PERSPECTIVE • **OPEN ACCESS**

A hybrid deployment model for generative artificial intelligence in hospitals

MACHINE LEARNING
Health

OPEN ACCESS

RECEIVED
2 April 2025

REVISED
12 May 2025

ACCEPTED FOR PUBLICATION
21 May 2025

PUBLISHED
31 July 2025

PERSPECTIVE

A hybrid deployment model for generative artificial intelligence in hospitals

Maxime Griot[1,2,*], **Coralie Hemptinne**[1], **Jean Vanderdonckt**[2] and **Demet Yuksel**[1,3]

[1] Institute of NeuroScience, Université catholique de Louvain, Brussels, Belgium
[2] Louvain Research Institute in Management and Organizations, Université catholique de Louvain, Louvain-la-Neuve, Belgium
[3] Medical Information Department, Cliniques Universitaires Saint-Luc, Brussels, Belgium
[*] Author to whom any correspondence should be addressed.

E-mail: maxime.griot@uclouvain.be

Keywords: hybrid, deployment, models, generative, artificial intelligence, hospitals

Abstract

This paper analyzes challenges in deploying generative AI in healthcare, examining inadequate evaluation methods, vendor-provided use cases, and adaptation needs. We propose a hybrid framework that uses vendor models for non-medical applications while implementing hospital-managed infrastructure for clinical use cases. This approach balances innovation with patient safety, providing control over adaptation while mitigating biases, ensuring regulatory compliance, and maintaining long-term stability for sustainable healthcare integration.

1. Introduction

The rapid development of Artificial Intelligence (AI) along with the availability of large amounts of data in healthcare, is driving a rapid growth of applications and research in applied AI for medicine [1]. The impressive capabilities of Generative AI (GenAI) to process natural language along with the vast amounts of free text available in EHRs were quickly recognized and appeared as a promising solution to the challenges faced by healthcare to process this type of data. A major challenge with the introduction of EHRs is the duplication of information through structured and unstructured data. Clinicians often rely on unstructured data such as clinical notes found in the EHRs, partly due to the increased workload, which makes data sharing and secondary use inefficient [2, 3]. Another challenge is the heavy reliance on large amounts of information in various formats, particularly free text, and on multiple platforms, requiring clinicians to spend up to 50% of their working time in the EHR or accessing external resources to assist with clinical decision [4].

The difficulty of leveraging unstructured natural language has motivated foundational model trainers to investigate the capabilities of GenAI in healthcare, but these investigations were performed in a context of heavy competition between technology companies, which led to limited evaluation methodologies often without the implication of medical professionals [5, 6]. The focus on automated evaluation methodologies and the lack of clinical studies to assess the relevance of AI models in clinical practice contribute to a misinterpretation of the capabilities of Generative AI for healthcare, with studies demonstrating low performance on more realistic scenarios [7]. Additionally, LLMs can exhibit extreme biases toward certain ethnic groups or genders [8]. Previous work demonstrated that models are unable to appropriately represent the demographic diversity in medical conditions and more concerning make incorrect recommendations due to the patient's demographic [9].

In this work, we first analyze current evaluation methods for GenAI in medicine, examine vendor-provided use cases, highlight adaptation challenges, and then propose a hybrid framework that balances innovation with patient safety.

2. Challenges with AI vendors

2.1. Evaluation of generative AI in medicine

Benchmarks such as MedQA, based on multiple-choice medical exams, dominate AI evaluation but have limited real-world validity [5, 10]. Model developers such as OpenAI, Microsoft or Google report the medical performance of their models by evaluating the accuracy using these benchmarks with MedQA being the most commonly reported benchmark [6, 11, 12]. These actors have greatly influenced the testing methodologies for the medical domain and shaped what is now the standard reported metrics in the field [13]. Furthermore, the widespread adoption of these benchmarks as the standard for evaluating GenAI in medicine has had unintended consequences. Specifically, these benchmarks are increasingly being treated not just as tools for evaluation but as goals in themselves, with leaderboards ranking model performance [13]. This phenomenon is exemplified by MedPrompt, which employs complex strategies to achieve higher scores on medical benchmarks, reaching over 90% accuracy when combining all strategies [14]. While competition fosters innovation, the availability of test data raises concerns about contamination, where models may have been exposed to questions during training, undermining reliability. This overemphasis on benchmark scores can mislead stakeholders, as high accuracy does not always translate to real-world clinical effectiveness.

