Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Adaptive Temporal Latent Architecture using Structured Electronic Health Records: A Temporal Physiologic-State Centric Foundation Model

David Chen

CHEND3@ccf.org

Cleveland Clinic https://orcid.org/0000-0001-5531-9180

Xiaotan Sun

Cleveland Clinic

H.W. Wilson Tang

Cleveland Clinic https://orcid.org/0000-0002-8335-735X

Po-Hao Chen

Cleveland Clinic

Christopher Nguyen

Cleveland Clinic https://orcid.org/0000-0003-1475-2329

Xiaofeng Wang

Cleveland Clinic https://orcid.org/0000-0001-8212-6931

Deborah Kwon

Cleveland Clinic

Article

Keywords:

Posted Date: November 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11353338/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Yes there is potential Competing Interest. Xiaofeng Wang is a statistical editor for Nature Medicine

Title: **Adaptive Temporal Latent Architecture using Structured Electronic Health Records: A Temporal Physiologic-State Centric Foundation Model**

Authors:
Xiaotan Sun MS[1,2], Wilson Tang MD[2], Po-Hao Chen MD[3,4], Christopher Nguyen PhD[1,2,3], Xiaofeng Wang PhD[5], *Deborah Kwon MD[2,3], *David Chen PhD[1,3,4]

Author affiliation:
[1]Cardiovascular Innovation Research Center, Cleveland Clinic, Cleveland, Ohio, USA
[2]Heart Vascular Thoracic Institute, Cleveland Clinic, Cleveland, Ohio, USA
[3]Diagnostics Institute, Cleveland Clinic, Cleveland, Ohio, USA
[4]Center for Diagnostic Artificial Intelligence, Cleveland Clinic, Cleveland, Ohio, USA
[5]Lerner Research, Cleveland Clinic, Cleveland, Ohio, USA

*Joint supervision

Corresponding Authors:

Dr. David Chen
chend3@ccf.org
Cardiovascular Innovation Research Center
Center for Diagnostic Artificial Intelligence
Imaging Institute
Cleveland Clinic
9500 Euclid Ave, JJ5S
Cleveland, OH 44195

Abstract

Electronic health records (EHRs) contain rich longitudinal observations of patient state. Yet clinical modelling commonly reduces these data to static time, task-specific prediction problems, losing much of the rich information in the high dimensional and temporally varying patient history. Previously developed temporal models have only been demonstrated on focused tasks. Here we introduce adaptive temporal latent architecture using structured electronic health records (ATLAS-EHR), a joint-embedding predictive architecture for learning broadly reusable representations from routinely collected structured EHRs. ATLAS-EHR integrates concurrently recorded measurements without imposing within-timestamp ordering, contextualizes these states over irregular time, and learns through self-supervised prediction of masked numeric events in latent space. We trained the model on the public Medical Information Mart for Intensive Care (MIMIC)-IV dataset and demonstrated that a single frozen backbone supported seven heterogeneous clinical prediction tasks, outperforming the best task-specific tabular model on six of seven primary metrics, with a mean relative improvement of 5.0%. Cross-institutional transfer to the independent eICU dataset demonstrated similar results. We further explored ATLAS-EHR as an imputation engine to probe learned physiologic signals. ATLAS-EHR demonstrated a standardized mean squared error of 0.322 across 828 variables, compared with 0.717 for last observation carried forward with training-mean fallback. Latent-state and perturbation analyses identified clinically coherent physiological structure together with patterns related to the clinical observation process, and revealed distinct quantitative evidence profiles for chronic kidney disease and severe sepsis. These findings support the potential for ATLAS-EHR to enable real-time diagnostic and predictive capabilities across a broad range of diseases.

1. Introduction

The universal adoption of electronic health records (EHR) systems has created an unprecedented longitudinal archive of human health data, consisting of multi-channel streams of extraordinary density and complexity[1]. A single hospital admission may produce thousands of timestamped observations distributed across information systems and clinical workflows. Yet the translation of this data abundance into generalizable clinical intelligence has lagged far behind its scale[2]. Much of this information is still used through task-specific pipelines that transform longitudinal records into fixed feature tables and train a separate model for each clinical endpoint. This contrasts with recent advances in language and vision, where self-supervised pretraining has shifted model development from repeatedly engineering one-off features for specific tasks toward learning reusable representations from large-scale data. A comparable approach to structured EHRs could allow the same longitudinal patient representation to support multiple clinical questions.

