npj | digital medicine

News & Views

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-02312-8

Foundation model embeddings for multimodal oncology data integration

Tara P. Menon, Arjun Mahajan & Dylan Powell

Cancer care generates vast quantities of data including clinical records, pathology images, radiology scans, and molecular profiles, yet these modalities are rarely integrated in a systematic, automated manner within routine clinical workflows, remaining largely siloed across separate departmental and technical systems. Foundation model-driven embeddings—or numerical representations (vectors) that summarize complex data such as text, images, and molecular profiles — offer a framework to integrate these data streams into unified patient representations. Here we examine the HONeYBEE platform's approach to multimodal integration in oncology, situate it within broader developments in representation learning, and clinical and technical challenges that may shape its path to implementation[1].

Cancer care is increasingly driven and defined by data—radiology, pathology, genomics, and electronic health records all capture different dimensions of the same disease, yet rarely connect meaningfully. The scale and fragmentation of this data have created both an opportunity and an unmet need: how to synthesize these modalities into a unified representation of the patient. The contemporary cancer patient generates an evolving digital footprint—a dynamic record encompassing both structured data (cancer stage, performance status, and treatment history) and unstructured data (narrative patient notes capturing symptom evolution and clinical reasoning)[2]. A pathologist can review a gigapixel whole-slide image revealing a patient's histological architecture and cellular detail, while a radiologist interprets volumetric CT and MRI scans tracking tumor burden and treatment response. Genomic sequencing further extends this digital footprint to the molecular drivers of disease[3].

Yet despite the coexistence in electronic health records, these modalities are rarely interpreted together. Radiologists often review images without detailed molecular context, and pathologists assess tissue samples with limited integration of corresponding imaging or longitudinal clinical data. This separation persists at the digital level, where oncology data remain siloed across platforms, formats, and infrastructure, limiting their collective potential in precision oncology.

This fragmentation highlights a central unmet need in oncology: a unified framework capable of harmonizing heterogeneous data into a shared analytical space.

A recent article in npj Digital Medicine, Tripathi et al. propose HONeYBEE (Harmonized Oncology Biomedical Embedding Encoder), an open-source platform that helps solve these integration challenges through foundation model-based embeddings[1]. The platform generates numerical representations (vectors) from clinical text, pathological reports, radiology images, whole-slide images, and molecular profiles. The article sheds light on multimodal integration as a feasible and predictive technique for cancer classification and patient similarity retrieval without requiring a full dataset for each patient. However, it also raises questions about the platform's applicability in real-world settings. Here we examine the HONeYBEE platform's approach to multimodal integration in oncology, situate it within broader developments in representation learning, and clinical and technical challenges that may shape its path to implementation.

The data fragmentation problem

The need to integrate cancer datasets has been recognized for more than a decade. Early studies sought to link patient-level clinical outcomes with one type of complex dataset, such as gene expression profiles or pathology imaging, rather than combining multiple data types together[4,5]. Despite growing interest in multimodal and foundation model–based approaches, current multimodal AI development requires coordinating separate tools, each with distinct file formats, preprocessing requirements, and programming interfaces. For example, researchers must often rely on disparate platforms— OpenSlide for pathology images, PyDicom for radiology scans, OHDSI for structured clinical data, and Bioconductor for genomics[6]. Harmonizing metadata across these platforms, managing version dependencies, resolving identifier mapping inconsistencies, and ensuring computational reproducibility introduce overhead that slows model development and creates barriers to collaboration.

The Cancer Genome Atlas (TCGA) illustrates both the potential and challenges of multimodal cancer datasets. While it provides comprehensive, accessible data across 33 cancer types, its use requires complex workflows and modality-specific preprocessing[7]. Quality control and endpoint definitions help identify low-quality samples and genes, but these approaches are often project-specific and difficult to apply elsewhere[8].

HONeYBEE's modular architecture

The HONeYBEE framework addresses key infrastructure challenges by using a modular design that standardizes how each type of oncology data is processed[1]. It makes use of specialized foundation models that have been pre-trained on large-scale, domain-specific datasets: GatorTron for clinical text, UNI for pathology, RadImageNet for radiology, and SeNMo for molecular (genomic) data. Each data type is carefully preprocessed—such as stain normalization and subdividing pathology images or adjusting radiology scans for consistent scale—before being converted into fixed-length numerical vectors ("embeddings") by these models.