The use of standardized questions allows for quick, automated tests but is based on two hypotheses. First, that standardized tests are a reliable method to assess medical knowledge and capabilities, [15] and secondly, GenAI models can be evaluated using similar modalities as humans [16]. Both hypotheses are often taken at face value, and few studies attempt to validate them in practice. Existing literature on standardized testing found no link between USMLE scores and the selection of the chief resident, but there was predictive value in USMLE step 1 scores and board pass rates [17, 18]. While some studies do find associations with future career steps in a physician's career, a key limitation is assessing if this association is caused by knowledge of the material or other characteristics that are good predictors of career development. For example, it has been shown that IQ scores are good predictors of job performance, which could be the confounding factor contributing to the identified correlations between USMLE scores and career advancement [19]. Similarly, the second hypothesis that assumes models can be evaluated using the same tools as used for human candidates lacks supporting evidence; there is, on the contrary, supporting evidence that these tools are not a reliable methodology to assess these models [20]. For instance, an experiment showed that models achieved high scores on multiple-choice questions about fictional medical scenarios, whereas medical experts performed no better than random guessing [21].

Beyond these concerns, unexpected behaviors can appear in models such as major biases in terms of race and gender. For instance, GPT-4 exhibits extreme biases when generating clinical vignettes of diseases such as sarcoidosis or even COVID [8]. These biases could lead to real consequences and adverse outcomes for patients, as models may diagnose and treat certain populations better than others—for example, by prescribing advanced examinations at a higher rate [7]. Models can also be misleading, providing correct answers for incorrect or spurious reasons, such as hallucinating findings in a CT image that were not provided or misunderstanding a differential diagnosis [22].

GenAI applications for medicine are already available on the market without any supervision or regulatory approval, often directly available to patients who lack the knowledge to critically analyze [23]. This demonstrates the disconnect between the careful medical approach of historical actors and the new actors, who may not understand the requirements and best practices associated with healthcare.

2.2. Vendor-provided use cases and implementation challenges

The creation and validation of relevant use cases remain critical. External vendors, such as EHR providers, have demonstrated agility in iterating on various applications, including chart summarization, patient message responses, and note generation from transcriptions. The swift deployment of these use cases reflects not only the vendors' commitment to supporting GenAI, but also the substantial resources invested in this ongoing effort. Additionally, vendors are prioritizing usability, striving for seamless integration with existing tools. However, this rapid pace of innovation—driven by global excitement and aggressive competition—has often come at the expense of thorough validation. Notably, only 1.9% of FDA-approved AI medical devices are released with supporting scientific publications validating their use, and 43% lack published validation data altogether [24, 25].

This development strategy raises several concerns. First, the proposed use cases may not align with the needs of all healthcare providers. For example, in our institution, patients cannot message clinicians directly, rendering this feature irrelevant. Second, the lack of prior validation places hospitals in a precarious position, requiring them to commit resources to unvalidated products without robust evidence to justify these investments. Lastly, this top-down development approach often results in a one-size-fits-all product that fails to accommodate the diverse practices and requirements of different healthcare structures. For instance,

models may generate regionally inappropriate recommendations, such as providing U.S. helplines for non-U.S. patients.

A major challenge with vendor-provided AI tools is their lack of transparency. These models operate as 'black boxes,' generating output without revealing their underlying decision-making processes. This opacity makes it difficult for clinicians to trust or justify AI-assisted recommendations. Additionally, as models evolve beyond the control of healthcare institutions, the inability to scrutinize updates further erodes accountability and trust in AI-driven decisions.

The adoption of GenAI in healthcare also faces barriers in non-clinical applications due to the challenge of delineating medical from non-medical use cases. Introducing GenAI for non-medical purposes, such as administrative tasks like billing, scheduling, and resource planning holds significant promise for improving operational efficiency. Yet, this approach is not without risks. Increased familiarity with these tools may inadvertently encourage their use in borderline or unintended applications, potentially crossing into clinical domains with clinicians using these tools for clinical decision support or clinical documentation [26]. Such misuse could lead to data breaches, compromises of sensitive patient information, and even unintended patient safety risks [27].

2.3. Local adaptation

Beyond the potential misalignment between use cases and local needs, there is a risk of misalignment between the memorized clinical knowledge and local specificities. Healthcare is deeply influenced by regional epidemiology, cultural practices, and healthcare infrastructure. LLMs, primarily trained on English-language literature from high-income regions like the United States and Europe, may not adequately address regions with different epidemiology, cultural practices, and resource availability. For instance, the antibiotic resistance profile of pneumococcus differs significantly across regions and is influenced by local vaccination practices [28]. A system trained predominantly on North American and European data might recommend inappropriate treatments in such cases, potentially leading to suboptimal or harmful outcomes.

