errors and standardizing the evidence generation processes), and the ability to combine multimodal data sources (e.g., structured clinical variables, unstructured clinical texts, imaging, and omics), enabling more holistic assessments.

Generative AI has the potential to improve the efficiency and quality of RWE generation. It is well-known that unstructured notes in EHRs, such as radiology reports and physician notes, are important for RWE generation [30] and the use of LLMs can make information extraction from notes easier and more efficient [31]. This also leads to the development of domain-specific LLMs like GatorTron, GatorTronGPT and Me LLaMA, which are trained using large clinical texts [32-34]. Additionally, generative AI is increasingly being customized for images and other formats of real-world data to generate evidence [35]. However, one study also found that a range of foundation models had less than 50% accuracy in mapping descriptive text to the correct ICD and CPT codes [36]. Like other areas, this field is still nascent although likely to improve rapidly. AI, including ML methods, has already been widely used in predicting patient health outcomes and modeling disease progression, especially with RWD, both of which are important components of the evidence generation process; recent advancements in generative AI have made it possible to be more accurate and efficient in creating synthetic or virtual patient cohorts and simulate treatment outcomes [37]. This has potential applications such as creating synthetic or external control arms to optimize clinical trials, accelerating the evidence generation process [37]. However, as discussed in the limitations section below, many of these areas are still nascent and continue to need human oversight and validation within each step of the process.

Applications of generative AI to clinical trials

Clinical trials, the gold standard for measuring clinical efficacy and safety of medical interventions, are another important component of evidence generation. Applications of generative AI to improve and accelerate the planning and conduct of clinical trials are still nascent and are likely to expand [37-39]. First, in the area of clinical trial design, exploratory work is underway to leverage the summarizing and extracting capabilities of LLMs to optimize clinical trial protocols. This might be done by efficiently summarizing previous trials in the same disease area, reviewing their designs and outcomes [38], or by extracting the eligibility criteria of previous clinical trials from free text [40]. LLMs might also be used to aid the design of clinical trial eligibility criteria for patients. For example, AutoTrial generates eligibility criteria for clinical trials using natural language to prompt LLMs [41]. Second, LLMs are being explored quite extensively to improve and accelerate patient-matching for clinical trials. Several methods using LLMs have been proposed [42-44], some offering innovative approaches to increase patient confidentiality in the process [45]. One approach uses ChatGPT to extract patient information from unstructured clinical notes and generate queries to search for potentially eligible clinical trials [46]. Other approaches include InstructGPT to assist physicians in determining the eligibility of patients for clinical trials based on the patient's summarized medical profile [47] and TrialGPT to match patients to trials [48]. Future areas also include approaches to predict which patients might be more likely to complete a trial, and approaches to increase patient retention in trials, for example through the provision of specialized trained chatbots to answer patient questions [38]. Finally, foundation models are likely to be able to assist in clinical trial data management, for example by extracting data from clinical trial documents in structured format, such as Population, Intervention, Comparator, and Outcome (PICO) [49], matching images to text, or standardizing unstructured data but these remain exploratory approaches at this time. In a comprehensive review of ML techniques, Weissler et al. provide a helpful list of aspects of clinical trials that will likely be improved by ML tools which can be generalized to generative AI tools [50]. Overall, the integration of generative AI has the potential to significantly enhances clinical trial development and execution, including design, optimization, emulation, and recruitment, signaling a potentially transformative shift in the landscape of clinical trials.

Applications of generative AI to health economic modeling

Generative AI has the potential to support the phases of economic model development, including model conceptualization, parameterization, model implementation, and evaluation and validation of model results [51-53]. Chhatwal et al. explored the use of foundation models—GPT-4 and Bing Chat—in conceptualization of a Markov model for hepatitis C and parameterization of the model [54]. They found significant variability in how these models conceptualized disease progression. Despite GPT-4 performing better than Bing Chat, the study highlighted that foundation models alone are insufficient for robust conceptual model development. This underscores the likely need for specialized models that are specifically trained to develop more consistent and reliable health economic models. Reason et al. demonstrated the capability of GPT-4 to recreate published three-state partition survival models for non-small cell lung cancer and renal cell carcinoma [55]. While the model construction was replicated with high accuracy, this approach required human intervention to provide contextual information on top of information specifying the model assumptions, methods and parameter values. This suggests that although generative AI can create a model code, significant human expertise is still necessary to guide the model conceptualization and parameterization. Generative AI could play an important role in structural uncertainty

analysis, a typically resource-intensive process [52]. By automating this aspect, generative AI can help in identifying and mitigating potential weaknesses in model structure more efficiently.

In summary, generative AI has the potential to support all stages of health economic model development, from conceptualization to validation. This transformation could lead to more efficient model development, reducing time and resource expenditure. By automating different phases of the modeling, generative AI could lower the barriers to conducting comprehensive health economic evaluations, making these tools more accessible to a broader range of healthcare decision-makers. Future research should explore the accuracy and reliability of fully automating all phases of health economic modeling. As AI becomes more integrated into health economic modeling, it will be crucial to establish rigorous standards to ensure the validity and reliability of AI-generated models. Future collaborations should focus on establishing guidelines and frameworks for the review of such models.

Limitations of generative AI in HTA applications

While generative AI offers promising applications in areas relevant to HTA, caution should be exercised when evaluating studies employing these technologies [4]. Several frameworks have been developed to help guide the responsible use of AI in health care, many of which remain relevant to generative AI [56-58]. Here, we briefly review and focus on three different categories of limitations of generative AI and foundation models: 1) Scientific validity and reliability, 2) Bias, equity and fairness and 3) Regulatory and ethical concerns.

