

Knowledge Graph and Large Language Model Fusion for Explainable Medical Question Answering

Sofi Jorgensen

Department of Electrical Engineering and Computer Science, University of Missouri,
Columbia, MO, USA.

sjorgensen@missouri.edu

Abstract

Medical question answering has moved from retrieval of curated documents toward generative systems that synthesize evidence across heterogeneous clinical resources. Large language models offer fluent, context-sensitive responses, yet their probabilistic nature creates risks of hallucination, citation drift, and opaque reasoning. Knowledge graphs provide structured, auditable representations of biomedical entities and relations, but they are often incomplete, brittle, and disconnected from natural language interaction. This paper examines the fusion of knowledge graphs and large language models for explainable medical question answering as a socio-technical systems problem rather than a purely model-centric task. It analyzes architectural patterns that couple graph retrieval, semantic alignment, attention-based fusion, and constrained generation, with attention to trade-offs among expressiveness, latency, verifiability, and maintainability. The discussion then turns to explainability as an infrastructure requirement, arguing that clinical explanations must connect model outputs to provenance, evidence hierarchies, and workflow-specific decision thresholds. Governance, privacy, regulatory, and sustainability concerns are examined through the lens of deployment across hospitals, research networks, and patient-facing services. The paper also assesses fairness and robustness risks arising from incomplete graphs, biased documentation, and unequal data representation. It concludes that explainable medical question answering requires layered validation, human oversight, transparent data lineage, and policy mechanisms that treat knowledge graphs and language models as continuously governed clinical infrastructure. The central contribution is a system-level agenda for designing fusion architectures that are clinically useful, institutionally accountable, and resilient under real-world uncertainty.

Keywords

medical question answering; knowledge graphs; large language models; explainable artificial intelligence; retrieval-augmented generation; clinical governance; socio-technical systems; health informatics.

1. Introduction

Medical question answering is one of the most demanding applications of artificial intelligence because it combines open-domain language understanding with high-stakes clinical reasoning. Clinicians, patients, and administrators ask questions that range from medication interactions to differential diagnoses, eligibility criteria, and longitudinal care planning. Early systems relied on curated databases or information retrieval, but recent advances in large language models have made it possible to generate coherent, context-aware answers from heterogeneous text [1], [2]. Yet fluency is not equivalent to clinical validity. A generated answer may be plausible while omitting contraindications, misstating evidence

strength, or failing to distinguish consensus guidelines from preliminary findings [5], [6]. The result is a socio-technical gap: the model can converse, but the surrounding system may not be able to justify, audit, or constrain what it says.

Knowledge graphs offer a complementary representational strategy. They encode biomedical concepts, relations, and provenance in structured form, enabling queries that can be traced through explicit edges and source annotations [8], [9]. In domains such as drug safety, phenotype discovery, and clinical trial matching, graph structures support reasoning that is more transparent than dense vector similarity alone. However, knowledge graphs are difficult to maintain, often incomplete, and not naturally aligned with the ambiguity of clinical language [10], [11]. Their coverage may reflect institutional priorities, coding practices, and publication biases rather than the full spectrum of patient experience. A fusion architecture therefore cannot treat the graph as a static oracle. It must treat the graph as one evidence source among several, subject to uncertainty, versioning, and governance.

This paper argues that knowledge graph and large language model fusion for explainable medical question answering should be designed as layered clinical infrastructure. The relevant design questions are not only which model or retrieval method performs best on a benchmark, but how the system establishes provenance, manages conflicting evidence, supports clinician override, and remains sustainable after deployment. It examines architecture, explainability, governance, privacy, fairness, robustness, deployment economics, and policy. The discussion is intentionally system-level. It assumes that no single component can guarantee safety and that explainability emerges from the interaction of data pipelines, graph curation, retrieval strategies, generation constraints, interface design, and institutional accountability. The paper contributes a conceptual framework for evaluating fusion systems as long-lived, governed, and adaptive medical infrastructure.

2. Background and Related Work

Large language models have changed the technical baseline for medical question answering. Transformer architectures enable long-range contextualization and transfer learning across domains [15], while pretraining and few-shot adaptation allow models to answer tasks without task-specific supervision [16], [17]. In medicine, these capabilities have been tested on licensing-style examinations, differential diagnosis prompts, and patient communication scenarios [5], [6]. The results are often impressive in controlled settings, but they also reveal persistent limitations: sensitivity to prompt phrasing, tendency to fabricate citations, uneven calibration, and difficulty in recognizing when local institutional policy differs from general medical knowledge. Foundation model research has further shown that capabilities and risks scale together, making governance and evaluation inseparable from model development [18].