Even within the same healthcare system, hospital workflows, documentation styles, and EHR configurations can differ significantly. A single AI model may not seamlessly integrate across diverse clinical settings, leading to inefficiencies or inconsistencies. Additionally, individual physicians have unique documentation preferences, which current models do not accommodate, requiring extra time for manual corrections rather than improving efficiency.

Furthermore, the adaptation of AI models to local contexts is constrained by the opacity of commercial AI development. Most vendor-provided models operate as proprietary systems, limiting the ability of healthcare institutions to modify, retrain, or tailor them to their specific needs. Because models are typically trained externally on datasets that may not reflect local patient populations, the ability to adjust them post-deployment is often minimal. This lack of transparency makes it difficult to assess the appropriateness of AI recommendations for local practice, further exacerbating concerns about safety and clinical relevance.

2.4. Lack of visibility

Medicine faces a new challenge due to the speed at which AI develops. Historically, medical sciences have taken a slow, careful approach to novelty, with new drugs needing between 10 and 15 years to reach the market [31]. These processes contribute to patient safety and ensure that new interventions have demonstrated benefits for patients. However, generative AI models evolve at an unprecedented pace compared to traditional medical devices, often undergoing major updates every few months, creating uncertainty regarding their long-term viability and clinical reliability (table 1).

One of the most pressing concerns is the short lifespan of AI models and the frequency of updates, which contrast sharply with the long-term nature of medical validation. In traditional medicine, once a drug or treatment protocol is established, it remains relatively stable for years, allowing clinicians to build expertise and trust in its use. In contrast, AI models undergo frequent modifications, retraining, and version changes, often without clear documentation on how these updates impact medical performance. A model validated for clinical use today may be significantly altered or even deprecated within a year, requiring constant reevaluation and adaptation. This lack of stability raises concerns about the sustainability of AI-driven workflows in healthcare, as hospitals may need to continuously assess and recalibrate their integration strategies to keep up with evolving models.

Adding to this challenge is the uncertainty surrounding long-term support and model availability. Many AI vendors, including large technology firms, operate on commercial strategies that prioritize newer models over maintaining legacy versions. For instance, Azure OpenAI states that they will notify retirement of a model at least 60 days before the date of retirement, [32] which is insufficient for hospitals to conduct the necessary validation of candidate replacement models. If a healthcare institution invests significant time and resources into integrating an AI model, there is no guarantee that the model will remain accessible in the

Table 1. Comparison of key aspects of the development and support of the lifecycle between traditional medical devices and GenAI models.

Aspect	Traditional medical devices	GenAI models
Development timeline	3–7 years from concept to market [29]	Months between major model releases
Regulatory review	Comprehensive pre-market approval	Limited or non-existent for many applications [30]
Version stability	Post-market surveillance and compliance, clinical follow-up for entire lifecycle	Models may be retired from the market within 1–2 years
Update frequency	Scheduled, with clear documentation	Often continuous, with limited transparency
Cost predictability	Fixed licensing or per-unit costs	Usage-based pricing subject to frequent changes
Support commitments	Long-term support guarantees for at least 10 years for Medical Device Regulation compliance	Limited guarantees (e.g. 60 days' notice before retirement)
Backward compatibility	Typically maintained across versions	May require significant workflow adjustments

long term. This raises questions about continuity of care and the risks of adopting tools that may be discontinued or altered without sufficient notice. Unlike traditional medical devices or pharmaceuticals that have well-defined regulatory post-market surveillance requirements, AI models exist in a more fluid and less regulated space, making it difficult for hospitals to plan for long-term stability.

Despite AI's rapid evolution, regulatory frameworks struggle to keep pace, raising concerns about oversight and long-term governance. In response, the European Union has introduced regulations such as the Medical Device Regulation (MDR), which sets stringent requirements for AI-driven medical tools classified as high-risk, including those used for diagnosis, treatment, or monitoring (Box 1) [33]. The upcoming AI Act, which takes effect in 2026, will further reinforce these requirements [34]. Additionally, the General Data Protection Regulation (GDPR) ensures appropriate handling of patient data, mandating consent-based processing [35]. However, given the speed at which AI models evolve and retire, compliance with these regulations remains a challenge, as frequent updates may require repeated validation cycles, creating additional burdens for healthcare institutions.

Box 1: Excerpt from the MDR defining what qualifies as a Medical Device.