Scientific validity and reliability

Scientific validity and reliability are critical aspects of research. Because LLMs are often trained on large corpora of publicly available information from the web, errors may be introduced especially in domain specific areas like health [4]. A well noted limitation has been the generation of "hallucinations" where the model generates text that seems plausible, but it is in fact fabricated [20,59]. This phenomenon occurs because models generate context based on learned statistical associations between words and sentences, not from an understanding of the knowledge in the corpora. Several strategies might be employed to improve the factual correctness of model outputs. First, improving the guidance used in the prompts provided to the models, a process known as prompt engineering, has been shown to improve the accuracy of the results. For example, chain-of-thought prompting is a technique in which a model is guided to generate step-by-step explanation or its reasoning process while arriving at an answer [60]. Few-shot learning, a method where the model learns and generalizes from a very limited amount of data and examples, is another approach to improve model performance [61]. Second, the use of retrieval augmented generation (RAG) allows the model to retrieve relevant information from external sources [21]. This has been shown reduce the incidence of hallucinations and factually incorrect output [62]. Third, foundation models can be fine-tuned with domain-specific data and knowledge, thus improving their performance on answering domain-specific questions [32,63,64]. In science, findings should also be reproducible. Challenges to reproducibility with research using generative AI, include the randomness inherent in the training of foundation models and in the generation of the model parameters. Researchers are also proposing interesting approaches to support the reproducibility of studies and the accuracy of results. For example, Hasan et al. propose a framework for integrating foundation models into SLRs [25]. Reason et al. run their models over 20 times to compare the results from different runs [26]. Open sharing of data, code and results as well as the adoption of standards for reporting and transparency will likely help

improve reproducibility [65,66]. Finally, transparency in this area is challenging because of the 'black box' nature of close-source foundation models (e.g. GPT, Gemini), meaning it is unclear how models generate answers from the input queries and the model algorithms. Strategies proposed to improve transparency have included requirements for the outputs to cite which part of the dataset contributed to the answer, and 'explainable' AI [4]. To achieve these goals and address the challenges above, the involvement of all stakeholders, with consideration of human-AI teaming, including human oversight, is critically needed in using these models [67].

Bias, Equity and Fairness
Foundation models can propagate or amplify biases introduced at many steps from model inception to deployment, resulting in systematic differences that may exacerbate inequities and cause harms to individuals and communities [68,69]. First, systemic bias (or institutional or historical bias) may occur in the use of historical data to train foundation models, where marginalized groups may be underrepresented in datasets because of barriers to access to resources, perpetuating existing systemic biases in society [70]. Second, computational and statistical biases may stem from errors that occur when the sample is not representative of the population and the algorithm produces a result that differs from the true underlying estimate [71,72]. Bias may also occur when specific population groups are excluded from data collection, training, testing or subsequent analyses [68]. Several strategies have been proposed to assess and mitigate the risk of bias in foundation models. Surveys on this topic have been published [73-75]. One review of methods for addressing AI fairness and bias in biomedicine identified distributional and algorithmic approaches for addressing bias [73]. Distributional approaches include strategies such as data augmentation, data perturbation and data reweighting, such as the generation of synthetic datasets to ensure balanced representation of all demographic groups [76]. Federated learning may offer some potential to address biases, where models are trained on multi-institutional data without the need for data sharing, with the idea that access to more diverse populations in many different health systems may mitigate biases inherent to localized datasets [68]. Algorithmic approaches change the algorithm itself and include approaches such as adversarial learning and loss-based methods, where the loss function is adjusted to penalize biased predictions [73].

Regulatory and Ethical Considerations
Regulatory frameworks focused on generative AI in biomedical research are under development in many countries. However, established data privacy and confidentiality laws such as the Health Insurance Portability and Accountability Act (HIPAA) in the United-States and General Data Protection Regulation (GDPR) in the European Union remain applicable. However, foundation models require a large amount of training data which could contain Protected Health Information (PHI) where absolute de-identification is not attainable, and re-identification risks remain non-zero [77,78]. Further foundation models have the ability to memorize the data that they were trained on, posing a risk of reproducing PHI that they memorized during training [79]. Enhancements such as synthetic data that replicates real patient data without personal identifiers [80] and enabling computations on encrypted data [81], are being explored to augment privacy protections. The issue of informed consent for AI driven research is becoming increasingly critical and is underscored by recent legal disputes over the use of patient EHR data to train LLMs [82]. Ethical concerns will arise from the concerns around scientific validity, bias and fairness. Inaccurate information can lead to patient harm, for example if used to support an HTA submission that erroneously exclude underserved populations. In addition, emergent capabilities of LLMs may raise significant

ethical challenges, particularly when the model's accurate performance cannot be transparently explained. For example, Giyocha et al. found that the AI model could identify a patient's self-reported race from imaging alone, raising ethical issues that will need to be thoughtfully addressed [83]. In summary, given the early stages of the development and advances of generative AI, the themes discussed above will continue to be important and the object of warranted attention from many stakeholders within society.