Unlike language or images, EHRs consist of asynchronous and heterogeneous data streams generated across clinical, diagnostic, therapeutic, and administrative systems, each governed by distinct rules for what is documented, when it is documented, and how it is documented. EHR trajectories or "event streams"[3–5] are therefore shaped jointly by underlying physiology, clinical decision-making, and data-collection workflow. Measurements recorded at the same time may describe different aspects of a common clinical state, whereas successive observations may be separated by minutes, hours or days and individual variables may exhibit markedly different temporal dynamics[6]. Existing EHR pipelines make event streams computationally tractable through one-hot encoding, imputation[7], fixed-window aggregation[8], or naive event serialization[9,10]. These approaches impose assumptions about how observations should be aligned and summarized, and thereby partly determine what information is available for learning. Fixed windows can obscure short-timescale variation, imputation can replace informative absence with synthetic values[13], and sequential serialization can impose an arbitrary order on observations that were recorded concurrently. Such transformations may therefore simplify the record at the expense of temporal and multivariable structure that is itself clinically informative.

Structured-EHR representation learning has evolved from modelling individual variable trajectories toward learning shared representations of heterogeneous clinical event streams. Early recurrent neural network approaches[10–12] typically represented variables through separate trajectories, leading to high dimensional models which were largely sparse in the temporal direction. More recent work has developed along three complementary directions. One line of work focuses on representing the structure and scale of irregular clinical records: DuETT[13] explicitly capture interactions across temporal and variable dimensions, while EHRMamba[14] improves scalability to long event sequences. A second line uses self-supervision signals from information withheld from the record: ICU-BERT[14] reconstructs masked clinical variables and values, while masked-autoencoding approaches such as Lab-MAE[15] and AID-MAE[16] extend reconstruction to continuous measurements and incomplete EHRs. A third line of models has taken inspiration from natural language tokenization, treating events as a multi-dimensional token and performs future event prediction through next-token rollouts: MOTOR[17] uses time-to-event pretraining, whereas ETHOS[18] and ARES[19] generate future patient timelines for prediction and risk estimation. These approaches have advanced the representation, reconstruction and forecasting of structured EHRs, but continue to formulate the learning problem largely in terms of the record itself: how it is organized, what is missing from it, or what events occur next.

We take a different view of the learning problem. We hypothesized that observed clinical measurements provide partial and overlapping evidence of a shared patient state that evolves over time. Under this hypothesis, the statistical dependencies among measurements and across time provide a natural source of self-supervision: if the shared state is represented faithfully, an individual measurement should be recoverable by querying that representation with the identity of the variable being sought. This hypothesis motivates prediction in the latent space rather than direct reconstruction in the original input space. For structured EHRs, low-level reconstruction may place disproportionate emphasis on measurement noise, variable-specific value distributions and institution-specific documentation patterns. We therefore adopt a joint-embedding predictive architecture (JEPA)[20] to predict the latent representation of a withheld observation from its surrounding multivariable and longitudinal context.

A separate challenge is determining what information the representation should be allowed to learn from. EHRs contain not only measurements of patient state but also information that reflects clinical recognition, management and disposition[21]. Although highly predictive, such signals can obscure whether a model represents the evolving patient state or decisions already made in response to it[22]. We therefore asked whether reusable clinical structure could be learned from a deliberately constrained layer of routinely collected state observations while excluding these more explicit diagnostic, therapeutic and narrative signals.