[system]…intended by the manufacturer to be used, alone or in combination, for human beings for one or more of the following specific medical purposes:

- *Diagnosis, prevention, monitoring, prediction, prognosis, treatment or alleviation of disease;*
- *Diagnosis, monitoring, treatment, alleviation of, or compensation for, an injury or disability;*
- *Investigation, replacement or modification of the anatomy or of a physiological or pathological process or state;*
- *Providing information by means of in vitro examination of specimens derived from the human body, including organ, blood and tissue donations.*

Beyond the technical uncertainties, the cost of AI implementation remains unpredictable and highly variable. Unlike fixed medical device costs, GenAI pricing is dynamic and often unpredictable, posing budgeting challenges for hospitals. For instance, when GPT-4 was initially released, its cost was set at $60 per million output tokens. However, with the introduction of GPT-4.5, pricing has increased to $150 per million output tokens [36]. Given the high token consumption of complex AI tasks, such as reasoning-based models or Retrieval-Augmented Generation systems, which require additional indexing, embedding, and input token consumption, even short-term budget estimations become highly complex. The exponential growth in AI capabilities leads to increased computational demands, which, in turn, drive higher operational costs. Healthcare institutions face the challenge of managing these unpredictable expenses, making financial planning for AI adoption increasingly difficult.

Moreover, the pricing structures of AI services are subject to frequent modifications, often dictated by market dynamics rather than healthcare needs. Hospitals may find themselves locked into AI-driven workflows, only to face sudden cost increases that make continued usage prohibitive. This lack of cost stability is particularly concerning in public healthcare systems, where budgets are fixed on an annual or multi-year basis, limiting flexibility to accommodate fluctuating AI expenses.

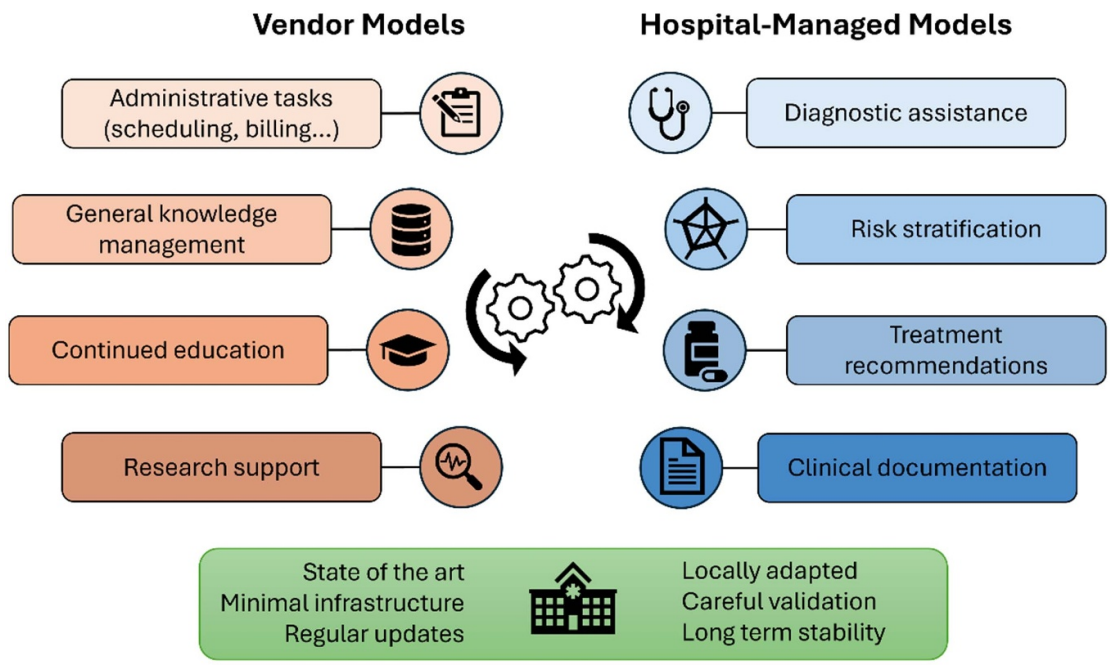

Figure 1. Selected use cases illustrating a hybrid model for generative AI deployment in healthcare. The diagram outlines examples of tasks suited to vendor-provided models versus those requiring hospital-managed models. Administrative, educational and research-oriented tasks, knowledge management and cohort identification, can often be supported by external solutions. In contrast, clinically sensitive applications, including diagnostic support, risk stratification (for example, anticipating care load or identifying patients at risk of clinical deterioration), treatment recommendations and documentation, are better suited to local deployment, where institutions can maintain control over clinical validity, data integrity and operational reliability. This framework aims to guide institutions in balancing external innovation with appropriate levels of clinical oversight.

