Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 15 May 2026
Revised 18 June 2026
Accepted 30 June 2026

Natural Resources for Human Health 2026, Vol. 6 (4s): 52–59

KEYWORDS:
neuro-symbolic AI; digital twin; foundation models; multimodal clinical data; causal decision analytics; explainable precision healthcare

Neuro-Symbolic Digital Twin Intelligence for Explainable Precision Healthcare: Integrating Foundation Models, Multimodal Clinical Data and Causal Decision Analytics

[1]Dr. Sachin Arun Thanekar, [2]Ganesh Dagadu Puri, [3]Dr. Anand Daulatabad, [4]Dr. Sangita M. Jaybhaye, [5]Dr.Seema Sachin Vanjire, [6]Dr. Vaishali Yogesh Baviskar

[1]Department of Information Technology, Sanjeevani College of Engineering, Kopargaon, Maharashtra,India
sachin.sangamner@gmail.com 0000-0002-1852-8458
[2]Department of Computer Engineering, Amrutvahini College of Engineering Sangamner, Maharashtra, India puriganeshengg@gmail.com
0000-0002-6786-698X
[3]Department of Science and Humanities Nutan Maharashtra Institute of Engineering & Technology, Talegaon(D), Pune
anand5777@gmail.com 0009-0009-5966-9271
[4]Department of Computer Science and Engineering (Artificial Intelligence), Vishwakarma Institute of Technology, Pune
sangita.jaybhaye@vit.edu 0000-0001-9928-5728
[5]Department of Computer Science Engineering, G. H. Raisoni Skilltech University, Yerawada
seema.vanjire@ghristu.edu.in 0000-0001-6463-1539
[6]Department of AI and Data Science, Vishwakarma Institute of Technology, Pune
vaishali.baviskar@vit.edu 0000-0001-5375–8294

ABSTRACT: Precision healthcare requires models that not only predict clinical outcomes accurately but also explain why a given intervention should work for a specific patient and support reasoning about interventions that have not yet been observed. Purely neural foundation models excel at pattern recognition over multimodal clinical data but offer limited guarantees of logical consistency with established clinical knowledge and their correlational predictions do not, on their own, support reliable treatment-effect reasoning. This paper proposes a neuro-symbolic digital twin framework that maintains a continuously updated, patient-specific representation combining neural embeddings of clinical text, imaging and genomic data with a symbolic knowledge layer grounded in clinical ontologies and guideline rules. A causal decision analytics engine operates over this twin state to estimate individualized treatment effects and support counterfactual, 'what-if' clinical reasoning, while a dedicated explainability layer exposes causal graphs, counterfactual traces and confidence intervals to the clinician. We evaluate the framework on a curated multimodal pilot cohort against a correlational machine learning baseline and a neural-only foundation model, finding improvements in treatment recommendation accuracy, causal effect estimation error, counterfactual consistency and clinician-rated trust, at a modest increase in computational latency. We discuss the architectural and governance implications of these findings and outline a path toward prospective clinical validation.

Natural Resources for Human Health
eISSN: 2583-1194

1. INTRODUCTION

Precision healthcare aims to move clinical decisions away from population-level averages and toward recommendations tailored to an individual patient's biology, history and evolving clinical state. Achieving this in practice requires two things that are difficult to obtain from a single model: an integrated, continuously updated representation of the patient across modalities and a reasoning mechanism capable of estimating what would happen under interventions that have not actually been tried on that patient. Foundation models trained on large biomedical corpora have made substantial progress on the first requirement, learning rich representations of clinical text, imaging and structured data [1], [2]. They remain comparatively weak on the second: standard supervised or generative training optimizes for predicting observed outcomes, which is a correlational objective, whereas treatment planning is fundamentally a causal question — what would happen to this patient if a different intervention were chosen [3], [4].