Policy landscape

A rapid review of the policy landscape on the use of generative AI, including foundation models, in healthcare was conducted. **Table 1** presents legal and policy documents related to the use of AI, generative AI, and foundation models from governments, regulatory bodies, and multi-stakeholder groups. Many governments have issued laws or policy guidance with respect to the use of AI in their countries that are not specific to healthcare or health research. The EU has issued the most comprehensive "horizontal" regulatory framework to date, the EU AI Act, which applies to all applications of AI across sectors [84]. The regulation is binding on its member states and offers a risk-based approach to the use of AI technologies, with higher risk technologies associated with the most stringent requirements. Other countries have adopted "vertical" regulation, which applies to specific applications of AI or a specific sector. The UK has opted for a more flexible regulatory approach, with no centralized regulatory body; instead, different regulators will oversee AI applications in different sectors [85,86]. In Japan, the government has encouraged voluntary governance by companies with sector-specific guidelines rather than AI-specific laws [87]. In the US, there is no comprehensive federal AI regulation framework, but the Biden administration has issued an executive order on safe, secure and trustworthy AI, tasking multiple agencies to develop guidelines for AI systems and address the risk associated with federal government use of AI [88,89]. International organizations have also issued guidance or statements on AI, including the Organization for Economic Co-operation and Development (OECD), the United Nations (UN) and United Nations Educational, Scientific and Cultural Organization (UNESCO) [90-92]. In the health care sector, the World Health Organization (WHO) has provided guidance to member states on large multimodal models (LMMs), describing the benefits and risks of deploying LMMs in health care [93,94].

In the regulatory space, the European Medicines Agency (EMA) has issued a reflection paper on the use of AI/ML in the medicinal product life cycle, including product development, authorization, and post-authorization, emphasizing the need to meet all existing legal requirements, employ a risk-based approach, and promote AI trustworthiness [95].

The FDA has issued several discussion documents related to AI, specifically on the use of AI/ML-based software as a medical device [96,97] and more generally on AI use across the drug development process, including drug discovery, nonclinical phase of drug development, the clinical phase including in clinical trials, post-marketing surveillance and in manufacturing [98-100].

In the field of HTA, we reviewed the websites of the main HTA bodies and did not identify statements or guidance on the use of AI in HTA submissions at the time of writing (May 2024). The ISPOR HTA Roundtable conducted an informal survey with 18 members from North America, Latin America, Australasia and Europe, in April 2024 to gauge the level of internal use of generative AI in HTA agencies (They were not queried about the use of AI in submissions). Several agencies mentioned that they were still in an exploratory phase with generative AI, seeking to better understand how it may support their work and be used by manufacturers in their

submissions. Some groups were not using AI at this time and one agency blocks the use of AI tools such as Chat-GPT at this time. Several agencies are evaluating software offered by private companies in the space of Systematic Literature Reviews (5 agencies) and one agency described experimenting with the use of generative AI in economic modelling. Finally, some agencies mentioned internal use of AI tools for translation and for general education purposes. The learnings from these explorations of AI tools will likely inform the methods and processes guidance HTA agencies will need to develop to guide companies in using generative AI in HTA submissions [101]. In Box 2, we provide suggestions for HTA agencies and policy makers for establishing procedures and processes to improve the use of generative AI in HTA submissions (**Box 2**). Multi-stakeholder groups are also working on frameworks for the responsible and trustworthy use of AI, including for example, the National Academies of Medicine, the US National Institute of Standards and Technology (NIST), the Coalition for Health AI and the European Commission's HTx H2020 (Next Generation Health Technology Assessment) project [56-58,72,101]. Finally, we note that scientific and medical journals are in the process of developing their policies with respect to the use of AI tools in publications submitted to their journals [102]. Of note, the new NEJM AI journal encourages the use of LLMs in submissions with the assumption that they will enhance the quality of research and support its democratization [103].


Conclusion
The field of generative AI provides potentially transformative tools to augment and support the efficient generation of evidence to support HTA. In this article, we have provided a brief overview of the development of generative AI and reviewed applications relevant to HTA including: evidence synthesis, evidence retrieval and generation, clinical trials and economic modeling. Despite their promise, it is important to acknowledge that these technologies, while rapidly improving, are still nascent and continued careful evaluation in their application to HTA is required. To ensure their responsible use and implementation, both developers and users of research incorporating these tools, should familiarize themselves with their current limitations, including the issues related to scientific validity, risk of bias, equity implications and ethical considerations. This is a transformative and exciting time in scientific and medical research, but scientific research is a complex endeavor that impacts human health and will continue to require careful validation and human oversight for the foreseeable future.

Glossary

- Artificial intelligence (AI): a broad field of computer science that aims to create intelligent machines capable of performing tasks typically requiring human intelligence.

- Deep Learning: a subset of machine learning algorithms that uses multilayered neural networks, called deep neural networks. Those algorithms are the core behind the majority of advanced AI models.

- Foundation model: a large scale pretrained models that serve a variety of purposes. These models are trained on broad data at scale and can adapt to a wide range of tasks and domains with further fine-tuning.

- Generative AI: AI systems capable of generating text, images, or other content based on input data, often creating new and original outputs.

- Generative Pre-trained Transformer (GPT): a specific series of LLMs created by OpenAI based on the Transformer architecture, which is particularly well-suited for generating human-like text.

- Large Language Model: a specific type of foundation model trained on massive text data that can recognize, summarize, translate, predict, and generate text and other content based on knowledge gained from massive datasets.

- Machine learning (ML): a field of study within AI that focuses on developing algorithms that can learn from data without being explicitly programmed.

- Multimodal AI: an AI model that simultaneously integrates diverse data formats provided as training and prompt inputs, including images, text, bio-signals, -omics data and more.

- Prompt: the input given to an AI system, consisting of text or parameters that guide the AI to generate text, images, or other outputs in response.

- Prompt engineering: creating and adapting prompts (input) to instruct AI models to generate specific output.