These learned representations (vectors) are stored in a format compatible with standard machine learning software, allowing them to be reused

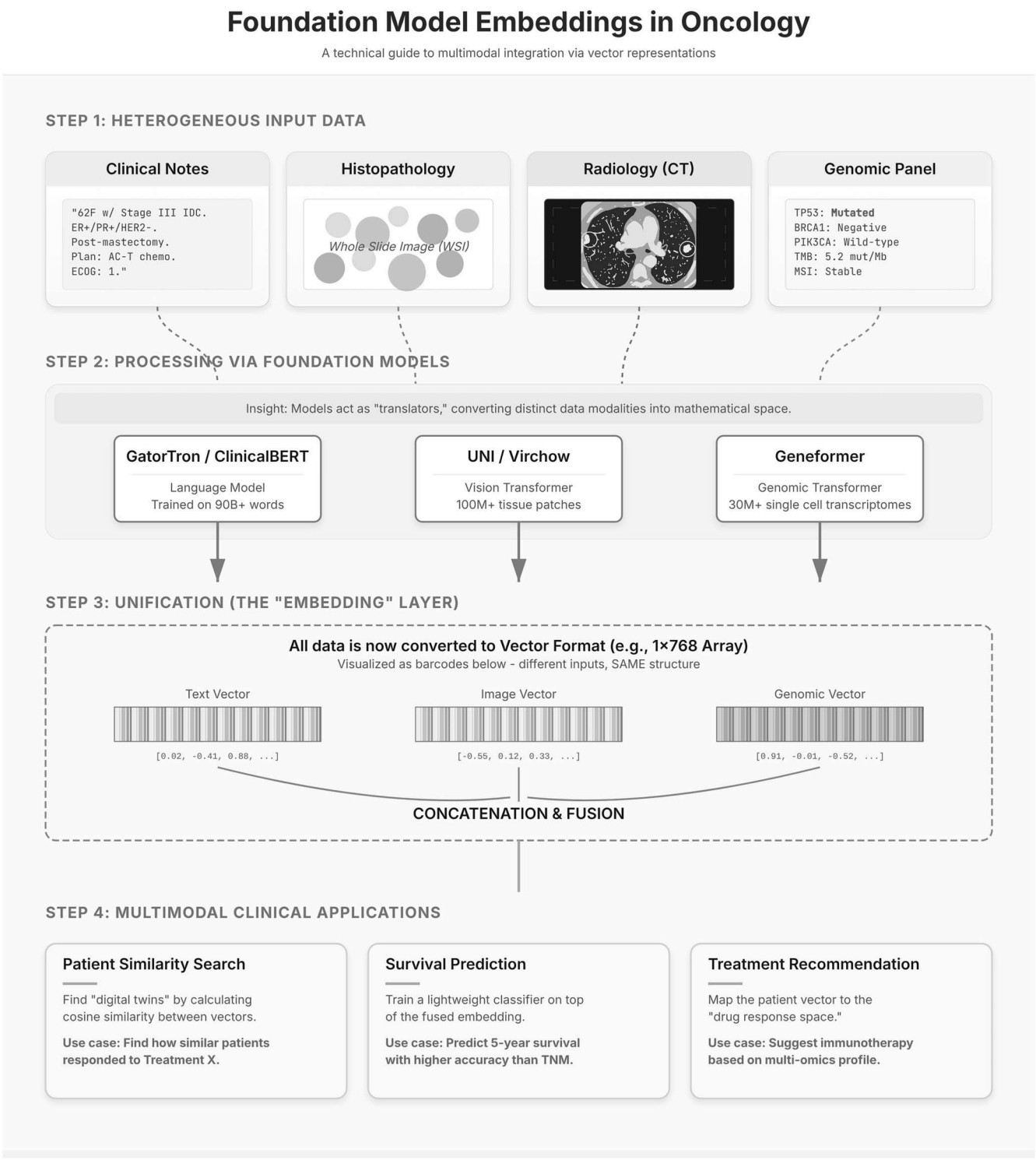

Fig. 1 | A clinician-oriented overview of foundation model embeddings in oncology.