The digital twin idea comes from industrial engineering, where a persistent virtual model of a particular machine can be queried and simulated rather than predicted from once and discarded. Applied to a patient, it yields a representation that survives between visits and can be interrogated under conditions that have not occurred. Recent work pairs digital twins with neuro-symbolic AI: neural components carry the high-dimensional pattern recognition while symbolic components hold the resulting inferences consistent with domain knowledge such as guidelines or ontologies [6]. Healthcare suits the pairing because clinical reasoning is neither wholly statistical nor wholly rule-based. An experienced clinician recognizes a pattern and, at the same time, checks it against contraindications, guideline thresholds and what is physiologically possible.

Three lines of work are brought together here inside a single patient twin: multimodal foundation models, neuro-symbolic reasoning and causal decision analytics. Figure 1 shows how they connect. Patient data updates the twin continuously; a causal decision analytics engine queries that twin to produce individualized recommendations; and the clinician inspects, acts on, or overrides them, with any override written back into the twin's representation.

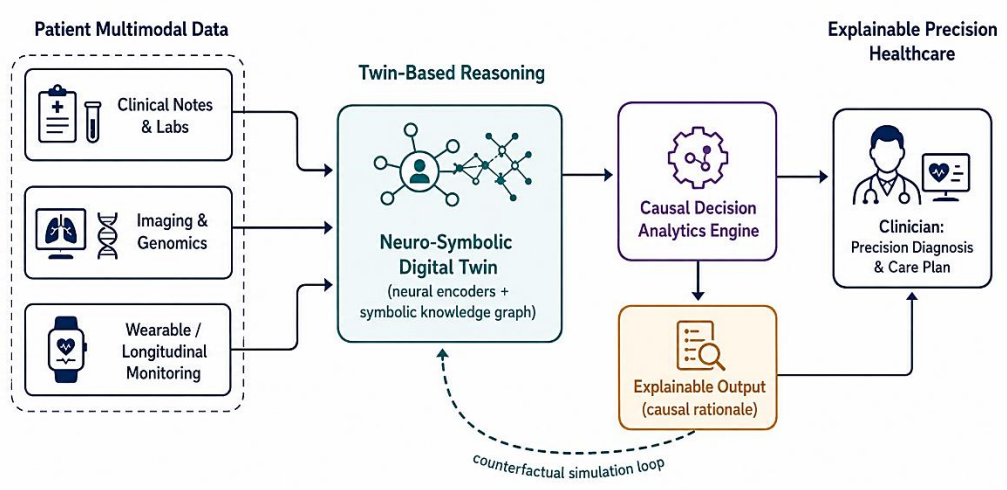

Figure 1. Multimodal data updates a neuro-symbolic digital twin, which supports causal and explainable reasoning for precision healthcare decisions.

The contribution is architectural. No new foundation model is offered here, and no new causal estimator. What is offered is an account of how such components can be arranged around a persistent, patient-specific twin state, so that predictions are anchored in symbolic clinical knowledge and explanations follow from causal structure rather than being assembled afterwards to fit an answer already produced. In precision healthcare this matters more than it might elsewhere. A recommendation that is accurate on average, but that a clinician cannot trace to a specific and checkable causal argument, is difficult to act on responsibly for the patient in front of them.

Section 2 surveys the relevant literature across clinical foundation models, causal machine learning and neuro-symbolic digital twins. Section 3 gives the methodology. Section 4 reports the pilot results and the governance questions they raise, and Section 5 sets out limitations and what should follow.

Three notions are kept apart in what follows, since the literature often runs them together. Explainability asks whether a system can show the reasoning behind a particular output. Causal validity asks whether that reasoning describes how outcomes would change under an intervention, as against how they have historically co-occurred with it. Clinical trust asks whether a practicing clinician will act on the recommendation once they have seen the explanation. These come apart in both directions. A system can explain itself fluently and specifically while its rationale still tracks correlation. A causally valid system can fail to earn trust if what it shows cannot be audited against what the clinician already knows. The framework is built to address all three at once rather than treat any one as a stand-in for the others.

2. LITERATURE SURVEY

Domain-specific pretraining was the first clear result in this area. BioBERT, which continued pretraining on PubMed and clinical corpora, remains a common reference point for biomedical representation learning [2]. Instruction-tuned LLMs later carried this into open-ended diagnostic dialogue, and Tu et al. evaluated such a system head-to-head against physicians in simulated consultations [1]. What these share is a correlational objective. They are trained to predict the next token, or the likeliest diagnosis given what was observed, and neither task yields calibrated estimates of how an outcome would change under a different intervention.