Figure 1: Relationship between AI, Gen AI, Foundation Models and LLMs
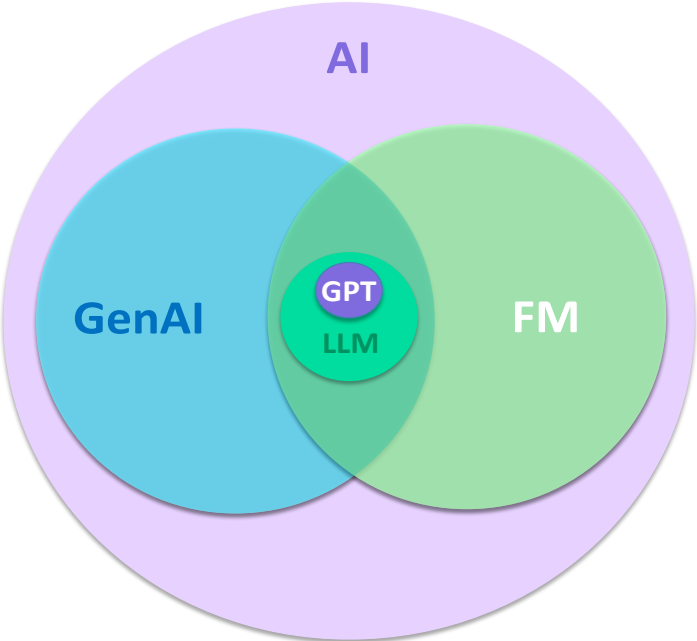
AI=Artificial Intelligence; FM= foundation model; Gen AI = generative artificial intelligence;
GPT= Generative Pre-Trained Transformer; LLM= large language model.

Table 1: Laws, Policies and Guidance on the use of Generative AI from governments, regulatory bodies, and multi-stakeholder groups

Entity	Availability of guidance on the use of AI and LLMs
Governments	
Canada	- The Artificial Intelligence and Data Act (AIDA) – Companion document (Canada, 2022)
European Union	- EU AI ACT (EU, 2024)
G7	- Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems (G7, 2023)
Japan	- Japan's Approach to AI Regulation and Its Impact on the 2023 G7 Presidency (Habuka, 2023)
United-Kingdom	- United-Kingdom, A pro-innovation approach to AI regulation: government response (UK Consultation, 2024) - United-Kingdom, A pro-innovation approach to AI regulation (UK, 2023)
United-States	- Executive Order of the President on safe, secure and trustworthy AI (EOP, 2024) - Blueprint for an AI Bill of Rights (EOP, 2022)
International organizations	
Organization for Economic Cooperation and Development (OECD)	- Recommendations from the council on artificial intelligence (OECD, 2023) - OECD principles overview (OECD, 2024)
United Nations	- Interim Report: Governing AI for Humanity (UN, 2024)
UNESCO	- Ethics of Artificial Intelligence (UNESCO, 2024)
World Health Organization (WHO)	- Regulatory considerations on artificial intelligence for health (WHO, 2023) - Ethics and governance of artificial intelligence for health, guidance on large multi-modal models (WHO, 2024)
Regulatory Agencies	
EMA: European Medicines	- Reflection paper on the use of Artificial Intelligence (AI) in the medicinal product lifecycle (EMA, 2023)

Agency (European Union)	
FDA: US Food and Drug Administration (United-States)	- AI and Medical products: How CBER, CDER and CDRH and OCP are working together (FDA, 2024) - Using AI/ML in the development of drug and biological products (FDA, 2023) - Artificial Intelligence in Drug Manufacturing (FDA, 2023) - Predetermined Change Control Plans for Machine Learning-Enabled Medical Devices: Guiding Principles (FDA, 2023) - AI/ML SAMD Action Plan (FDA, 2021) - Good Machine Learning Practices for Medical Device Development: guiding principles (FDA, 2021)
Medicines and Healthcare Products Regulatory Agency (MHRA) (UK)	- Impact of AI on the regulation of medical products: Implementing the AI White Paper principles (MHRA, 2024)
Multi-stakeholder groups	
Coalition for Health Artificial Intelligence (CHAI)	- Blueprint for trustworthy AI implementation guidance and assurance for healthcare (CHAI, 2023)
HTx H2020 (European Countries)	- Recommendations to overcome barriers to the use of artificial intelligence-driven evidence in health technology assessment (Zemplenyi, 2023)
NIST (United-States)	- Artificial Intelligence Risk Management Framework (NIST, 2023) - Towards a Standard for Identifying and Managing Bias in Artificial Intelligence (NIST, 2022)
National Academy of Medicine (United-States)	- AI in health, healthcare and biomedical science: an AI code of conduct (Adams et al. 2024)

AI= Artificial Intelligence; AI/ML= Artificial Intelligence/Machine Learning; CBER= Center for Biologics Evaluation and Research; CDER = Center for Drug Evaluation and Research; CDRH=

Center for Devices and Radiological Health; OCP= Office of Combination Products; SAMD= software as a medical device.

Box 1. Key Definitions*

Generative AI (Gen AI)

AI systems capable of generating text, images, or other content based on input data, often creating new and original outputs.

- Example: DALL-E generating images from text descriptions.

Foundation Models

Large scale pretrained models that serve a variety of purposes. These models are trained on broad data at scale and can adapt to a wide range of tasks and domains with further fine-tuning. While LLMs are a subset of foundation models focusing specifically on language, foundation models can encompass other modalities, such as vision and multimodal tasks.