3. A case for a hybrid approach to generative AI in medicine

Considering the issues identified, we propose a hybrid approach to GenAI adoption in healthcare organizations to benefit from the rapid advancements from vendors and the security provided by medical validation protocols. This approach is based on the distinction between non-medical and medical applications, as depicted in figure 1. However, some use cases exist in a gray area where it is difficult to clearly categorize them as either medical or non-medical. Examples include AI-generated discharge summaries, clinical documentation support, or language translation of patient instructions. In such borderline cases, decisions should be guided by a combination of factors: the degree of clinical influence the tool exerts, the presence and sensitivity of patient-identifiable data, applicable regulatory definitions, the potential consequences of errors, and the level of localization required. For instance, if a tool influences patient understanding or clinician decisions without formal review, it should be treated as a medical application and deployed locally. Conversely, tasks that are always verified by clinicians and do not modify the clinical workflow directly may be safely managed through vendor solutions. This framework helps ensure that institutions maintain control over sensitive or high-risk functions while still benefiting from external innovation where appropriate.

For non-medical applications, we propose the use of state-of-the-art models and integrated solutions that enable the rapid deployment of new features benefiting end users. Access to these tools provides valuable insights into which developments are practically useful in production, ensuring that innovation aligns with operational needs.

3.1. Local models

For medical use cases, we propose that hospitals manage their own infrastructure and models. This approach addresses many issues found in vendor solutions—such as local adaptation, stable versioning, and transparent updates—but it also entails higher internal development costs and, at times, slightly lower performance relative to cutting-edge vendor LLMs. Hospital-managed AI models provide greater control over adaptation, ensuring alignment with local needs while mitigating vendor-induced biases or frequent deprecations [37].

Leveraging internal data allows hospitals to develop relevant benchmarks and refine evaluations based on real-world usage [38]. This approach enables finetuning for local populations and clinician preferences,

improving practical integration into workflows [39]. Additionally, the ability to continuously refine models based on open clinician feedback enhances both reliability and long-term sustainability [40]. With efficient parameter adaptation techniques, hospitals can deploy models on commercially available hardware without extensive computational costs [41], offering a more predictable and scalable alternative to vendor-dependent AI solutions.

Managing AI infrastructure internally also provides long-term cost visibility with predictable expenses and scalable capacity. Unlike vendor models, where costs can fluctuate based on usage, token consumption, or licensing changes, an in-house solution allows hospitals to control expenses and plan for sustainable AI adoption. Additionally, a hospital-managed approach enables deliberate and structured change management, ensuring that AI tools are thoroughly validated before deployment. By maintaining control over their models, hospitals are not exposed to the risk of model retirement, unexpected updates, or sudden shifts in vendor pricing structures. This stability is crucial for medical applications, where reliability, compliance, and continuity are essential for patient safety and operational efficiency.

3.2. Feasibility

Deploying an AI infrastructure in hospitals is increasingly feasible, provided that five core engineering elements are addressed: selecting safe and high-performing algorithms, assembling governed big data corpora, ensuring adequate computing capacity, operating within a secure deployment context, and maintaining reliable energy and operational support [42]. Implementing and maintaining AI systems demands expertise in computing resources, software integration, and regulatory compliance. However, we argue that such an infrastructure is now accessible, given recent advancements in model efficiency and the increasing availability of open-source alternatives.

Modern open-weight models can achieve performance comparable to commercial LLMs and run effectively on commercially available hardware. For instance, quantized versions of models like Llama 3 70B [43] or Mixtral 8x7B [44] can now run on consumer-grade hardware. Hospitals could invest either in a cloud or local machine with GPU capabilities. Compared to the costs of commercial API usage, this represents a predictable and potentially more economical long-term approach for high-volume usage scenarios. Hardware capable of running these models in production starts at $20 000 but can cost up to $500 000 for a server powerful enough to train or run the largest models. Most hospitals do not need training capabilities or the largest models and can therefore purchase hardware for under $100 000 to meet their inference needs.

A significant challenge for hospital-managed GenAI is the required technical expertise. We estimate that a 0.5 FTE DevOps profile is required to set up the basic infrastructure and install the required software stack needed to run a pilot in a medium–sized hospital with an EHR supporting SMART on FHIR launch. However, this gap can be addressed through partnerships with academic computer science departments, targeted hiring of AI specialists with healthcare experience, training programs for existing IT staff, participation in healthcare-specific AI consortia, and utilization of increasingly user-friendly open-source tools that reduce technical barriers. While building this expertise requires investment, it creates long-term capabilities that extend beyond any single GenAI implementation. Furthermore, as these models become increasingly integrated into healthcare operations, developing internal expertise becomes less a luxury and more a necessity for effective governance, regardless of whether models are developed internally or externally.