Retrieval-augmented generation was introduced to reduce reliance on parametric memory by conditioning generation on retrieved documents [7]. In medical settings, retrieval can draw from guidelines, systematic reviews, electronic health records, drug labels, and institutional protocols. However, retrieval quality depends on indexing, chunking, embedding, and query interpretation. A semantically similar passage may not be clinically applicable to the patient at hand, and a retrieved guideline may be outdated or superseded. Multi-source retrieval therefore creates a reconciliation problem. Systems must decide how to weight evidence, how to expose conflicts, and how to preserve attribution when several documents support or contradict a generated statement. Recent work on unified semantic spaces and attention-based fusion across heterogeneous sources addresses part of this problem by aligning representations and dynamically weighting evidence during generation [14].

Knowledge graphs provide an alternative or complementary substrate for medical reasoning. They can represent diseases, symptoms, drugs, genes, procedures, outcomes, and provenance relations in a graph structure that supports traversal and querying [8], [9]. Linked data principles and semantic web standards allow distributed biomedical resources to be connected through shared identifiers, improving interoperability across institutions and data silos [10]. The FAIR principles further emphasize findability, accessibility, interoperability, and reuse, which are essential when graphs are assembled from multiple sources [11]. Yet graph construction is labor-intensive, and biomedical ontologies evolve rapidly. A graph that is accurate for one specialty may be misleading in another, and an edge that lacks temporal context may be unsafe in clinical decision support. Explainability research in healthcare has repeatedly shown that explanations must match the decision context rather than satisfy abstract interpretability criteria alone [12], [13].

3. Architecture of Knowledge Graph and Large Language Model Fusion

A fusion architecture for medical question answering typically begins with a query understanding layer that maps natural language questions into clinical intents, entities, constraints, and temporal scopes. This layer may use named entity recognition, ontology linking, and specialty-specific normalization. The quality of this mapping determines whether graph retrieval and document retrieval target the right concepts. For example, a question about whether a patient with renal impairment can receive a particular anticoagulant requires recognition of the drug, the indication, the comorbidity, and the relevant thresholds. If the query is mapped only to the drug name, the system may retrieve general pharmacology but miss the contraindication. A robust architecture therefore treats query interpretation as a governed clinical task rather than a preprocessing detail.

The next layer performs multi-source retrieval. Structured sources such as knowledge graphs, drug databases, and coding terminologies provide explicit relations, while unstructured sources such as guidelines, trial reports, and notes provide nuance and context. A recent unified semantic space and attention-based fusion framework [14] illustrates how heterogeneous sources can be aligned and weighted during retrieval-augmented generation. The architectural implication is important: fusion is not merely concatenating graph triples with text passages. It involves representation alignment, source reliability estimation, temporal validity checks, and conflict detection. If a graph edge and a guideline paragraph disagree, the system must decide whether the disagreement reflects a knowledge update, a population difference, or an error. Without such reconciliation, fusion can amplify inconsistency rather than reduce it.

Generation in a fusion system should be constrained by the retrieved evidence. Constrained decoding, citation grounding, and answer templates can reduce hallucination, but they also introduce trade-offs. Strong constraints improve verifiability and safety for well-defined questions, yet they may limit the system's ability to synthesize novel evidence or respond to rare presentations. Weak constraints preserve flexibility but increase the risk of unsupported claims. A practical design may use tiered generation, where high-stakes questions require explicit evidence links and clinician review, while lower-stakes educational questions permit broader synthesis with clear disclaimers. This tiering must be visible to users and adjustable by institutional policy, because the same question may carry different risk depending on whether it is asked for learning, triage, or prescribing. The final layer is the clinical interface and feedback loop. Explainability is not delivered only as a textual rationale; it is embedded in the display of evidence, uncertainty, alternatives, and provenance.

4. Explainability Mechanisms and Clinical Reasoning

Explainability in medical question answering must answer several different questions. A clinician may ask why a recommendation was made, a patient may ask what it means for daily life, a regulator may ask whether the system's behavior is consistent with approved labeling, and an institution may ask whether the answer can be defended in an audit. These audiences require different forms of explanation, even when they refer to the same underlying computation. A graph can provide an explicit path from symptom to disease to treatment, but the path may be incomplete or probabilistic. A language model can provide a fluent rationale, but the rationale may be post hoc and not faithful to the actual computation. Fusion systems can improve faithfulness by forcing generated claims to correspond to retrieved graph edges or document spans, yet faithfulness is not guaranteed merely because a citation is present [12], [13].