- Examples: GPT-4, Gemini, BERT, CLIP (Contrastive Language-Image Pretraining), LLama,

Large Language Models (LLMs)

A specific type of foundation model trained on massive text data that can recognize, summarize, translate, predict, and generate text and other content based on knowledge gained from massive datasets.

- Examples: GPT-4, Gemini, BERT, LLama, GatorTron.

***First draft of definitions developed using Chat-GPT4o, May 2024**

Box 2 . Suggestions to HTA Agencies and Policy Makers on the use of Large Language Models (LLMs)

- Develop guidance for assessing the appropriate use of LLMs for health technology assessment or endorse existing guidance.

- Provide examples of what constitutes acceptable use of LLMs in health technology assessments, including through use-cases and "non-use cases", i.e areas where it is not recommended to use LLMs in HTAs.

- Collaborate with other HTA agencies and regulatory agencies to harmonize standards, processes, and reporting requirements for the use of LLMs for health technology assessment.

- Develop guidance on and/or checklists for reporting the use of LLMs, including providing checks on validation, reproducibility, bias.

- Assess whether LLM qualification or accreditation schemes, by vendor or dataset, would improve the efficiency and quality of submissions to HTA agencies.

- Engage in multi-stakeholder collaborations to support shared learning around the use of LLMs for HTA, develop best practices, and include the patient perspective.

- Ensure that health equity considerations are appropriately included in the guidance around the use of LLMs in HTA, such as ensuring that populations that have been historically marginalized are represented.

- Invest in training and preparing the HTA workforce and partners (e.g., academic collaborators, committee members) in the area of LLMs. Involve relevant experts where possible.

References

1.   McCarthy J. What is Artificial Intelligence? Accessed 21 May 2024, http://jmc.stanford.edu/artificial-intelligence/what-is-ai/index.html

2.   Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med*. Jan 2022;28(1):31-38. doi:10.1038/s41591-021-01614-0

3.   Howell MD, Corrado GS, DeSalvo KB. Three Epochs of Artificial Intelligence in Health Care. *JAMA*. Jan 16 2024;331(3):242-244. doi:10.1001/jama.2023.25057

4.   Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. Aug 2023;29(8):1930-1940. doi:10.1038/s41591-023-02448-8

5.   Telenti A, Auli M, Hie BL, Maher C, Saria S, Ioannidis JPA. Large language models for science and medicine. *Eur J Clin Invest*. Feb 21 2024:e14183. doi:10.1111/eci.14183

6.   The Economist. Could AI transform life in developing countries? The Economist: The Economist; 2024.

7.   Zhao WX, Zhou K, Li J, et al. A survey of large language models. *arXiv preprint arXiv:230318223*. 2023;

8.   Turing AM. *Computing machinery and intelligence*. Springer; 2009.

9.   Weizenbaum J. ELIZA—a computer program for the study of natural language communication between man and machine. *Commun ACM*. 1966;9(1):36–45. doi:10.1145/365153.365168

10.   Shortliffe E. MYCIN: A Rule-Based Computer Program For Advising Physicians Regarding Antimicrobial Therapy Selection. 01/01 1976:739-739. doi:10.1145/1408800.1408906

11.   Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. *Neural Comput*. Jul 2006;18(7):1527-54. doi:10.1162/neco.2006.18.7.1527

12.   Ferrucci DA. Introduction to "This is Watson". *IBM Journal of Research and Development*. 05/01 2012;56:1:1-1:15. doi:10.1147/JRD.2012.2184356

13.   Silver D, Huang A, Maddison CJ, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*. 2016/01/01 2016;529(7587):484-489. doi:10.1038/nature16961

14.   Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. Aug 2021;596(7873):583-589. doi:10.1038/s41586-021-03819-2

15.   Economist T. Remarkable progress has been made in understanding the folding of proteins. *The Economist*. July 30 2021 2021;

16.   Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Advances in neural information processing systems*. 2017;30

17.   Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. *Nature*. Apr 2023;616(7956):259-265. doi:10.1038/s41586-023-05881-4

18.   The Economist. AIs will make health care safer and better. *The Economist*. March 27 2024 2024;

19.   Qureshi R, Shaughnessy D, Gill KAR, Robinson KA, Li T, Agai E. Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation? *Syst Rev*. Apr 29 2023;12(1):72. doi:10.1186/s13643-023-02243-z

20.   Jin Q, Leaman R, Lu Z. Retrieve, Summarize, and Verify: How Will ChatGPT Affect Information Seeking from the Medical Literature? *J Am Soc Nephrol*. Aug 1 2023;34(8):1302-1304. doi:10.1681/ASN.0000000000000166

21.   Wei CH, Allot A, Lai PT, et al. PubTator 3.0: an AI-powered literature resource for unlocking biomedical knowledge. *Nucleic Acids Res*. Apr 4 2024;doi:10.1093/nar/gkae235

22.   Khraisha Q, Put S, Kappenberg J, Warraitch A, Hadfield K. Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. *Res Synth Methods*. Mar 14 2024;doi:10.1002/jrsm.1715

23.   Guo E, Gupta M, Deng J, Park Y-J, Paget M, Naugler C. Automated paper screening for clinical reviews using large language models. *arXiv preprint arXiv:230500844*. 2023;

24.   Robinson A, Thorne W, Wu BP, et al. Bio-sieve: Exploring instruction tuning large language models for systematic review automation. *arXiv preprint arXiv:230806610*. 2023;

25.   Hasan B, Saadi S, Rajjoub NS, et al. Integrating large language models in systematic reviews: a framework and case study using ROBINS-I for risk of bias assessment. *BMJ Evid Based Med*. Feb 21 2024;doi:10.1136/bmjebm-2023-112597