Despite the growing feasibility of local deployment, hospitals must navigate significant trade-offs. Talent scarcity remains a critical barrier, as institutions compete with the private sector to recruit and retain AI specialists. Resource constraints further limit the capacity to invest in infrastructure without diverting funds from core clinical services. Moreover, while internal governance processes enhance safety and accountability, they may also slow the pace of innovation relative to vendor platforms, which benefit from rapid iteration cycles. A credible deployment strategy must therefore reconcile the imperative for control and transparency with the realities of operational capacity, financial stewardship, and institutional agility. Collaborative models—such as hospital consortia or regional networks—may offer a viable path forward by pooling resources, expertise, and infrastructure to overcome these limitations.

3.3. Governance

Local deployment of LLMs in hospital environments demands a distinct governance model. Unlike vendor-hosted tools, which externalize much of the compliance and oversight burden, local LLMs place operational, ethical, and regulatory responsibility on the institution itself. To address this, we review governance recommendations for local deployments [45].

First, institutions should establish a digital committee responsible for evaluating proposed LLM use cases prior to deployment. This committee, ideally composed of clinicians, informaticians, patient representatives, and legal or ethical advisors, should assess both the intended functionality and the risk class of each application. Evaluation should consider the potential for direct clinical impact, exposure to protected health

information, and the appropriateness of the model and proposed training for end users. Retrospective and prospective studies on patients must also be reviewed by the ethics committee in addition to the data protection Office. This process aligns with existing review mechanisms used for decision support tools and embedded algorithms in EHRs [46].

Second, before any local LLM is released into clinical workflows, it should undergo a shadow deployment phase, when possible, in which the system generates outputs in real-time but remains hidden from end users. During this period, output quality, factual consistency, and task suitability can be audited retrospectively against clinician-authored notes, established gold standards, or previous iterations of the LLM application. Shadow deployments provide empirical grounding for safety and performance claims while reducing the likelihood of patient-facing harm during early-stage testing. Existing frameworks such as DECIDE-AI can be used to implement this phase [47].

Once in active use, local deployments must incorporate continuous monitoring [48]. This includes automated logging of model outputs, real-time error or anomaly detection, and audit trails for prompt modifications or system overrides. Metrics such as user correction frequency, response latency, and escalation rates should be collected and reviewed regularly, preferably on a monthly or quarterly cycle, by the digital committee. In parallel, frontline users should be provided with clear channels for feedback and error reporting, including low-friction interfaces for flagging unsafe or irrelevant outputs and defined timelines for follow-up action.

Governance also requires strict model versioning controls [49]. Every modification, whether a new model checkpoint, prompt reconfiguration, or software patch, should be recorded in a local update registry with rollback capability. This ensures traceability in the event of adverse events and supports compliance with post-market surveillance requirements under frameworks such as the EU MDR or forthcoming AI-specific regulations.

4. Conclusion

Generative AI is advancing rapidly, offering both enormous potential and significant challenges for healthcare. Our proposed hybrid framework—using vendor-based solutions for non-clinical tasks and hospital-managed AI for patient care—provides a practical blueprint to harness innovation without compromising safety or sustainability. By selectively adopting vendor offerings for lower-risk applications while building local infrastructure and expertise for clinical use cases, hospitals can retain greater control over data, adapt models to local conditions, and maintain compliance with evolving regulations. This dual approach is not only feasible but increasingly vital, ensuring that AI-driven progress remains aligned with patient well-being, clinical standards, and institutional priorities. The care providers are in the loop, interact with the models, and monitor the results. Hospitals should consider embracing this strategy as they chart a path forward in integrating generative AI responsibly.

Data availability statement

No new data were created or analysed in this study.

Acknowledgment

This work was supported by the Fondation Saint-Luc Grant Numbers 467E and the Fédération Wallonie-Bruxelles through the Fond Spécial de Recherche of Université catholique de Louvain.

Author contributions

M G wrote the main manuscript text and prepared the figure. C H provided administrative support and contributed to critical revisions of the manuscript. J V offered supervision and critically revised the manuscript for intellectual content. D Y secured funding, provided administrative oversight, and contributed significantly to manuscript revisions. All authors discussed the ideas presented, critically reviewed, and approved the final manuscript.

Conflict of interest

The authors declare that they have no conflicts of interest relevant to the content of this article.