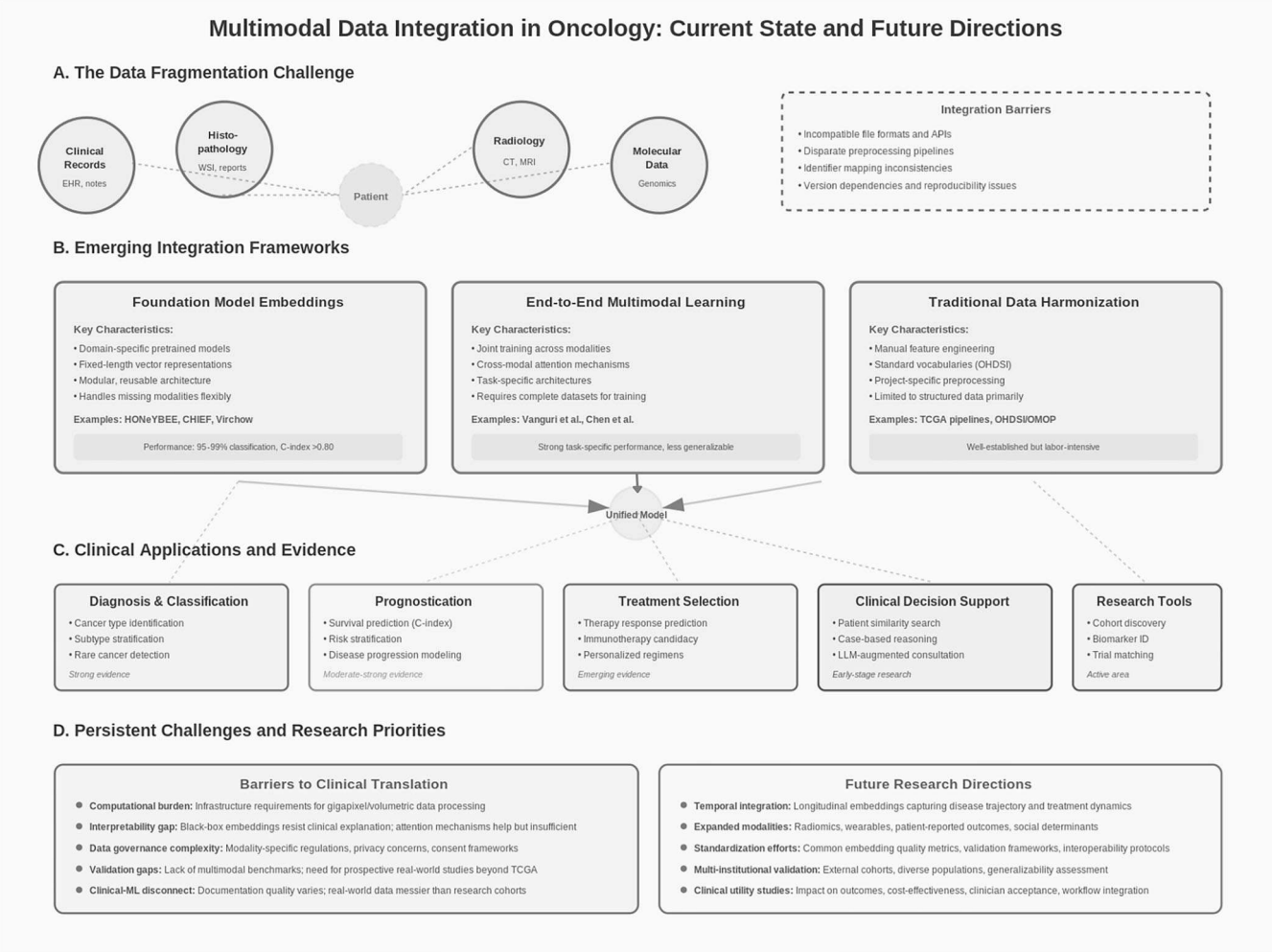

Fig. 2 | Overview of current challenges, integration frameworks, and clinical applications for multimodal data integration in oncology.

across multiple predictive tasks[1]. Because the embeddings are independent of the downstream analysis tools, newer foundation model architectures can be adopted without redesigning predictive pipelines[1]. These compact representations are also easier to share than raw datasets and are better suited to privacy requirements[1]. Importantly, the framework can accommodate patients with incomplete data by flexibly combining whatever modalities are available for analysis—a realistic reflection of clinical practice, where not every patient has every type of data available.