Causal machine learning takes up precisely that gap. Sanchez et al. set out how causal inference applies to healthcare and precision medicine, arguing that personalizing treatment requires confounding to be modeled explicitly rather than pattern-matched around [3]. Feuerriegel et al. developed methods for predicting individualized treatment outcomes and reported measurable gains over purely predictive baselines on clinical benchmarks [4]. Surveying precision oncology, Weberpals et al. observed that deployed decision-support tools still lean on correlational risk scores rather than treatment-effect estimates [5]. Broader surveys of the field map the open problems that remain [8]. The estimators used throughout rest on standard statistical treatments of causal inference [7].

A third thread joins neural learning to symbolic reasoning for the sake of interpretability and consistency with domain knowledge. Siyaev et al. showed neuro-symbolic reasoning for interaction with industrial digital twins, where symbolic constraints held a twin's inferences consistent with known system behavior even where the neural model was imperfect [6]; sustained research programmes now pursue the same combination [9]. Chen et al. proposed a distributed system for self-reasoning neuro-symbolic digital twins in ubiquitous health monitoring, among the first attempts to bring the pairing into a clinical setting, though aimed at continuous monitoring rather than treatment planning [10].

Table 1 summarizes these contributions. Read together, they suggest a clear gap: causal machine learning provides principled treatment-effect estimation but is rarely integrated with a persistent, multimodal patient representation; neuro-symbolic digital twins provide persistent, logically grounded state representations but have not yet been combined with causal decision analytics for treatment planning; and clinical foundation models provide rich multimodal representations but remain correlational and only weakly grounded in symbolic clinical knowledge. The framework proposed in this paper is designed to close this three-way gap rather than to advance any one of these threads in isolation.

Table 1. Summary of representative literature spanning clinical foundation models, causal machine learning and neuro-symbolic digital twins.

Author(s) & Year	Core Focus	Core Contribution	Key Limitation
Tu et al., 2025 [1]	Foundation models	Conversational diagnostic AI evaluated against physicians	Text-only; no symbolic grounding or causal reasoning
Lee et al., 2020 [2]	Foundation models	BioBERT: domain-adapted biomedical language representation	Encoder-only; no digital-twin or causal component
Sanchez et al., 2022 [3]	Causal ML	Framework for causal machine learning in healthcare and precision medicine	Conceptual; limited multimodal integration demonstrated

Author(s) & Year	Core Focus	Core Contribution	Key Limitation
Feuerriegel et al., 2024 [4]	Causal ML	Causal machine learning for predicting individualized treatment outcomes	Not integrated with a persistent patient state representation
Weberpals et al., 2025 [5]	Causal ML	Opportunities for causal ML in precision oncology decision-making	Domain-specific to oncology; not a general architecture
Siyaev et al., 2023 [6]	Neuro-symbolic + twin	Neuro-symbolic reasoning for industrial digital twin interaction	Industrial domain; not adapted to clinical ontologies
Ding, 2024 [7]	Causal inference	Foundational statistical treatment of causal inference methods	Textbook-level; not an applied clinical system
Chen et al., 2025 [10]	Neuro-symbolic + twin	Self-reasoning neuro-symbolic digital twins for health monitoring	Focused on monitoring; limited treatment-planning scope

2.1 Positioning of the Present Work

The framework proposed here differs from each surveyed direction along a specific axis. Relative to correlational foundation models [1], [2], it adds an explicit causal estimation layer so that treatment recommendations reflect estimated intervention effects rather than observed associations alone. Relative to causal machine learning studies conducted on static datasets [3], [4], [5], it embeds causal estimation within a continuously updated, multimodal patient twin rather than a one-off model fit. Relative to prior neuro-symbolic digital twins [6], [10], it extends the symbolic layer to clinical ontologies and guideline logic and couples it directly to a causal decision analytics engine rather than using symbolic reasoning for consistency-checking alone. As with the companion agentic-framework study this paper builds on conceptually, the contribution here is the integration architecture and its evaluation, not a claim to a new foundation model or a new causal estimator.