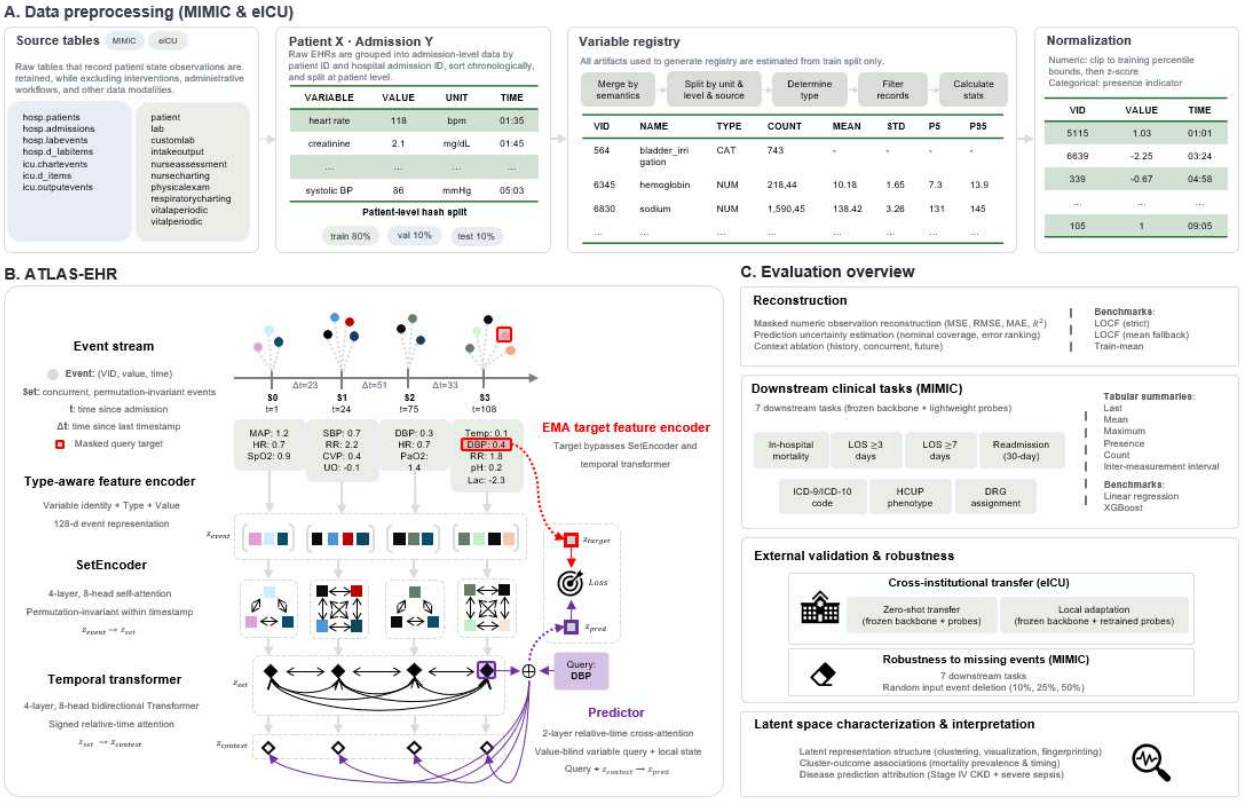


Fig. 1 | Overview of ATLAS-EHR data processing, architecture and evaluation framework. A) Structured patient-state observations from MIMIC-IV and eICU are organized into admission-level event streams and split at the patient level. A training-derived variable registry harmonizes semantically related measurements, separates variables by unit, source and level where appropriate, assigns numeric or categorical type, and estimates normalization statistics. Numeric observations are clipped to training-derived percentile bounds and standardized, whereas categorical variables are represented by presence indicators. B) ATLAS-EHR first encodes each event using variable identity, type and value to obtain an event-level representation. Concurrent

observations at each timestamp are integrated using a permutation-invariant SetEncoder, producing local timestamp-level states that are subsequently contextualized across the admission by a bidirectional temporal Transformer with signed relative-time attention. During self-supervised pretraining, selected numeric observations are masked from the online pathway. An exponential-moving-average (EMA) target feature encoder generates the corresponding event-level latent target directly from the true masked observation, bypassing the SetEncoder and temporal Transformer, while a relative-time predictor estimates this target from a value-blind variable query, the local state and longitudinal contextual representations. C) The evaluation framework tests masked-observation imputation, uncertainty estimation and contextual dependence; reuse of the frozen backbone across seven downstream clinical tasks in MIMIC-IV; cross-institutional transfer to eICU through zero-shot transfer and local adaptation; robustness to random event deletion; and latent-space analyses assessing recurrent state structure, outcome associations and disease-specific probe sensitivity.