26.   Reason T, Benbow E, Langham J, Gimblett A, Klijn SL, Malcolm B. Artificial Intelligence to Automate Network Meta-Analyses: Four Case Studies to Evaluate the Potential Application of Large Language Models. *Pharmacoecon Open*. Mar 2024;8(2):205-220. doi:10.1007/s41669-024-00476-9

27.   Gartlehner G, Kahwati L, Hilscher R, et al. Data extraction for evidence synthesis using a large language model: A proof-of-concept study. *Res Synth Methods*. Mar 3 2024;doi:10.1002/jrsm.1710

28.   Schopow N, Osterhoff G, Baur D. Applications of the Natural Language Processing Tool ChatGPT in Clinical Practice: Comparative Study and Augmented Systematic Review. *JMIR Med Inform*. Nov 28 2023;11:e48933. doi:10.2196/48933

29.   Fleurence R, Kent S, Adamson B, et al. Assessing Real-World Data from Electronic Health Records for Health Technology Assessment – The SUITABILITY Checklist: A Good Practices Report from an ISPOR Task Force. *Value in Health*. 2024;In Press

30.   Keloth VK, Banda JM, Gurley M, et al. Representing and utilizing clinical textual data for real world studies: An OHDSI approach. *J Biomed Inform*. Jun 2023;142:104343. doi:10.1016/j.jbi.2023.104343

31.   Hu Y, Chen Q, Du J, et al. Improving large language models for clinical named entity recognition via prompt engineering. *J Am Med Inform Assoc*. Jan 27 2024;doi:10.1093/jamia/ocad259

32.   Yang X, Chen A, PourNejatian N, et al. A large language model for electronic health records. *NPJ Digit Med*. Dec 26 2022;5(1):194. doi:10.1038/s41746-022-00742-2

33.   Peng C, Yang X, Chen A, et al. A study of generative large language model for medical research and healthcare. *npj Digital Medicine*. 2023/11/16 2023;6(1):210. doi:10.1038/s41746-023-00958-w

34.   Xie Q, Chen Q, Chen A, et al. Me LLaMA: Foundation Large Language Models for Medical Applications. *arXiv preprint arXiv:240212749*. 2024;

35.   Xu H, Usuyama N, Bagga J, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*. May 22 2024;doi:10.1038/s41586-024-07441-w

36.   Soroush A, Glicksberg BS, Zimlichman E, et al. Large Language Models Are Poor Medical Coders — Benchmarking of Medical Code Querying. *NEJM AI*. 2024;1(5):AIdbp2300040. doi:doi:10.1056/AIdbp2300040

37.   Zhang B, Zhang L, Chen Q, Jin Z, Liu S, Zhang S. Harnessing artificial intelligence to improve clinical trial design. *Communications Medicine*. 2023/12/21 2023;3(1):191. doi:10.1038/s43856-023-00425-3

38.   Hutson M. Cutting to the chase. *Nature*. 2024;

39.   Ghim JL, Ahn S. Transforming clinical trials: the emerging roles of large language models. *Transl Clin Pharmacol*. Sep 2023;31(3):131-138. doi:10.12793/tcp.2023.31.e16

40.   Datta S, Lee K, Paek H, et al. AutoCriteria: a generalizable clinical trial eligibility criteria extraction system powered by large language models. *J Am Med Inform Assoc*. Jan 18 2024;31(2):375-385. doi:10.1093/jamia/ocad218

41.   Wang Z, Xiao C, Sun J. AutoTrial: prompting language models for clinical trial design. *arXiv preprint arXiv:230511366*. 2023;

42.   Unlu O, Shin J, Mailly CJ, et al. Retrieval Augmented Generation Enabled Generative Pre-Trained Transformer 4 (GPT-4) Performance for Clinical Trial Screening. *medRxiv*. Feb 8 2024;doi:10.1101/2024.02.08.24302376

43.   Park J, Fang Y, Ta C, et al. Criteria2Query 3.0: Leveraging generative large language models for clinical trial eligibility query generation. *J Biomed Inform*. Apr 30 2024:104649. doi:10.1016/j.jbi.2024.104649

44.   Nievas M, Basu A, Wang Y, Singh H. Distilling large language models for matching patients to clinical trials. *Journal of the American Medical Informatics Association*. 2024:ocae073.

45.   Yuan J, Tang R, Jiang X, Hu X. LLM for patient-trial matching: Privacy-aware data augmentation towards better performance and generalizability. 2023:

46.   Peikos G, Symeonidis S, Kasela P, Pasi G. Utilizing ChatGPT to enhance clinical trial enrollment. *arXiv preprint arXiv:230602077*. 2023;

47.   Hamer DMd, Schoor P, Polak TB, Kapitan D. Improving Patient Pre-screening for Clinical Trials: Assisting Physicians with Large Language Models. *arXiv preprint arXiv:230407396*. 2023;

48.   Jin Q, Wang Z, Floudas CS, et al. Matching patients to clinical trials with large language models. *ArXiv*. 2023;

49.   Ghosh M, Mukherjee S, Ganguly A, Basuchowdhuri P, Naskar SK, Ganguly D. AlpaPICO: Extraction of PICO frames from clinical trial documents using LLMs. *Methods*. Jun 2024;226:78-88. doi:10.1016/j.ymeth.2024.04.005

50.   Weissler EH, Naumann T, Andersson T, et al. The role of machine learning in clinical research: transforming the future of evidence generation. *Trials*. 2021/08/16 2021;22(1):537. doi:10.1186/s13063-021-05489-x