ORCID iDs

Maxime Griot https://orcid.org/0009-0000-8180-4737
Coralie Hemptinne https://orcid.org/0000-0003-0717-3804
Jean Vanderdonckt https://orcid.org/0000-0003-3275-3333
Demet Yuksel https://orcid.org/0009-0009-0954-9781

References

[1] Warraich H J, Tazbaz T and Califf R M 2024 FDA perspective on the regulation of artificial intelligence in health care and biomedicine *JAMA* **333** 241–7
[2] Capurro D, PhD M Y, van Eaton E, Black R and Tarczy-Hornoch P 2014 Availability of structured and unstructured clinical data for comparative effectiveness research and quality improvement: a multisite assessment *Egems* **2** 1079
[3] Joukes E, Abu-Hanna A, Cornet R and de Keizer N F 2018 Time spent on dedicated patient care and documentation tasks before and after the introduction of a structured and standardized electronic health record *Appl. Clin. Inform.* **9** 46–53
[4] Pinevich Y, Clark K J, Harrison A M, Pickering B W and Herasevich V 2021 Interaction time with electronic health records: a systematic review *Appl. Clin. Inform.* **12** 788–99
[5] Jin D, Pan E, Oufattole N, Weng W H, Fang H and Szolovits P 2020 What disease does this patient have? A large-scale open domain question answering dataset from medical exams (arxiv:2009.13081) (Accessed 7 December 2023)
[6] Singhal K *et al* 2023 Large language models encode clinical knowledge *Nature* **620** 172–80
[7] Hager P *et al* 2024 Evaluation and mitigation of the limitations of large language models in clinical decision-making *Nat. Med.* **30** 2613–22
[8] Zack T *et al* 2024 Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study *Lancet Digit. Health* **6** e12–e22
[9] Johnson D *et al* 2023 Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the chat-GPT model *Res. Square* rs-3
[10] NBME 2024 United states medical licensing examination (available at: www.usmle.org/) (Accessed 19 October 2023)
[11] Saab K, Tu T, Weng W H, Lee-Messer C, Ré C and Rubin D L,2024 Capabilities of gemini models in medicine (arXiv:2404.18416)
[12] Nori H, King N, McKinney S M, Carignan D and Horvitz E 2023 Capabilities of GPT-4 on medical challenge problems (arXiv:2303.13375)
[13] Pal A, Minervini P, Motzfeldt A G and Alex B 2024 Openlifescienceai/open_medical_llm_leaderboard (available at: https://huggingface.co/spaces/openlifescienceai/open_medical_llm_leaderboard)
[14] Nori H, Lee Y T, Zhang S, Lengerich B J, Nori H, Painter I, Souter V and Caruana R,2023 Can generalist foundation models outcompete special-purpose tuning? case study in medicine (arXiv:2311.16452)
[15] Raji I D, Daneshjou R and Alsentzer E 2025 It's time to bench the medical exam benchmark *NEJM AI* **2** AIe2401235
[16] Griot M, Hemptinne C, Vanderdonckt J and Yuksel D 2025 Large language models lack essential metacognition for reliable medical reasoning *Nat. Commun.* **16** 642
[17] Sutton E, Richardson J D, Ziegler C, Bond J, Burke-Poole M and McMasters K M 2014 Is USMLE Step 1 score a valid predictor of success in surgical residency? *Am. J. Surg.* **208** 1029–34
[18] Cohen E R, Goldstein J L, Schroedl C J, Parlapiano N, McGaghie W C and Wayne D B 2020 Are USMLE scores valid measures for chief resident selection? *J. Grad. Med. Educ.* **12** 441–6
[19] Zimmer P and Kirkegaard E O W 2023 Intelligence really does predict job performance: a long-needed reply to richardson and norgate *OpenPsych* (https://doi.org/10.26775/OP.2023.02.12)
[20] Li W, Li L, Xiang T, Liu X, Deng W and Garcia N 2024 Can multiple-choice questions really be useful in detecting the abilities of LLMs? *Proc. 2024 Joint Int. Conf. on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* ed N Calzolari, M Y Kan, V Hoste, A Lenci, S Sakti and N Xue (*17 December 2024*) (ELRA and ICCL) pp 2819–34 (available at: https://aclanthology.org/2024.lrec-main.251)
[21] Griot M, Vanderdonckt J, Yuksel D and Hemptinne C 2024 Multiple choice questions and large languages models: a case study with fictional medical data (arXiv:2406.02394)
[22] Jin Q *et al* 2024 Hidden flaws behind expert-level accuracy of multimodal GPT-4 vision in medicine *npj Digit. Med.* **7** 1–6
[23] Freyer O, Wiest I C and Gilbert S 2025 Policing the boundary between responsible and irresponsible placing on the market of large language model health applications *Mayo Clin. Proc. Digit. Health* **3** 100196
[24] Muralidharan V *et al* 2024 A scoping review of reporting gaps in FDA-approved AI medical devices *npj Digit. Med.* **7** 1–9
[25] Chouffani El Fassi S, Abdullah A, Fang Y, Natarajan S, Masroor A B, Kayali N, Prakash S and Henderson G E 2024 Not all AI health tools with regulatory authorization are clinically validated *Nat. Med.* **30** 2718–20
[26] Blease C R, Locher C, Gaab J, Hägglund M and Mandl K D 2024 Generative artificial intelligence in primary care: an online survey of UK general practitioners *BMJ Health Care Inform.* **31** e101102
[27] Wu X, Duan R and Ni J 2024 Unveiling security, privacy, and ethical concerns of ChatGPT *J. Inf. Intell.* **2** 102–15
[28] Belman S *et al* 2024 Geographical migration and fitness dynamics of Streptococcus pneumoniae *Nature* **631** 386–92
[29] Fargen K M, Frei D, Fiorella D, McDougall C G, Myers P M, Hirsch J A and Mocco J 2013 The FDA approval process for medical devices: an inherently flawed system or a valuable pathway for innovation? *J. Neurointerv. Surg.* **5** 269–75
[30] Weissman G E, Mankowitz T and Kanter G P 2025 Unregulated large language models produce medical device-like output *npj Digit. Med.* **8** 1–5
[31] Tamimi N A M and Ellis P 2009 Drug development: from concept to marketing! *Nephron Clin. Pract.* **113** c125–31
[32] mrbullwinkle 2025 Azure OpenAI Service model retirements—Azure OpenAI (available at: https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/model-retirements) (Accessed 3 March 2025)
[33] Niemiec E 2022 Will the EU medical device regulation help to improve the safety and performance of medical AI devices? *Digit. Health* **8** 20552076221089079
[34] Aboy M, Minssen T and Vayena E 2024 Navigating the EU AI Act: implications for regulated digital medical products *npj Digit. Med.* **7** 1–6

