OPEN ACCESS

Artificial Intelligence COMMENTARY

The Innovation Informatics

Causal representation learning for medical time series: Toward mechanism-aware intelligence

Wenzhuo Wang,[1] Xueyan Lyu,[1] Xiaohui Li,[1] and Lin Wang[1,*]

[1]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China

*Correspondence: lin.wang1@siat.ac.cn

Received: June 4, 2026; Accepted: November 5, 2026; Published Online: November 5, 2026; https://doi.org/10.59717/j.xinn-inform.2026.100070

© 2026 The Author(s). This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

Citation: Wang W., Lyu X., Li X., et al. (2026). Causal representation learning for medical time series: Toward mechanism-aware intelligence. *The Innovation Informatics* 2:100070.

Medical time-series artificial intelligence is moving from task-specific prediction models toward large-scale pretrained temporal models. However, better sequence modeling does not automatically yield clinically meaningful intelligence. In observational clinical data, physiological dynamics, treatment decisions, measurement processes, and institutional workflows are tightly entangled, allowing models to achieve high predictive accuracy while relying on unstable or non-causal signals. Causal representation learning offers a promising route to address this limitation by encouraging models to learn latent patient states that are more robust, intervention-aware, and mechanistically interpretable. In this commentary, we argue that the next stage of medical time-series AI should not be defined by scale alone, but by the ability to represent evolving patient states in relation to disease mechanisms, clinical interventions, and observation processes.

FROM PREDICTIVE SEQUENCES TO PATIENT STATES

Artificial intelligence is reshaping biomedical research and clinical decision-making. With the rise of large-scale pretraining, self-supervised learning, multimodal learning, and foundation models, medical time series—including electrocardiography, electroencephalography, photoplethysmography, intensive-care-unit vital signs, laboratory trajectories, wearable sensor streams, and longitudinal electronic health records—are becoming an increasingly important substrate for clinical AI.[1,2]

Despite this progress, a central limitation remains. Many deep learning models for medical time series are still predominantly association-based: they learn statistical regularities between observed signals and clinical outcomes, often under the assumption that training and deployment data are comparable. Such models may achieve strong benchmark performance, yet predictive accuracy alone does not ensure that the learned representations correspond to disease mechanisms, physiological states, treatment effects, or causal drivers of clinical deterioration. This distinction is particularly important in medicine, where observations are irregular, interventions are frequent, data-generating processes are heterogeneous, and decisions require clinical accountability.

Causal representation learning provides a framework for addressing this gap. Rather than learning latent features that only summarize statistical associations, it aims to recover representations that reflect aspects of the underlying data-generating process.[3,4] In clinical time series, such representations should help separate latent disease severity from treatment policy, physiological change from measurement intensity, and stable patient state from institution-specific workflow. The convergence of medical time-series modeling, foundation models, and causal representation learning therefore suggests a broader transition in clinical AI: from systems that primarily predict outcomes from observed sequences toward systems that represent evolving patient states in relation to biological mechanisms, clinical interventions, and measurement processes.

BEYOND SCALE IN MEDICAL TIME-SERIES AI

The dominant trajectory of modern artificial intelligence has been characterized by scale. Medical time-series modeling is beginning to follow the same path, with increasing interest in large pretrained temporal encoders that can be adapted to diverse downstream clinical tasks.[1,2]

This trajectory has clear value for medical time-series modeling. Medical time series are highly complex, and large-scale pretraining can help models capture temporal dependencies, multiscale dynamics, missingness patterns, and cross-variable interactions. However, scale alone does not solve the problem of causality. A large model trained on large observational datasets can still learn high-dimensional correlations rather than causal structure. Indeed, when confounding, selection bias, treatment-policy variation, and measurement artifacts are present, larger models may become even more capable of exploiting shortcuts that improve in-distribution performance but fail under real-world deployment.

This distinction is particularly important for medical applications. A model may accurately predict mortality because it has learned that certain treatments are associated with poor outcomes. However, in observational data, aggressive treatment is often administered to sicker patients. Without causal adjustment, the model may encode treatment as a risk marker rather than as an intervention whose effect depends on the patient state. Similarly, a model may learn that measurement frequency, hospital identity, device type, or missingness patterns predict outcomes, even when these features reflect clinical workflow rather than pathophysiological mechanisms.