51.   Caro JJ, Briggs AH, Siebert U, Kuntz KM. Modeling good research practices--overview: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force--1. *Value Health*. Sep-Oct 2012;15(6):796-803. doi:10.1016/j.jval.2012.06.012

52.   Briggs AH, Weinstein MC, Fenwick EA, Karnon J, Sculpher MJ, Paltiel AD. Model parameter estimation and uncertainty: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force--6. *Value Health*. Sep-Oct 2012;15(6):835-42. doi:10.1016/j.jval.2012.04.014

53.   Eddy DM, Hollingworth W, Caro JJ, Tsevat J, McDonald KM, Wong JB. Model transparency and validation: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force--7. *Value Health*. Sep-Oct 2012;15(6):843-50. doi:10.1016/j.jval.2012.04.012

54.   Chhatwal J, Yildrim IF, Balta D, et al. Can Large Language Models Generate Conceptual Health Economic Models? . presented at: ISPOR 2024; 2024; Atlanta, Georgia.

55.   Reason T, Rawlinson W, Langham J, Gimblett A, Malcolm B, Klijn S. Artificial Intelligence to Automate Health Economic Modelling: A Case Study to Evaluate the Potential Application of Large Language Models. *Pharmacoecon Open*. Mar 2024;8(2):191-203. doi:10.1007/s41669-024-00477-8

56.   Adams L, Fontaine E, Lin S., Crowell T., Chung VCH., Gonzalez A. Artificial Intelligence in Health, Health Care, and Biomedical Science: An AI Code of Conduct Principles and Commitments Discussion Draft. *NAM Perspectives, .* 2024;doi:https://doi.org/10.31478/202403a

57.   National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. 2023:42. January 2023. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

58.   Coalition for Health AI. *Blueprint for trustworthy AI implementation guidance and assurance for healthcare*. 2023:25. April 2023 https://www.coalitionforhealthai.org/papers/blueprint-for-trustworthy-ai_V1.0.pdf

59.   Alkaissi H, McFarlane SI. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus*. Feb 2023;15(2):e35179. doi:10.7759/cureus.35179

60.   Lin Z. How to write effective prompts for large language models. *Nat Hum Behav*. Mar 4 2024;doi:10.1038/s41562-024-01847-2

61.   Wang S, Scells H, Zhuang S, Potthast M, Koopman B, Zuccon G. Zero-shot Generative Large Language Models for Systematic Review Screening Automation. *arXiv preprint arXiv:240106320*. 2024;

62.   Yu H, Gan A, Zhang K, Tong S, Liu Q, Liu Z. Evaluation of Retrieval-Augmented Generation: A Survey. *arXiv preprint arXiv:240507437*. 2024;

63.   Saab K, Tu T, Weng W-H, et al. Capabilities of gemini models in medicine. *arXiv preprint arXiv:240418416*. 2024;

64.   Yang L, Xu S, Sellergren A, et al. Advancing Multimodal Medical Capabilities of Gemini. *arXiv preprint arXiv:240503162*. 2024;

65.   Beam AL, Manrai AK, Ghassemi M. Challenges to the Reproducibility of Machine Learning Models in Health Care. *JAMA*. Jan 28 2020;323(4):305-306. doi:10.1001/jama.2019.20866

66.   Kapoor S, Cantrell EM, Peng K, et al. REFORMS: Consensus-based Recommendations for Machine-learning-based Science. *Sci Adv*. May 3 2024;10(18):eadk3452. doi:10.1126/sciadv.adk3452

67.   National Academies of Sciences EaM. *Human-AI Teaming: State-of-the-Art and Research Needs*. 2022.

68.   Drukker K, Chen W, Gichoya J, et al. Toward fairness in artificial intelligence for medical image analysis: identification and mitigation of potential biases in the roadmap from data collection to model deployment. *J Med Imaging (Bellingham)*. Nov 2023;10(6):061104. doi:10.1117/1.JMI.10.6.061104

69.   Gervasi S, Chen I, Smith-McLallen A, et al. The Potential For Bias In Machine Learning And Opportunities For Health Insurers To Address It. *Health Affairs*. 2022;41(2):212-218. doi:10.1377/hlthaff.2021.01287

70.	Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. Oct 25 2019;366(6464):447-453. doi:10.1126/science.aax2342

71.	Parikh RB, Teeple S, Navathe AS. Addressing Bias in Artificial Intelligence in Health Care. *JAMA*. Dec 24 2019;322(24):2377-2378. doi:10.1001/jama.2019.18058

72.	National Institute of Standards and Technology. *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*. 2022:86. March 2022.

73.	Yang Y, Lin M, Zhao H, Peng Y, Huang F, Lu Z. A survey of recent methods for addressing AI fairness and bias in biomedicine. *J Biomed Inform*. Apr 25 2024:104646. doi:10.1016/j.jbi.2024.104646

74.	Xu J, Xiao Y, Wang WH, et al. Algorithmic fairness in computational medicine. *EBioMedicine*. Oct 2022;84:104250. doi:10.1016/j.ebiom.2022.104250

75.	Huang Y, Guo J, Chen WH, et al. A scoping review of fair machine learning techniques when using real-world data. *J Biomed Inform*. Mar 2024;151:104622. doi:10.1016/j.jbi.2024.104622

76.	Juwara L, El-Hussuna A, El Emam K. An evaluation of synthetic data augmentation for mitigating covariate bias in health data. *Patterns (N Y)*. Apr 12 2024;5(4):100946. doi:10.1016/j.patter.2024.100946