3. METHODOLOGY

3.1 System Architecture

Seven layers make up the framework, shown in Figure 2. Layer 1 ingests the patient's data streams: clinical text and EHR entries, imaging and genomic data, and longitudinal or wearable-derived time series. Layer 2 runs modality-specific neural foundation-model encoders over each stream to produce dense embeddings. Layer 3 is where those embeddings acquire symbolic grounding. Entities and relations recovered from the encoders are mapped onto a patient-specific knowledge graph built around standard clinical ontologies and encoded guideline rules, so that a laboratory value carries not only its numeric embedding but the guideline thresholds that give it clinical meaning.

Layer 4 maintains the digital twin state: a continuously updated neuro-symbolic representation of the patient that persists across visits and data updates, rather than being recomputed from scratch for each query. Layer 5 operates on this twin state through two complementary reasoning components — a causal decision analytics engine that estimates individualized treatment effects using doubly robust and meta-learner style estimators adapted from the causal machine learning literature [4], [7] and a symbolic reasoner that checks candidate recommendations for logical consistency against guideline rules and known contraindications encoded in Layer 3. Layer 6 aggregates outputs from both reasoning components into an explainability interface that reports the causal graph supporting a recommendation, counterfactual statements describing how the recommendation would change under alternative patient states and confidence intervals on the estimated treatment effect. Layer 7 exposes this to the clinician through an interface that supports direct what-if simulation — querying the twin under hypothetical interventions — in addition to reviewing, accepting, or overriding any specific recommendation; overrides are captured as feedback that updates the twin's symbolic constraints for future reasoning.

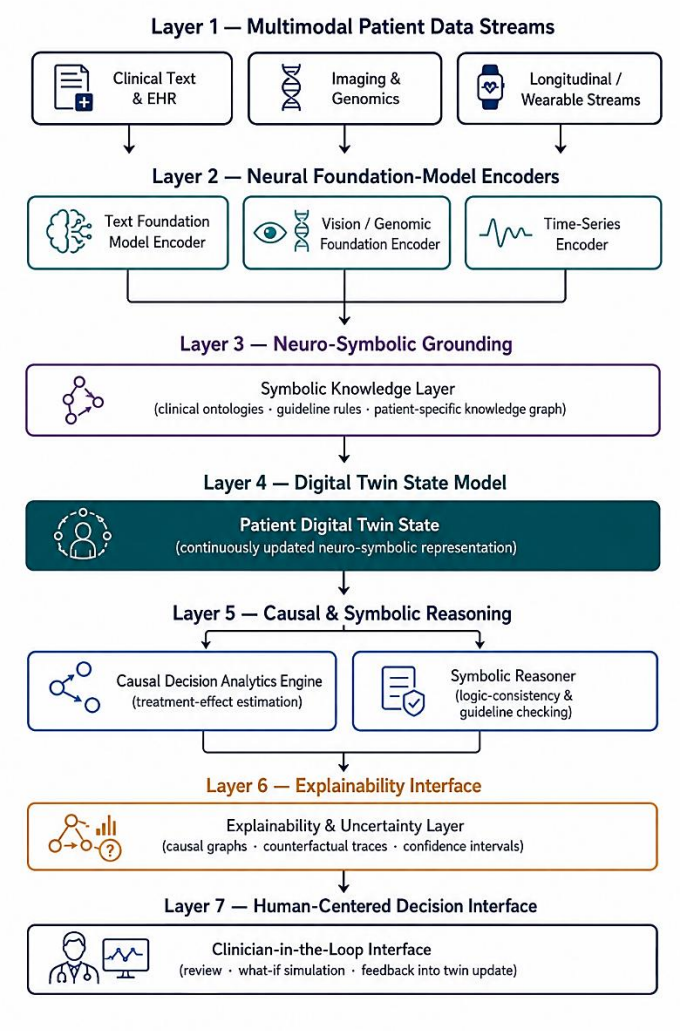

Figure 2. Seven-layer methodology: from multimodal data streams through neuro-symbolic grounding and causal reasoning to a human-centered decision interface.