We introduce Adaptive Temporal Latent Architecture for Structured Electronic Health Records (ATLAS-EHR), a JEPA-style foundation model for structured EHRs (Figure 1). ATLAS-EHR follows three design principles: 1) concurrent observations are represented jointly without imposing an intrinsic order; 2) successive clinical states are related over their recorded irregular timestamps; and 3) withheld observations are predicted through value-blind variable queries in latent space. During self-supervised pretraining, the model learns a multivariable and longitudinal clinical context from which the latent representation of an individual observation becomes predictable, without optimizing any downstream clinical endpoint. The resulting representation can then be queried for any supported variable, yielding a context-conditioned distribution over plausible values through a non-parametric mapping to the variable's learned latent manifold, without training variable-specific decoders.

ATLAS-EHR was evaluated in a variety of standard downstream tasks relevant to the ICU setting in both a held-out test set and an external test set. First, masked-observation imputation tested whether latent predictions retained recoverable quantitative information. We next tested whether the same frozen representation retained information useful across heterogeneous clinical endpoints. Finally, disease-specific perturbations assessed whether phenotype probes were sensitive to clinically concordant evidence. Together, these analyses move evaluation beyond asking only what an EHR representation can predict toward testing what information the representation contains, where that information arises, and how it can be reused.

2. Results

2.1 ATLAS-EHR imputes masked clinical measurements with informative uncertainty

Across the complete test cohort of 1,704,552 masked observations from 42,655 admissions, ATLAS-EHR achieved a standardized mean squared error (MSE) of 0.322, with an RMSE of 0.568, mean absolute error (MAE) of 0.320 and R^2 of 0.762 across 828 clinical variables. On the same queries, the training-set mean yielded an MSE of 1.342 and last observation carried forward (LOCF) with training-mean fallback yielded an MSE of 0.717. Among the 1,377,384 masked observations for which the same variable had been recorded at a strictly earlier timestamp within the admission, ATLAS-EHR achieved an MSE of 0.285 across 740 clinical variables, compared with 0.560 for strict LOCF (Supplementary Table 1a).

Predicted uncertainty was evaluated across the complete masked-query cohort. Empirical coverage exceeded nominal coverage for intermediate prediction intervals: the nominal 68% interval contained

85.0% of observed values, whereas the nominal 95% interval achieved 94.8% coverage (Fig. 2a). The resulting uncertainty estimates were therefore conservative over narrower intervals while approaching nominal coverage at 95% (Supplementary Table 1b).

Predicted uncertainty showed strong error-discrimination and risk-ranking ability (Fig. 2b). Mean predicted standard deviation increased from 0.280 in the lowest-uncertainty decile to 0.803 in the highest, accompanied by an increase in observed RMSE from 0.127 to 1.096. Predictions assigned greater uncertainty were therefore substantially more likely to incur large imputation errors. This suggests uncertainty prediction provided substantial discrimination between lower- and higher-error imputations (Supplementary Table 1c).

2.2 Imputation gains arise from shared context and depend on variable dynamics

The advantage over strict LOCF varied substantially across variables (Fig. 2c and Supplementary Table 2a,b). Among 312 variables meeting the prespecified eligibility criteria, 177 favored ATLAS-EHR, 68 favored strict LOCF and 67 were inconclusive. The largest ATLAS-EHR advantages were observed for several blood-gas and electrolyte measurements, including arterial and venous oxygen pressure and sodium measurements, whereas strict LOCF performed better for several enzyme measurements and device- or rate-related variables, including lipase, creatine kinase and replacement- or infusion-rate measurements.

This heterogeneity was systematically related to short-lag temporal persistence (Fig. 2d). The relative advantage of ATLAS-EHR decreased as within-admission persistence increased (Spearman ρ=-0.572). At the prespecified persistence threshold of 0.5, 133 low-persistence variables favored ATLAS-EHR and 17 favored LOCF, whereas 38 high-persistence variables favored ATLAS-EHR and 44 favored LOCF. ATLAS-EHR therefore showed its largest relative advantage among lower-persistence variables, whereas strict LOCF remained competitive for measurements with stronger short-lag persistence.

Factorial replay was then used to localize the contextual information supporting imputation (Fig. 2e and Supplementary Table 3). Removing all contextual information increased MSE from 0.319 under full context to 1.487 with the query alone. Among the three context sources, same-timestamp co-measurements made the largest average contribution to error reduction (Shapley value, 0.528), exceeding both historical (0.325) and future (0.314) context. Consistent with this result, combining history and future information without same-timestamp measurements remained substantially less accurate than combinations that included concurrent observations. Together, these analyses show that ATLAS-EHR imputation is driven by variable-specific combinations of concurrent and longitudinal information, rather than by persistence of the preceding value alone.