[35] General Data Protection Regulation (GDPR)—Official Legal Text 2023 General data protection regulation (GDPR) (available at: https://gdpr-info.eu/) (Accessed 22 August 2023)
[36] OpenAI 2025 (available at: https://openai.com/api/pricing/) (Accessed 4 March 2025) Pricing | OpenAI
[37] Dennstädt F, Hastings J, Putora P M, Schmerder M and Cihoric N 2025 Implementing large language models in healthcare while balancing control, collaboration, costs and security *npj Digit. Med.* **8** 1–4
[38] Mathias R, Vasey B, Chalkidou A, Riedemann L, Melvin T and Gilbert S 2025 Safe AI-enabled digital health technologies need built-in open feedback *Nat. Med.* **31** 370–5
[39] Griot M, Vanderdonckt J, Yüksel D and Hemptinne C 2024 Physician in the Loop Design of Interactive Agents (Engineering Interactive Systems Embedding AI Technologies) (available at: http://hdl.handle.net/2078.1/289403)
[40] Griot M, Hemptinne C, Vanderdonckt J and Yuksel D 2024 Impact of high-quality, mixed-domain data on the performance of medical language models *J. Am. Med. Inform. Assoc.* **31** ocae120
[41] Hu E J *et al* 2021 LoRA: low-rank adaptation of large language models (arXiv:2106.09685)
[42] Lee J, Su H, Ji D Y and Minami T 2025 Engineering artificial intelligence: framework, challenges, and future direction (arXiv:2504.02269)
[43] Dubey A *et al* 2024 The Llama 3 herd of models (arXiv:2407.21783)
[44] Jiang A Q *et al* 2024 Mixtral of experts (arXiv:2401.04088)
[45] World Health Organization Ethics and governance of artificial intelligence for health: guidance on large multi-modal models (available at: www.who.int/publications/i/item/9789240084759) (Accessed 5 May 2025)
[46] Lekadir K *et al* 2025 FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare *BMJ* **388** e081554
[47] Vasey B *et al* 2022 Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI *BMJ* **377** e070904
[48] Andersen E S *et al* Monitoring performance of clinical artificial intelligence… : JBI Evidence Synthesis (available at: https://journals.lww.com/jbisrir/fulltext/2024/12000/monitoring_performance_of_clinical_artificial.2.aspx) (Accessed 5 May 2025)
[49] Garg S, Pundir P, Rathee G, Gupta P K, Garg S and Ahlawat S 2021 On continuous integration/continuous delivery for automated deployment of machine learning models using MLOps *2021 IEEE* 4th *Int. Conf. on Artificial Intelligence and Knowledge Engineering (AIKE)* pp 25–28