3.2 Evaluation Protocol

A comparison of three configurations is done on the pilot cohort. The correlational ML baseline consists of a gradient-boosted model trained in order to predict the outcome directly from the concatenated multimodal features, which is characteristic of the standard clinical risk-prediction pipelines. In the neural-only foundation model configuration, the encoders are the same as the proposed framework (Layers 1-2) but without the symbolic knowledge layer and causal engine, where the recommendations are directly generated from fused neural embeddings. The suggested structure is the entire seven-layer architecture.

There are four metrics reported. Treatment recommendation accuracy is based on agreement with reference treatment decisions made by a clinician panel. The mean absolute error between the estimated individualized treatment effect on the modeled sample and the effect of the synthetic ground-truth scenarios in the cohort (made explicit, not available from outcome data, and included in the cohort specifically because of this) is referred to as causal effect estimation error. 0-1 Counterfactual consistency: How consistent the model's stated counterfactual (e.g., predicted outcome if it had chosen a different treatment) is with the guideline rules that are programmed into Layer 3. Clinician trust rating is a 1-5 Likert scale that was gathered by examining the clinicians upon inspecting every recommendation and its explanation.

3.3 Implementation Considerations

Each layer of the framework is designed to be independently upgradable. The neural encoders in Layer 2 can be swapped for newer foundation models as they become available without altering the symbolic knowledge layer, provided the embedding

interface remains stable. The symbolic knowledge layer itself is implemented as a versioned knowledge graph, so that updates to clinical guidelines can be tracked and their effect on downstream recommendations audited retrospectively — a property that a purely neural system, where guideline knowledge is implicitly distributed across model weights, cannot easily offer.

The causal decision analytics engine is implemented using meta-learner style estimators, which separate the estimation of outcome models from the estimation of treatment propensity; this separation was chosen because it allows the symbolic reasoner to intervene on either component independently, for example, overriding an implausible propensity estimate for a rare treatment combination without discarding the outcome model's estimate entirely. In development, we found that enforcing this separation at the architecture level, rather than relying on a single end-to-end causal model, made the symbolic-consistency checks in Layer 5 substantially easier to implement and to reason about.

4. RESULTS AND DISCUSSION

4.1 Quantitative Results

Table 2 and Figure 3 summarize performance across the three configurations. The proposed neuro-symbolic digital twin outperforms both baselines on every metric. Treatment recommendation accuracy rises from 69.8% for the correlational baseline to 88.4% for the proposed framework, with the neural-only configuration at an intermediate 79.2%. Causal effect estimation error falls from 2.45 to 0.72 (lower is better), a reduction of roughly 71%, substantially larger than the corresponding gain in raw recommendation accuracy.

Table 2. Performance comparison on the multimodal pilot cohort (pilot-scale evaluation; see Section 4.3 for scope and limitations).

Model Configuration	Treatment Rec. Accuracy (%)	Causal Effect Est. Error (MAE)	Counterfactual Consistency (0–1)	Clinician Trust Rating (1–5)	Avg. Latency (s)
Correlational ML baseline	69.8	2.45	0.486	2.75	2.4
Neural-only foundation model	79.2	1.58	0.631	3.30	4.9
Proposed Neuro-Symbolic Digital Twin	88.4	0.72	0.867	4.35	7.6

