Abstract: The emergence of generative large language model (LLM) artificial intelligence (AI) represents one of the most profound developments in healthcare in decades, with the potential to create revolutionary and seismic changes in the practice of medicine as we know it. However, significant concerns have arisen over questions of liability for bad outcomes associated with LLM AI-influenced medical decision making. Although the authors were not able to identify a case in the United States that has been adjudicated on medical malpractice in the context of LLM AI at this time, sufficient precedent exists to interpret how analogous situations might be applied to these cases when they inevitably come to trial in the future. This commentary will discuss areas of potential legal vulnerability for clinicians utilizing LLM AI through review of past case law pertaining to third-party medical guidance and review the patchwork of current regulations relating to medical malpractice liability in AI. Finally, we will propose proactive policy recommendations including creating an enforcement duty at the US Food and Drug Administration (FDA) to require algorithmic transparency, recommend reliance on peer-reviewed data and rigorous validation testing when LLMs are utilized in clinical settings, and encourage tort reform to share liability between physicians and LLM developers.

Keywords: artificial intelligence; ChatGPT; large language models; liability; medical malpractice

Sources of liability in medical decision making influenced by AI based on historical review

The roots of the legal definition of malpractice are encapsulated in English jurist William Blackstone’s Commentaries on the Laws of England [6], where he characterized it as a “misdemeanor and offense at common law, whether it be for curiosity and experiment, or by neglect; because it breaks the trust which party had placed in his physician, and tends
to the patient’s destruction.” Patients retain, and frequently exercise, a cause of action in civil courts for the redress of injuries resulting from medical malpractice. As a general rule, a medical provider engages in malpractice if their conduct deviates from the ordinary standard of care in their jurisdiction, and this can differ widely between jurisdictions [7]. Therefore, a physician’s use of LLM AI in treating a patient will likely be analyzed through the lens of the prevailing standard of care [5]. Although the novelty of LLM AI in clinical practice lends uncertainty as to what sort of evidence the courts will treat as being probative of its proper use, it seems most likely that LLM AI-generated advice will be treated as third-party medical guidance. Fortunately, a practitioner’s reliance on (or disregard for) third-party medical guidance is not a novel topic for the courts.

It is important to note that publicly available LLMs such as Chat-GPT are not exclusively utilizing expert-reviewed data to produce medical guidance. In order to make a more direct comparison to the examples that follow, let us suppose that a future LLM marketed directly to physicians will.

Our first analogus situation involves the degree to which package inserts drafted by pharmaceutical manufacturers can be utilized to establish the standard of care concerning the administration of a drug. Drug package inserts are derived from source information and clinical studies reviewed by physicians; however, they may not be authored by them. In Julian vs. Barker, the Supreme Court of Idaho held that a trial court erred in barring the admission of an information sheet that provided directions on the proper administration of sodium pentothal, as the manufacturer was presumed to be qualified to give directions concerning the use of its product [8]. In coming to this conclusion, the Julian Court noted that a drug manufacturer’s directions represented prima facie (“facially sufficient” – a legal term meaning that an argument or position is sound enough at first glance to be presumed valid) evidence of the drug’s proper administration. In Mueller vs. Mueller, the Supreme Court of South Dakota upheld a jury instruction that directed the jury to consider evidence that a physician deviated from a drug manufacturer’s instructions on the proper administration of its product as evidence of negligence [9]. The Mueller Court supported its decision by noting that drug manufacturers were increasingly being found liable for defective products, rendering their instructions more reliable. Also, per the Mueller Court, a busy modern physician had no choice but to rely upon a manufacturer’s instructions, as they could not be expected to independently verify the propriety of a drug’s particular application.

Touching upon physicians’ usage of third-party medical guidance even more closely analogous to LLM AI, other cases have probed the extent to which adherence to a point-of-care decision resource can be considered standard of care. In Spensieri vs. Lasky, the Court of Appeals of the State of New York rejected an argument that drug information compiled in the Physician’s Desk Reference was prima facie evidence of the standard of care, noting that a patient’s individual circumstances were vital to that analysis [10]. Ultimately, the Lasky Court held that Physician’s Desk Reference could properly be incorporated into an expert’s testimony but was not standalone proof of the standard of care.

As evidenced by the split in authority discussed above, some jurisdictions may deem an AI utilizing expert-reviewed data to give medical guidance as representing the standard of care, whereas others may generally reject its applicability. A hybrid approach is also possible, in which courts permit the admission of a generative AI’s response to an inquiry by a physician but require supplemental testimony from a qualified medical expert. Uncertainty and doubt that the courts will evolve, as courts frequently modify or clarify established precedent based upon the unique facts of a particular case [9]. In Lhotka vs. Larson, for instance, the Court noted that a jury instruction that deviation from a manufacturer’s directions constituted evidence of negligence would be appropriate only if said directive was clear and unambiguous. However, because the directive was ambiguous, such an instruction was not warranted [11]. As medical guidance given by generative AI approximates the tone of a human consultant, in practice, it is rarely unambiguous.