One useful distinction is between evidence explanation and reasoning explanation. Evidence explanation identifies the sources and data points that support an answer. Reasoning explanation describes how those sources were combined, weighted, and interpreted. In a fusion system, evidence explanation can be supported by graph provenance, source metadata, and retrieval traces. Reasoning explanation can be supported by attention-based fusion weights, conflict-resolution rules, and uncertainty estimates. However, exposing all internal weights may overwhelm users and create false confidence in numerical precision. A better approach is to provide layered explanations that allow users to drill down from a concise summary to a detailed evidence trail. The interface should distinguish between what the system knows from structured sources, what it infers from text, and what remains uncertain [22]. Clinical reasoning is also temporal and contextual. Explainability must include temporal and patient-specific qualifiers rather than flattening them into a single recommendation [4], [12].

5. Data Governance, Privacy, and Regulatory Infrastructure

Medical question answering systems depend on data that are sensitive, distributed, and regulated. Knowledge graphs may combine public ontologies, proprietary drug databases, institutional formularies, and patient-derived information. Large language models may be pretrained on broad corpora and then adapted to local data. Each data source carries legal, ethical, and operational obligations. Privacy regulations such as the Health Insurance Portability and Accountability Act in the United States and the General Data Protection Regulation in Europe impose constraints on collection, use, disclosure, and retention. Even when data are de-identified, re-identification risks remain, especially when graphs link rare conditions, geographies, and temporal patterns [20]. Governance must therefore address not only access control but also purpose limitation, minimum necessary use, and auditability.

The FAIR principles provide a useful starting point for data stewardship, but they do not resolve questions of consent, equity, or institutional power [11]. A knowledge graph may be findable and accessible while still encoding biased terminology or excluding underrepresented populations. A language model may be interoperable while generating recommendations that are inappropriate for local practice. Regulatory frameworks for medical software add another layer. A system that provides diagnosis or treatment recommendations may be classified as a medical device, requiring validation, post-market surveillance, and change control. A system that retrieves educational information may face lighter requirements, yet the boundary is often blurry in user experience. Infrastructure for governance must include data lineage tracking, versioned knowledge graphs, model cards, audit logs, and human oversight committees.

Because fusion systems combine multiple changing components, governance cannot rely on one-time certification. It requires continuous monitoring, periodic revalidation, and clear responsibility assignment [21].

6. Deployment, Sustainability, and Operational Trade-offs

Deployment of knowledge graph and language model fusion systems involves trade-offs among accuracy, latency, cost, maintainability, and clinician burden. A high-throughput emergency department may require sub-second responses, while a rare disease consultation may tolerate longer retrieval and reasoning. A patient-facing application may need to run on low-cost infrastructure, while a hospital system may have substantial compute resources but strict security requirements. These differences mean that a single architecture is unlikely to serve all contexts. Deployment should be staged, with clear use-case boundaries and escalation pathways. A system that answers medication questions for clinicians may be valuable even if it cannot safely answer diagnostic questions. The key is to align capabilities with risk and to make limitations visible rather than hiding them behind conversational fluency.

Sustainability is often overlooked in artificial intelligence research. Large language models require significant energy and computational resources for training and inference, while knowledge graphs require continuous curation by domain experts. If a system depends on frequent re-embedding of entire corpora or human review of every graph update, its long-term viability may be limited. Sustainable design favors modular updates, incremental indexing, caching of common queries, and selective use of large models for difficult cases. It also favors reuse of standardized ontologies and shared governance protocols rather than bespoke pipelines for every institution. The environmental and financial costs of deployment are not separate from clinical value; they shape which systems can persist long enough to be evaluated and improved [19]. Operational integration is another critical factor. A system embedded in the electronic health record, clinical decision support workflow, or patient portal has a better chance of being used, but it also raises concerns about alert fatigue, automation bias, and documentation burden. The interface must therefore support calibrated trust through uncertainty, inspectable evidence, and rapid override [22].