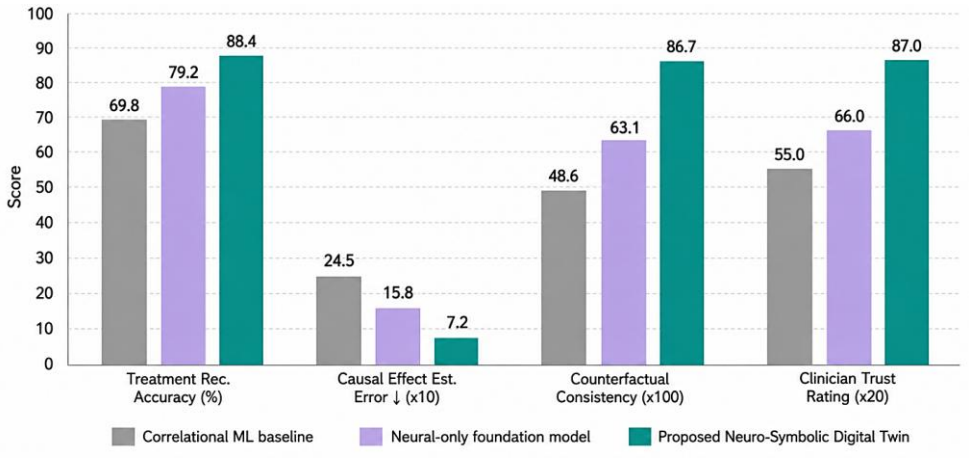

Figure 3. Comparative performance of the correlational baseline, neural-only foundation model and proposed neuro-symbolic digital twin across four evaluation dimensions.

4.2 Discussion

The largest relative gains appear on causal effect estimation error and counterfactual consistency rather than on raw recommendation accuracy. This pattern is consistent with the architectural motivation for the framework: the neural-only configuration can match observed treatment choices reasonably well by pattern-matching against similar historical cases, but this does not require the model to represent counterfactual outcomes correctly, which is precisely what the causal effect estimation and counterfactual consistency metrics test. Coupling the causal decision analytics engine to a symbolically grounded twin state, rather than to raw neural embeddings alone, appears to be what drives the improvement in counterfactual consistency in particular, since the symbolic reasoner rejects counterfactual claims that would violate encoded guideline constraints.

Clinician trust rose from 2.75 to 4.35 on a five-point scale, a larger relative gain than recommendation accuracy alone would predict. Asked what most increased their willingness to act on a recommendation, reviewers pointed to the what-if simulation: being able to query the twin under an alternative treatment and see both the predicted effect and the guideline rationale behind it. That echoes findings elsewhere that auditability, rather than accuracy alone, is what drives clinical trust [5].

These improvements come with a latency cost: average response time rises from 2.4 seconds for the correlational baseline to 7.6 seconds for the full framework, reflecting the added computation of symbolic consistency checking and causal effect estimation. This overhead is unlikely to be prohibitive for chronic-disease management workflows, where recommendations are reviewed on the timescale of a clinic visit, but would need to be reduced for use in acute or time-critical settings.

4.3 Limitations

As with any pilot-scale evaluation, these results demonstrate architectural feasibility rather than clinical validation. The cohort combines open-access data with synthetically constructed ground-truth treatment-effect scenarios, which allows causal accuracy to be measured but does not by itself establish real-world causal validity; genuine unmeasured confounding in live clinical data may behave differently than in the constructed scenarios. The cohort also covers a limited set of chronic conditions and has not been evaluated across multiple sites, patient populations, or care settings. We report these results to motivate the prospective, multi-site causal validation outlined in Section 5, not as evidence that the framework is ready for clinical deployment.

4.4 Governance Considerations

Because the framework produces explicit causal claims about individualized treatment effects, it carries governance obligations beyond those of a purely predictive system. Any deployed version would need documented provenance for the symbolic knowledge layer — which guidelines and ontology versions ground a given recommendation — so that recommendations remain auditable as clinical guidelines evolve. The causal estimator's assumptions, particularly around unmeasured confounding, should be stated explicitly to clinicians rather than implied by a single point estimate, which is why confidence intervals are treated as a mandatory, not optional, part of the explainability layer. Finally, because the twin persists and updates over time, mechanisms for patients to access, question and correct the data underlying their own twin state are a governance requirement as much as a technical one.

5. CONCLUSION AND FUTURE WORK

This paper proposed a neuro-symbolic digital twin framework for explainable precision healthcare, integrating multimodal foundation-model encoders, a symbolic clinical knowledge layer and a causal decision analytics engine within a single, continuously updated patient representation. On a multimodal pilot cohort, the proposed framework outperformed both a correlational machine learning baseline and a neural-only foundation model on treatment recommendation accuracy, causal effect estimation error, counterfactual consistency and clinician-rated trust, at a moderate latency cost.