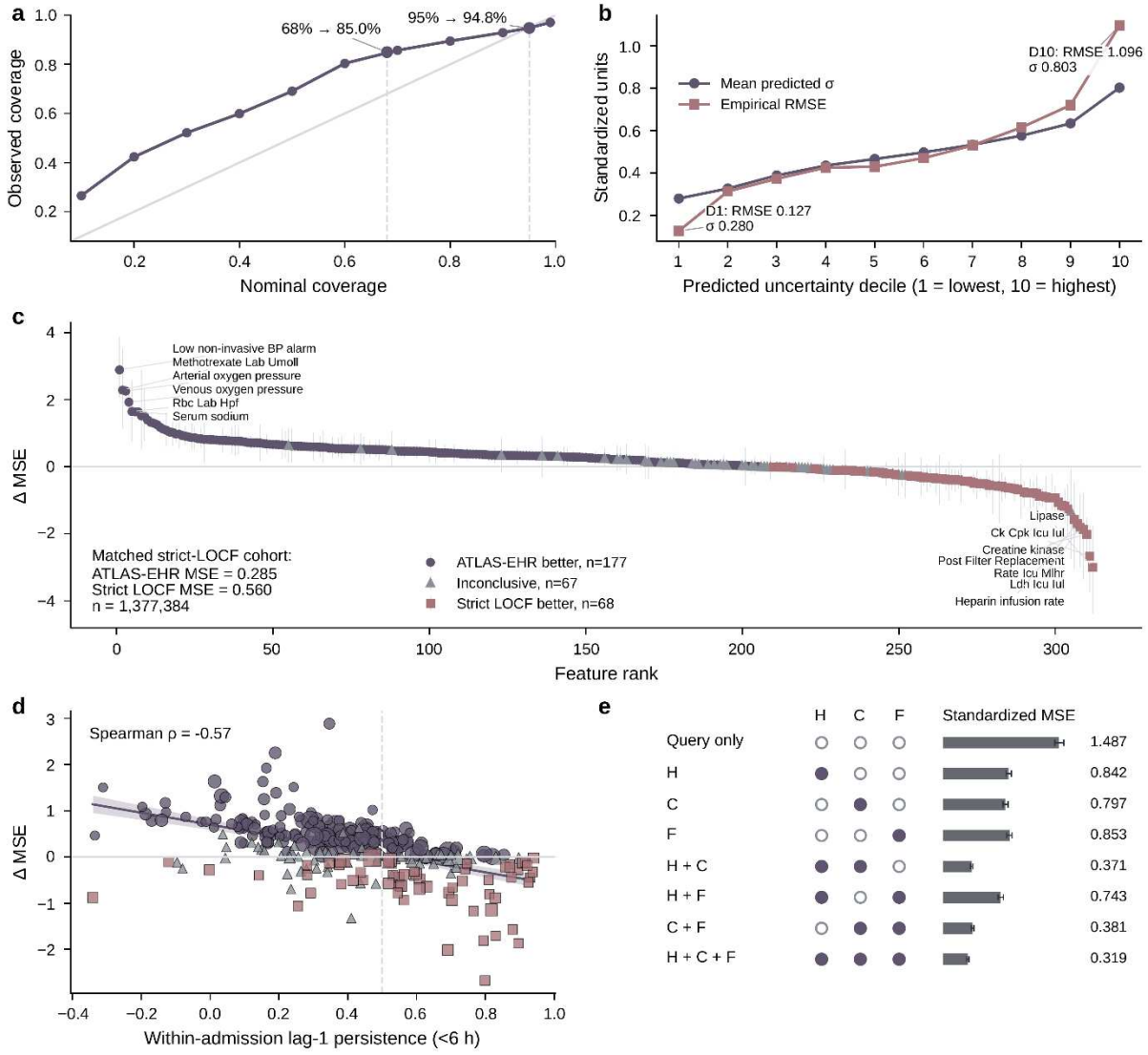


Fig. 2 | Masked-observation imputation, uncertainty and contextual dependence. a) Observed versus nominal coverage of imputation intervals across masked observations; the diagonal indicates ideal calibration. b) Mean predicted uncertainty (σ) and empirical RMSE across deciles of predicted uncertainty. c) Variable-level difference in MSE between strict last observation carried forward (LOCF) and ATLAS-EHR among 312 eligible variables. Positive ΔMSE favors ATLAS-EHR; error bars denote 95% confidence intervals. d) Relationship between ΔMSE and within-admission lag-1 persistence (<6 h); the dashed line marks persistence = 0.5. e) Factorial replay of historical (H), concurrent (C) and future (F) context. Filled circles denote included information sources; bars show standardized MSE.