Therefore, the future of medical time-series intelligence should not be defined only by larger temporal models. It should also be defined by whether these models can learn clinically meaningful latent states, distinguish causes from consequences, separate disease mechanisms from nuisance variation, and reason about interventions. This requirement naturally positions causal representation learning as a bridge between powerful deep learning architectures and clinically meaningful intelligence.

DISTINCTIVE CHALLENGES IN MEDICAL TIME-SERIES MODELING

Medical time-series data differ from many conventional machine-learning datasets because they are partial observations of dynamic biological systems embedded in clinical practice. Their temporal structure reflects not only physiology, but also how, when, and why measurements are obtained. Variables are sampled at different frequencies across patients and institutions, and observations may be triggered by clinical deterioration, physician decisions, resource constraints, or local protocols.

Informative missingness is a direct consequence of this observation process. The absence, timing, or frequency of a laboratory test or vital-sign measurement may carry clinical meaning: unstable patients may be monitored more intensively, whereas stable patients may generate fewer observations. Treating missingness only as a preprocessing problem can therefore obscure information about both patient state and clinical workflow.

Medical time series are also shaped by intervention and feedback. Patient states influence treatment decisions, and treatments subsequently alter future patient states. Associations between treatment and outcome may therefore reflect baseline severity, treatment policy, or causal treatment effects. These interactions also contribute to distribution shift, as hospitals, devices, populations, and care pathways differ in measurement practices, treatment preferences, coding standards, and disease prevalence.

These intertwined biological, observational, and interventional processes also complicate model interpretation. Feature-attribution methods may identify variables that support a prediction, but they do not necessarily clarify whether these variables are causes, consequences, confounders, proxies, or workflow markers.

CAUSAL REPRESENTATION AS A MECHANISM-AWARE LAYER

The value of causal representation learning lies not in simply adding another modeling component, but in reorganizing how medical time-series models represent clinical dynamics. Its main contribution is to provide a mechanism-aware layer between large-scale temporal models and clinical decision-making. In this view, medical time-series models do not only encode observed sequences; they infer latent patient states shaped by physiology,

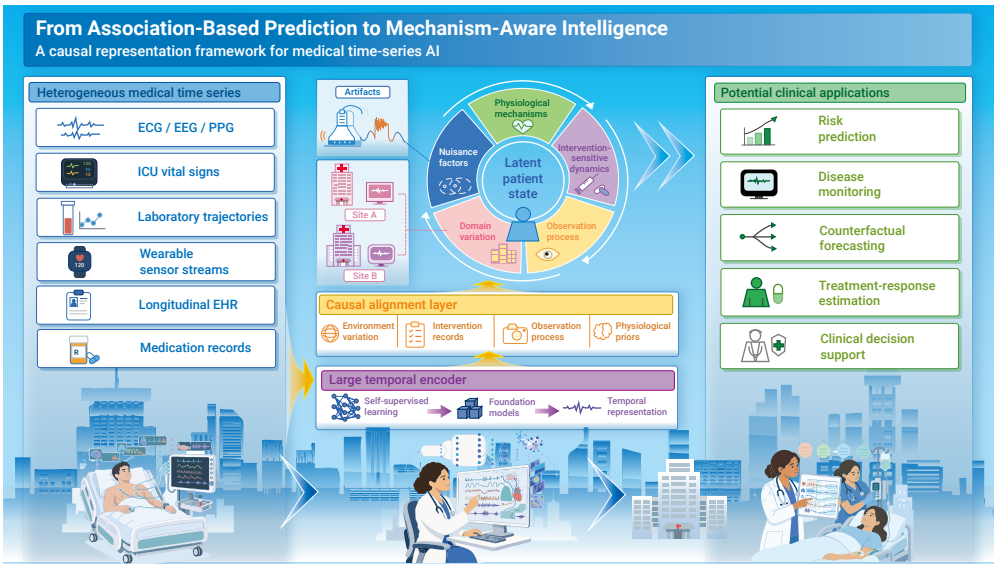


Figure 1. A causal representation framework for medical time-series AI.

interventions, observation processes, and clinical environments.

This perspective reframes a central problem in medical AI. Predictive features in observational data may reflect causes, consequences, confounders, proxies, or workflow artifacts. We define a mechanism-aware representation as a latent state $z_t = f(H_t)$, where H_t includes measurements, interventions, and observation indicators. Under explicitly stated assumptions, z_t should support a modular transition model, yield intervention-consistent changes, and separate physiological dynamics from treatment policy and measurement behavior. These are testable properties, not evidence that the true biological mechanism has been recovered.