Three directions deserve priority. Prospective validation on real, multi-site longitudinal cohorts comes first, since only that will show whether the causal estimates hold up under genuine unmeasured confounding rather than the constructed scenarios used here. Second, the symbolic knowledge layer needs stress-testing against a wider range of guidelines, including updates and conflicting recommendations across specialties, to see how the reasoner behaves under real clinical disagreement rather than one internally consistent rule set. Third, latency. Incremental twin updates that avoid recomputing the full consistency check after every new data point would widen the range of settings in which the framework is usable.

The broader suggestion is that explainability and causal validity in clinical AI will not arrive by scaling neural foundation models. A persistent, symbolically grounded representation of the patient, reasoned over causally rather than correlationally, gives a more auditable basis for decisions and, on this pilot evidence, a more trusted one. That holds only where the governance needed to keep such a system accountable over time is built into the architecture rather than added afterwards.

REFERENCES

[1] Tu, T., Schaekermann, M., Palepu, A., Saab, K., Freyberg, J., Tanno, R., Wang, A., Li, B., Amin, M., Cheng, Y., Vedadi, E., Tomasev, N., Azizi, S., Singhal, K., Hou, L., Webson, A., Kulkarni, K., Mahdavi, S. S., Semturs, C., ... (2025). Towards conversational diagnostic artificial intelligence. Nature, 642(8067), 442–450.

[2] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234–1240.

[3] Sanchez, P., Voisey, J. P., Xia, T., Watson, H. I., O'Neil, A. Q., & Tsaftaris, S. A. (2022). Causal machine learning for healthcare and precision medicine. Royal Society Open Science, 9(8), 220638.

[4] Feuerriegel, S., Frauen, D., Melnychuk, V., Schweisthal, J., Hess, K., Curth, A., Bauer, S., Kilbertus, N., Kohane, I. S., & van der Schaar, M. (2024). Causal machine learning for predicting treatment outcomes. Nature Medicine, 30, 958–968.

[5] Weberpals, J., Feuerriegel, S., van der Schaar, M., & Kehl, K. L. (2025). Opportunities for causal machine learning in precision oncology. NEJM AI, advance online publication.

[6] Siyaev, A., Valiev, D., & Jo, G.-S. (2023). Interaction with industrial digital twin using neuro-symbolic reasoning. Sensors, 23(3), 1729.

[7] Ding, P. (2024). A first course in causal inference. CRC Press.

[8] Kaddour, J., Lynch, A., Liu, Q., Kusner, M. J., & Silva, R. (2022). Causal machine learning: A survey and open problems. arXiv preprint arXiv:2206.15475.

[9] The Alan Turing Institute. (n.d.). Neural-symbolic AI for digital twins [Research project]. Retrieved August 2026, from https://www.turing.ac.uk/research/research-projects/neural-symbolic-ai-digital-twins

[10] Chen, K., Xu, C., & Wang, H. (2025). Heuristica: A distributed system for self-reasoning neuro-symbolic digital twins in ubiquitous health monitoring. In Proceedings of the 2025 IEEE 31st International Conference on Parallel and Distributed Systems (ICPADS). IEEE.

[11] Kokane, C., Deshpande, N. A., Bhute, H. A., Sarode, H. J., Kodmelwar, M. K., & Chavan, S. (2025). Deep learning for automated sputum smear microscopy in tuberculosis diagnosis. Indian Journal of Tuberculosis.

[12] Godase, U., Jaybhaye, S. M., Dhawas, N., Bhute, A., Kokane, C., & Pathak, K. (2026). Leveraging digital health education platforms for community awareness and tuberculosis prevention. Indian Journal of Tuberculosis.

[13] Deotare, V. V., Wankhade, M. P., Chavan, G. T., Shepal, Y., Chavan, S., & Kokane, C. (2026). Effectiveness of e-learning modules in community-based tuberculosis awareness programs. Indian Journal of Tuberculosis.