7. Fairness, Robustness, and Policy Implications

Fairness in medical question answering is multidimensional. A system may perform equally well on average while failing for patients with rare diseases, limited English proficiency, disabilities, or non-standard gender identities. Knowledge graphs can encode historical biases in diagnosis codes, research funding, and clinical documentation. Language models can reproduce stereotypes or generate recommendations based on majority-population evidence. If a graph underrepresents a condition in a particular group, retrieval may fail to surface relevant evidence, and the language model may fill the gap with generic or incorrect information. Fairness therefore requires disaggregated evaluation, community engagement, and deliberate supplementation of underrepresented knowledge [3]. It cannot be achieved by removing protected attributes from inputs if the underlying data and relations remain biased.

Robustness is equally important. Medical language is ambiguous, misspelled, abbreviated, and context-dependent. A system may encounter adversarial prompts, outdated protocols, or conflicting guidelines. It may be deployed during a pandemic when evidence changes rapidly, or in a rural clinic with limited connectivity. A robust fusion architecture should detect when it lacks sufficient evidence, when sources conflict irreconcilably, and when a query falls

outside its validated scope. It should degrade gracefully, perhaps by providing evidence without a recommendation or by escalating to a human. Robustness also includes resistance to data poisoning, because a knowledge graph or retrieval index can be manipulated if governance is weak. Security, privacy, and clinical safety are therefore interconnected properties of the same infrastructure [19], [20]. Policy implications follow from these technical realities. Regulators may need to assess not just the language model but the entire fusion pipeline, including retrieval sources, graph curation, and interface design. Professional bodies may need to define acceptable explanations and required human review levels. Policy should also support shared public infrastructure for biomedical knowledge graphs and evaluation benchmarks [21].

8. Evaluation, Case Illustrations, and Cross-Domain Comparisons

Evaluating fusion systems requires metrics that go beyond answer accuracy. Clinical correctness, evidence sufficiency, citation fidelity, calibration, and safety may all matter, and their relative importance depends on the use case. A system that answers a factual question about drug metabolism may be assessed mainly on correctness and citation fidelity. A system that supports triage must also be assessed on sensitivity, specificity, and escalation behavior. Human evaluation remains essential because automated metrics may reward fluent but unsafe answers. Clinician reviewers can assess whether an answer is actionable, whether it acknowledges uncertainty, and whether it would change management. Patient reviewers can assess clarity, emotional tone, and whether the answer supports informed decision-making. A multi-layer evaluation framework should include simulation, retrospective chart review, prospective shadow mode, and controlled clinical trials [4], [6].

Case illustrations help clarify the trade-offs. Consider a question about managing hypertension in a patient with chronic kidney disease and diabetes. A graph might provide explicit relations among comorbidities, drug classes, and contraindications. A language model might synthesize guideline recommendations and explain lifestyle factors. A fusion system could retrieve the relevant graph paths, identify conflicting recommendations between cardiology and nephrology guidelines, and present both with their evidence levels. In another case, a patient asks whether a new symptom is related to a medication. The graph may link the drug to known adverse events, but the symptom is non-specific. A robust system should not simply assert causality. It should provide possible associations, prevalence information, and advice to seek clinical evaluation. Cross-domain comparisons with finance, law, and engineering maintenance show a shared need for provenance, temporal validity, and accountability, while medicine differs in the intimacy of patient data and the immediacy of harm [14].

9. Future Directions and Research Agenda

Future research should advance fusion architectures that are both more capable and more governable. One direction is dynamic knowledge graph construction, where clinical documents, guidelines, and patient data are continuously transformed into graph structures with provenance and uncertainty annotations. This would reduce reliance on static curated graphs while raising new challenges for validation and version control. Another direction is uncertainty-aware generation, where the language model explicitly distinguishes between evidence-supported claims, inferred associations, and unanswered questions. Such systems could improve calibration and reduce hallucination, but they require training data and evaluation methods that reward appropriate abstention rather than fluent guessing. Research

should also explore modular architectures that allow graph components, retrievers, and generators to be updated independently without retraining the entire system.

Human-computer interaction is another critical frontier. Explainable medical question answering will succeed only if explanations are usable in real workflows. This requires studying how clinicians and patients interpret evidence traces, confidence indicators, and conflicting recommendations. It also requires designing interfaces that avoid automation bias while reducing cognitive burden. Feedback mechanisms should be co-designed with users so that overrides and corrections improve the system without creating new liabilities. Policy and infrastructure research must keep pace with technical development. Shared benchmarks, standard evidence schemas, and interoperable provenance formats can reduce duplication and improve comparability. Public investment in biomedical knowledge graphs and evaluation datasets can support equitable access, especially for rare diseases and underserved populations. Ultimately, the research agenda should treat explainable medical question answering as a long-term socio-technical system whose success depends on technical innovation, institutional accountability, and sustained public trust [5], [18].