Such a representation connects several goals that are often treated separately. If stable physiological state can be distinguished from hospital-specific workflow, models may generalize better across institutions. If treatment-sensitive dynamics can be separated from baseline severity, models may support counterfactual forecasting and treatment-response estimation. If missingness and measurement intensity are modeled as part of the observation process, models may be less likely to confuse clinical monitoring behavior with disease progression.

This framing is particularly timely in the era of medical foundation models. Large-scale self-supervised pretraining can produce powerful temporal encoders, but it does not guarantee that the latent space is causal, transportable, or intervention-aware. Mechanism-aware claims therefore require explicit inductive biases or additional information, including temporal ordering, sparse or modular transitions, interventions or natural experiments, environment variation, sampling models, multimodal consistency, or physiological priors.[3] Each source resolves a different ambiguity; none guarantees causal recovery.

FUTURE DIRECTIONS AND CLINICAL TRANSLATION

Causal representation learning offers a compelling direction for medical time-series modeling, but its promise depends on careful methodological and clinical grounding. Causal representations are difficult to identify from observational data alone. Without interventions, environmental variation, temporal structure, weak supervision, clinical priors, or other inductive biases, multiple latent structures may explain the same observed distribution.[3] Even with these biases, hidden confounding, positivity violations, timing errors, selection bias, and incorrect priors can produce false structure; methods can fail even in simple systems.[5] Claims about causal patient states therefore require explicit assumptions and evaluation strategies that go beyond standard predictive metrics.

Evaluation will be central to the development of this field. AUROC, AUPRC, and forecasting error remain useful, but they are insufficient to establish causal validity or clinical reliability. Cross-site shifts can test invariance; sampling-policy changes can test separation of observation from physiology; interventions or target-trial emulation can test response; and simulations with known mechanisms can test latent recovery. Translation should proceed from retrospective stress tests to prospective silent deployment and then impact studies, with uncertainty estimates and abstention outside supported treatment regimes.

Three CRL-specific agendas merit priority: intervention-target discovery tested with perturbational or quasi-experimental data; LLM-derived graph or temporal priors whose sensitivity is measured; and personalized causal discovery that combines population mechanisms with patient-specific parameters while reporting unsupported counterfactuals.

The goal is therefore not to presume that causal representations will directly enable reliable clinical decision support, but to determine the conditions under which they provide clinically useful information beyond conventional predictive models. Claims concerning patient mechanisms, alternative interventions, or treatment responses should be evaluated using external, prospective, and intervention-informed data, together with explicit uncertainty estimates and predefined failure criteria. Until such evidence is available, mechanism-aware representations should be regarded as testable modeling hypotheses rather than established descriptions of patient physiology or ready-to-deploy clinical decision tools.

REFERENCES

1. Lutsker G., Sapir G., Shilo S., et al. (2026). A foundation model for continuous glucose monitoring data. *Nature* **650**:978–986. DOI:10.1038/s41586-025-09925-9
2. El Ouahidi Y., Lys J., Tholke P., et al. (2025). REVE: A foundation model for EEG—adapting to any setup with large-scale pretraining on 25,000 subjects. *Adv. Neural Inf. Process. Syst.* **38**. DOI:10.52202/085713-0760
3. Li Z., Fu M., Huang J., et al. (2025). Towards identifiability of hierarchical temporal causal representation learning. *Adv. Neural Inf. Process. Syst.* **38**. DOI:10.52202/085713-0662
4. Yao D., Huang S., Cadei R., et al. (2025). The third pillar of causal analysis? A measurement perspective on causal representations. *Adv. Neural Inf. Process. Syst.* **38**. DOI:10.52202/085713-1444
5. Gamella J.L., Bing S. and Runge J. (2025). Sanity checking causal representation learning on a simple real-world system. *Proc. Mach. Learn. Res.* **267**:18143–18169. DOI:10.48550/arXiv.2502.20099

FUNDING AND ACKNOWLEDGMENTS

This work was supported in part by National Key Research and Development Program of China (2024YFE0216500), in part by National Natural Science Foundation of China (62573402). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscripts.

AUTHOR CONTRIBUTIONS

Wenzhuo Wang: Conceptualization, Writing – original draft, Visualization. Xueyan Lyu and Xiaohui Li: Writing – review & editing. Lin Wang: Supervision, Writing – review & editing. All authors contributed to the manuscript and approved the final version.

DECLARATION OF INTERESTS

The authors declare no competing interests.