The clinical data dominance question

Clinical embeddings emerge as a particularly influential modality in the HONeYBEE evaluation, but this observation must be interpreted within context. The analyses reported by Tripathi et al. are conducted using TCGA, a well-curated, research-grade cohort with relatively structured and standardized clinical variables. In contrast, real-world clinical documentation is often inconsistent and fragmented across unaffiliated notes, with key details buried in unstructured text or outdated templates propagated through "copy-and-paste" practices[7,9]. In such settings, multimodal fusion may prove more valuable than TCGA results suggest, as radiology, pathology, and molecular data can compensate for missing or low-quality clinical inputs. Foundation model embeddings standardize diverse data into comparable numerical representations for joint processing[10]. Prior EHR studies demonstrate that integrating structured variables, narrative text, labs, and medications improves prediction accuracy[11]. Extending this approach to include imaging and molecular data—and allowing embeddings to update as new modalities become available—enables more realistic longitudinal analyses, particularly for rare or incompletely characterized cancers.

Implementation challenges and limitations

Several obstacles remain before embedding-based multimodal systems enter routine clinical use. Foundation models require significant computational resources for inference, particularly with gigapixel whole-slide images and volumetric scans[11]. Although embeddings can be precomputed, initial generation may exceed smaller institutions' capacity, and cloud-based processing adds costs and latency. Explainability is another barrier, as self-supervised models learn latent patterns that improve performance but limit interpretability[12]. Clinicians may find such systems opaque, especially when fused modalities defy simple explanation. To address this conceptual opacity, Fig. 1 provides a clinician-oriented schematic illustrating how foundation model embeddings transform heterogeneous oncology data into a unified representation that can support downstream clinical decision-making.

Multimodal integration further complicates data governance, as regulatory and ethical requirements differ across data types: genetic data carry distinct consent and re-identification risks, imaging data are subject to modality-specific archival and sharing standards, and clinical notes raise unique privacy concerns related to free-text identifiers[13–15]. Finally, validation demands extend beyond domain-specific standards to frameworks assessing clinical utility rather than technical accuracy[16].

Future directions

Several extensions would strengthen the clinical applicability of multimodal embedding frameworks. Incorporation of temporal information to model disease progression and response to therapy may enable a more dynamic model of patients. Incorporating additional information modalities such as radiomics (quantitative image features) can further enhance model representations and prognostic value[17]. Wearable sensor data reflecting physical activity level, sleep patterns, and physiological variables may add value to traditional clinical assessments[18,19]. Patient self-reports capturing symptoms, health-related quality of life, and physical function can also offer information not captured in clinical documentation. Lastly, standards for embeddings for quality, validation, and interoperability can make adoption easier. Just as DICOM and HL7 are established standards for image transfer and clinical messaging, respectively, a common framework could enable consistent evaluation and comparison of embeddings and support a modular ecosystem[20]. An overview of the current state of multimodal integration, persistent challenges, and potential research directions is summarized in Fig. 2.

Conclusion

Foundation model-based embeddings provide a promising technological platform through which cancer's heterogeneous data landscape can be integrated into a patient framework that can enable predictive, retrieval, and analysis capabilities. Tripathi et al. show that a foundation model-based approach can yield success in a variety of cancers and that this success can still retain a level of modularity and extensibility[1]. Nevertheless, for its successful application in cancer care, this platform must address a variety of challenges related to computational needs, limitations in interpretability and regulations, as well as validation challenges extending beyond technical performance metrics. As more advanced multimodal AI approaches are developed in this field, they have a great potential to counter fragmentation in cancer data and improve patient care.

Data availability

No datasets were generated or analysed during the current study.