77.	Benitez K, Malin B. Evaluating re-identification risks with respect to the HIPAA privacy rule. *J Am Med Inform Assoc*. Mar-Apr 2010;17(2):169-77. doi:10.1136/jamia.2009.000026

78.	Simon GE, Shortreed SM, Coley RY, et al. Assessing and Minimizing Re-identification Risk in Research Data Derived from Health Care Records. *EGEMS (Wash DC)*. Mar 29 2019;7(1):6. doi:10.5334/egems.270

79.	Li H, Moon JT, Purkayastha S, Celi LA, Trivedi H, Gichoya JW. Ethics of large language models in medicine and medical research. *Lancet Digit Health*. Jun 2023;5(6):e333-e335. doi:10.1016/S2589-7500(23)00083-3

80.	Mosquera L, El Emam K, Ding L, et al. A method for generating synthetic longitudinal health data. *BMC Med Res Methodol*. Mar 23 2023;23(1):67. doi:10.1186/s12874-023-01869-w

81.	Brannvall R, Forsgren H, Linge H. HEIDA: Software Examples for Rapid Introduction of Homomorphic Encryption for Privacy Preservation of Health Data. *Stud Health Technol Inform*. May 18 2023;302:267-271. doi:10.3233/SHTI230116

82.	Duffourc MN, Gerke S. Health Care AI and Patient Privacy-Dinerstein v Google. *JAMA*. Mar 19 2024;331(11):909-910. doi:10.1001/jama.2024.1110

83.	Gichoya JW, Banerjee I, Bhimireddy AR, et al. AI recognition of patient race in medical imaging: a modelling study. *Lancet Digit Health*. Jun 2022;4(6):e406-e414. doi:10.1016/S2589-7500(22)00063-2

84.	European Union. EU Artificial Intelligence Act Accessed May 22, 2024. https://artificialintelligenceact.eu/the-act/

85.	UK Government. A pro-innovation approach to AI regulation: government response. https://www.gov.uk/government/consultations/ai-regulation-a-pro-innovation-approach-policy-proposals/outcome/a-pro-innovation-approach-to-ai-regulation-government-response

86.	UK Government. *A pro-innovation approach to AI regulation*. 2023:93. March 2023. Accessed 22 May 2024. https://assets.publishing.service.gov.uk/media/64cb71a547915a00142a91c4/a-pro-innovation-approach-to-ai-regulation-amended-web-ready.pdf

87.   Habuka H. Japan's Approach to AI Regulation and Its Impact on the 2023 G7 Presidency. https://www.mofa.go.jp/files/100573473.pdf

88.   The White House. *Blueprint for an AI Bill of Rights*. 2022:73. October 2022.

89.   The White House. FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence. Accessed 22 May 2024. https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/

90.   Organization for Economic Cooperation and Development (OECD). OECD AI Principles overview. Accessed 22 May, 2024. https://oecd.ai/en/ai-principles

91.   United Nations Educational SaCOU. Ethics of Artificial Intelligence Accessed 22 May 2024. https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

92.   United Nations. *Interim Report: Governing AI for Humanity* 2023:35. December 2023. https://www.un.org/sites/un2.un.org/files/un_ai_advisory_body_governing_ai_for_humanity_interim_report.pdf

93.   World Health Organization. *Regulatory considerations on artificial intelligence for health* 2023:80. https://iris.who.int/bitstream/handle/10665/373421/9789240078871-eng.pdf?sequence=1

94.   World Health Organization. *Ethics and governance of artificial intelligence for health. Guidance on large multi-modal models*. 2024. Accessed 22 May 2024. https://iris.who.int/bitstream/handle/10665/375579/9789240084759-eng.pdf?sequence=1

95.   European Medicines Agency. Reflection paper on the use of Artificial Intelligence (AI) in the medicinal product lifecycle Accessed 22 May 2024. https://www.ema.europa.eu/en/use-artificial-intelligence-ai-medicinal-product-lifecycle

96.   US Food and Drug Administration. *Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan*. 2021. January 2021. https://www.fda.gov/media/145022/download

97.   US food and Drug Administration. Predetermined Change Control Plans for Machine Learning-Enabled Medical Devices: Guiding Principles. Accessed 22 May, 2024. https://www.fda.gov/media/173206/download?attachment

98.   US Food and Drug Administration. *Artificial Intelligence and Medical Products: How CBER, CDER, CDRH, and OCP are Working Together* 2024. https://www.fda.gov/media/177030/download?attachment

99.   US Food and Drug Administration. *Using Artificial Intelligence and Machine Learning in the development of drug and biological products*. 2023. https://www.fda.gov/media/167973/download

100.   US Food and Drug Administration. *Artificial Intelligence in Drug Manufacturing*. 2023:17. 2023. Accessed 29 May 2024. https://www.fda.gov/media/165743/download

101.   Zemplenyi A, Tachkov K, Balkanyi L, et al. Recommendations to overcome barriers to the use of artificial intelligence-driven evidence in health technology assessment. *Front Public Health*. 2023;11:1088121. doi:10.3389/fpubh.2023.1088121

102.   Flanagin A, Pirracchio R, Khera R, Berkwits M, Hswen Y, Bibbins-Domingo K. Reporting Use of AI in Research and Scholarly Publication-JAMA Network Guidance. *Jama*. Apr 2 2024;331(13):1096-1098. doi:10.1001/jama.2024.3471

103.   Koller D, Beam A, Manrai A, et al. Why We Support and Encourage the Use of Large Language Models in NEJM AI Submissions. *Nejm Ai*. 2024;1(1)doi:10.1056/AIe2300128