10. Conclusion

Knowledge graph and large language model fusion offers a promising path toward explainable medical question answering because it combines the generative flexibility of language models with the structured provenance of graphs. Neither component is sufficient alone. Language models can synthesize and communicate, but they can also hallucinate and obscure their reasoning. Knowledge graphs can provide explicit relations and audit trails, but they are incomplete, context-dependent, and difficult to maintain. Fusion architectures can mitigate some weaknesses by retrieving heterogeneous evidence, aligning representations, detecting conflicts, and constraining generation. The most important design choices are not merely algorithmic; they concern how evidence is sourced, how uncertainty is represented, how explanations are delivered, and how accountability is assigned.

Realizing this vision requires a socio-technical approach that integrates architecture, governance, deployment, fairness, robustness, and policy. Systems must be evaluated in context, monitored continuously, and updated through transparent processes. They must support clinicians and patients without creating false confidence or new inequities. They must be sustainable in computational, financial, and human terms. The future of explainable medical question answering will not be determined by a single model or graph but by the institutional capacity to govern a complex information infrastructure. If designed with care, fusion systems can make medical knowledge more accessible, more transparent, and more useful while preserving the human judgment at the center of care.

References

1. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44-56. <https://doi.org/10.1038/s41591-018-0300-7>
2. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14), 1347-1358. <https://doi.org/10.1056/NEJMr1814259>
3. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447-453. <https://doi.org/10.1126/science.aax2342>

4. Shortliffe, E. H., & Sepulveda, M. J. (2018). Clinical decision support in the era of artificial intelligence. *JAMA*, 320(21), 2199-2200. <https://doi.org/10.1001/jama.2018.17163>
5. Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Seneviratne, M., Gambhir, S. S., Kelly, C., Nathanson, D., Heacock, L., Topol, E. J., Dean, J., & Kohane, I. (2023). Large language models encode clinical knowledge. *Nature*, 620(7972), 172-180. <https://doi.org/10.1038/s41586-023-06291-2>
6. Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature Medicine*, 29(8), 1930-1940. <https://doi.org/10.1038/s41591-023-02448-8>
7. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459-9474.
8. Nicholson, D. N., & Greene, C. S. (2020). Constructing knowledge graphs and their biomedical applications. *Computational and Structural Biotechnology Journal*, 18, 1414-1428. <https://doi.org/10.1016/j.csbj.2020.05.017>
9. Wang, Q., Mao, Z., Wang, B., & Guo, L. (2017). Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12), 2724-2743. <https://doi.org/10.1109/TKDE.2017.2754499>
10. Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data - The story so far. *International Journal on Semantic Web and Information Systems*, 5(3), 1-22. <https://doi.org/10.4018/jswis.2009081901>
11. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018. <https://doi.org/10.1038/sdata.2016.18>
12. Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, 20, 310. <https://doi.org/10.1186/s12911-020-01332-6>
13. Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11), e745-e750. [https://doi.org/10.1016/S2589-7500\(21\)00208-9](https://doi.org/10.1016/S2589-7500(21)00208-9)
14. Deng, Q., Liu, S., Qiu, H., Long, N., Lin, W., & Chen, S. (2026, April). Unified Semantic Space and Attention-Based Fusion Framework for Multi-Source Heterogeneous Retrieval-Augmented Generation. In *2026 6th International Conference on Artificial Intelligence and Industrial Technology Applications (AIITA)* (pp. 591-597). IEEE.
15. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998-6008.

16. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901.
17. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171-4186). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>
18. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv*. <https://doi.org/10.48550/arXiv.2108.07258>
19. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3645-3650). Association for Computational Linguistics. <https://doi.org/10.18653/v1/P19-1355>
20. Price, W. N., & Cohen, I. G. (2019). Privacy in the age of medical big data. *Nature Medicine*, 25(1), 37-43. <https://doi.org/10.1038/s41591-018-0272-7>
21. World Health Organization. (2021). Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization. <https://www.who.int/publications/i/item/9789240029200>
22. Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S. J., Lermer, E., Coughlin, J. F., Gutttag, J. V., Colak, E., & Ghassemi, M. (2021). Do as AI say: Susceptibility in deployment of clinical decision-aids. *npj Digital Medicine*, 4, 31. <https://doi.org/10.1038/s41746-021-00385-9>