Tara P. Menon[1], Arjun Mahajan[2] & Dylan Powell[3] ✉
[1]Virginia Tech Carilion School of Medicine, Roanoke, VA, USA. [2]Harvard Medical School, Boston, MA, USA. [3]Faculty of Health Sciences & Sport, University of Stirling, Stirling, UK. ✉e-mail: dylan.powell@stir.ac.uk

Received: 26 November 2025; Accepted: 21 December 2025;
Published online: 10 January 2026

References

1. Tripathi, A., Waqas, A., Schabath, M. B., Yilmaz, Y. & Rasool, G. HONeYBEE: enabling scalable multimodal AI in oncology through foundation model-driven embeddings. *npj Digit. Med.* **8**, 622 (2025).
2. Raghavan, P., Chen, J. L., Fosler-Lussier, E. & Lai, A. M. How essential are unstructured clinical narratives and information fusion to clinical trial recruitment? *AMIA* **2014**, 218–223 (2014).
3. Ghoreyshi, N. et al. Next-generation sequencing in cancer diagnosis and treatment: clinical applications and future directions. *Discov. Oncol.* **16**, 578 (2025).
4. Vanguri, R. S. et al. Multimodal integration of radiology, pathology and genomics for prediction of response to PD-(L)1 blockade in patients with non-small cell lung cancer. *Nat. Cancer* **3**, 1151–1164 (2022).
5. Mobadersany, P. et al. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc. Natl. Acad. Sci. USA* **115**, E2970–e9 (2018).
6. Reich, C. et al. OHDSI standardized vocabularies-a large-scale centralized reference ontology for international data harmonization. *J. Am. Med. Inform. Assoc.* **31**, 583–590 (2024).
7. Weinstein, J. N. et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
8. Liu, J. et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* **173**, 400–16.e11 (2018).
9. Vawdrey, D. K. et al. A practical approach for monitoring the use of copy-paste in clinical notes. *AMIA Annu. Symp. Proc.* **2021**, 1178–1185 (2021).
10. Timilsina, M. et al. Harmonizing foundation models in healthcare: a comprehensive survey of their roles, relationships, and impact in artificial intelligence's advancing terrain. *Comput. Biol. Med.* **189**, 109925 (2025).
11. Vorontsov, E. et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nat. Med.* **30**, 2924–2935 (2024).
12. Pahud de Mortanges, A. et al. Orchestrating explainable artificial intelligence for multimodal and longitudinal data in medical imaging. *npj Digit. Med.* **7**, 195 (2024).
13. Hansson, M. G. et al. The risk of re-identification versus the need to identify individuals in rare disease research. *Eur. J. Hum. Genet.* **24**, 1553–1558 (2016).
14. Bidgood, W. D. Jr, Horii, S. C., Prior, F. W. & Van Syckle, D. E. Understanding and using DICOM, the data interchange standard for biomedical imaging. *J. Am. Med. Inform. Assoc.* **4**, 199–212 (1997).
15. Negash, B. et al. De-identification of free text data containing personal health information: a scoping review of reviews. *Int. J. Popul. Data Sci.* **8**, 2153 (2023).
16. Kaczmarczyk, R., Wilhelm, T. I., Martin, R. & Roos, J. Evaluating multimodal AI in medical diagnostics. *npj Digit. Med.* **7**, 205 (2024).
17. Gillies, R. J., Kinahan, P. E. & Hricak, H. Radiomics: images are more than pictures, they are data. *Radiology* **278**, 563–577 (2016).
18. Daniore, P. et al. From wearable sensor data to digital biomarker development: ten lessons learned and a framework proposal. *npj Digit. Med.* **7**, 161 (2024).
19. Mahajan, A., Heydari, K. & Powell, D. Wearable AI to enhance patient safety and clinical decision-making. *npj Digit. Med.* **8**, 176 (2025).
20. Meskó, B. & Topol, E. J. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *npj Digit. Med.* **6**, 120 (2023).

Author contributions

T.P.M. drafted the manuscript and prepared Figs. 1 and 2. A.M. and D.P. contributed to the critical review, revision, and editing of the manuscript. All authors approved the final version of the manuscript.

Competing interests

The authors T.M. and A.M. declare no competing interests. D.P. is News & Views editor at npj Digital Medicine but played no role in the internal review or decision to publish this News & Views piece.

Additional information

Correspondence and requests for materials should be addressed to Dylan Powell.

Reprints and permissions information is available at http://www.nature.com/reprints