Brief review of existing laws and regulations applicable to AI

Legislative and regulatory efforts to address LLM AI have been limited and piecemeal so far, with the dynamic nature of these systems posing unique challenges for traditional regulatory approaches [12, 13]. Currently, the closest thing that exists to a comprehensive federal regulation to address liability for LLM AI-influenced medical decisions is a 2022 revision to the Affordable Care Act’s Section 1557, which states that physicians and covered entities are “liable for medical decisions made in reliance with clinical algorithms [14].” This regulation is problematic, because while the intent of this rule is to address and prevent discrimination
against historically marginalized communities that can result from AI algorithms [15], reasonable interpretation could apply this statement broadly to any medical decision made utilizing AI, including guidance given by LLMs.

The US Food and Drug Administration (FDA) is beginning to evaluate some AI systems as medical devices [16] and has published nonbinding recommendations for classifying clinical decision support (CDS) software as medical devices or not [17]. These recommendations interpret the scope of section 520(o)(1)(I)(i) of the Federal Food, Drug, and Cosmetic Act (FD&C Act), and when applied to a LLM, appear to exempt these systems from being classified as ‘devices’ (i.e., a nondevice CDS). This determination rests on the observation that LLMs, when utilized clinically, are designed for decision support and not (at least currently) involved in acquisition or processing of diagnostic images [17]. Importantly, this section also parallels the Lasky court in stating that recommendations given by a CDS tool should be independently considered by the physician in view of the individual patient and not utilized as the sole determination of diagnostic or treatment decisions [17].

In Congress, no significant legislation that specifically regulates LLM AI in healthcare has been proposed. The most potentially consequential bill is the Algorithmic Accountability Act of 2019 [18], which was recently re-introduced in 2023. This legislation aims to create protections for people affected in a negative way by utilization of AI in decisions on housing, credit, education, and other high-impact applications. If passed, it would create an enforcement duty at the Federal Trade Commission to require that automated systems be assessed for biases and hold bad actors accountable [19]. However, this bill does not mention medical malpractice and proposes no plans to assess clinically utilized LLM AI algorithms for reliability.

Discussion, limitations, and policy recommendations

It should be restated that the most significant limitation of this commentary is that to our knowledge, there are no current or previous cases litigated in the United States that specifically address malpractice liability for physicians utilizing LLM AI. As a result, the preceding legal review remains speculative by nature, representing our best guess at what evidence future courts may consider in these cases. Due to the paucity of legislation and regulatory efforts identified by our review in the previous section, we suggest that there is a critical need and opportunity for proactive action to address this issue through policy rather than waiting for resolution through the legal system.

To ensure the reliability of AI systems, protect patients, and promote the fair application of medical malpractice liability, federal policy should mandate rigorous validation and testing of AI tools before their deployment in clinical settings. The US FDA is the preferred agency to regulate clinical AI reliability, given its expertise in medical devices and software, and should extend this responsibility to LLM algorithms. This process could require AI developers to make their algorithms available for independent validation when utilized in clinical practice, ensuring that AI systems provide clear explanations for their recommendations with verified, peer-reviewed data. Finally, if utilization of high-quality LLM AI increasingly becomes considered the standard of care in most jurisdictions, liability reform will be needed to shift some responsibility for AI-generated medical guidance to algorithm developers, as shown relating to drug package inserts discussed previously. Achieving this may require state-level, rather than federal-level, tort reform, and could be an issue to be resolved later by the courts themselves if policy cannot be enacted first.

Conclusions

Until there is clarity or action in determining the scope of malpractice liability as a result of medical decisions influenced by AI, a significant barrier to the adoption and application to the full potential of this technology in medicine will remain. We recommend consideration and adoption of the policy recommendations given in this commentary as a proactive solution to protect patients and reduce the risk of malpractice liability for physicians that choose to take advantage of the potential of generative LLM AI for clinical applications.

Research ethics: Not applicable.
Informed consent: Not applicable.
Author contributions: Both authors provided substantial contributions to conception and design, acquisition of data, or analysis and interpretation of data; both authors drafted the article or revised it critically for important intellectual content; both authors gave final approval of the version of the article to be published; and both authors agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.
Competing interests: None declared.
Research funding: None declared.
Data availability: Not applicable.
Disclaimer: The views expressed in this material are those of the authors, and do not reflect the official policy or position of the U.S. Government, the Department of Defense, or the Department of the Air Force.

References