2.3 A single frozen representation supports heterogeneous clinical downstream tasks

We next examined whether a single frozen ATLAS-EHR representation could be reused across seven downstream tasks spanning binary outcomes, multilabel phenotyping and structured case-mix assignment. Task-specific probes trained on the frozen representation outperformed the best logistic regression across all seven tasks and the best XGBoost in six of seven tasks on the respective primary performance metrics (Fig. 3; Supplementary Tables 4a-c).

Across the four binary outcomes, ATLAS-EHR achieved higher AUPRC than the strongest tabular benchmark for mortality, LOS ⩾3 d, LOS ⩾7 d and readmission, while also producing lower Brier scores and expected calibration errors across all four tasks (Fig. 3a). Performance across the high-dimensional phenotype and case-mix tasks was more heterogeneous (Fig. 3b). ATLAS-EHR modestly improved ICD micro-AUPRC over the strongest tabular benchmark (0.257 versus 0.250), while HCUP macro-AUPRC was slightly lower (0.519 versus 0.526), representing the only task on which XGBoost retained the higher primary metric. The clearest gain was observed for DRG assignment, where micro-F1 increased from 0.234 to 0.262, accompanied by improvements in macro-F1 and top-k hit rates. Together, these results show that information encoded by the same frozen representation was reusable across clinically and statistically distinct prediction problems, although the magnitude of benefit remained task dependent.

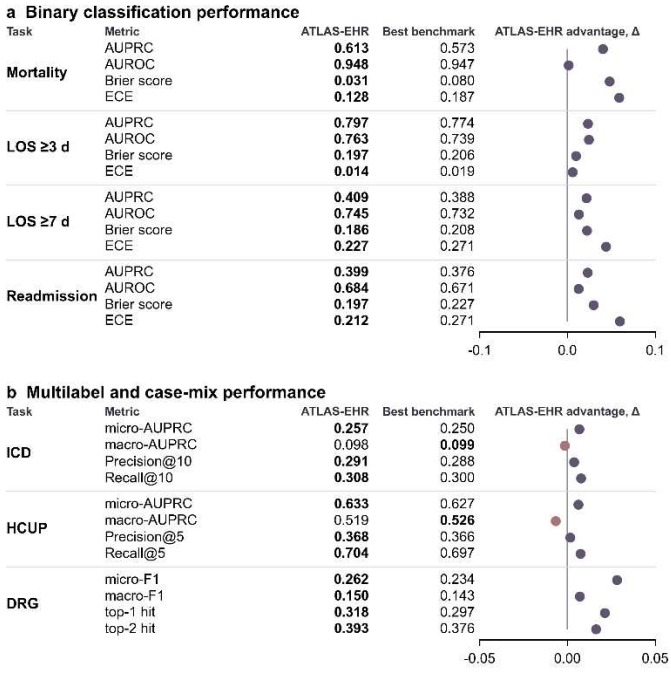


Fig. 3 | Frozen ATLAS-EHR representation across heterogeneous clinical prediction tasks in MIMIC. Performance of task-specific probes trained on the frozen ATLAS-EHR representation compared with the best logistic-regression or XGBoost benchmark. a, Binary classification performance for mortality, LOS ≥3 days, LOS ≥7 days and readmission. b, Multilabel and case-mix performance for ICD, HCUP and DRG prediction. Right-hand plots show direction-adjusted ATLAS-EHR advantage (Δ), such that positive values consistently favor ATLAS-EHR.

2.4 The pretrained representation supports cross-institutional adaptation and robustness varies by task

We evaluated whether the pretrained representations could be reused in eICU for hospital mortality and the two length-of-stay endpoints. Direct transfer of MIMIC-trained ATLAS-EHR probes was task dependent and did not outperform zero-shot XGBoost: mortality AUPRC was 0.372 versus 0.522, while the differences were smaller for LOS ⩾3 d (0.792 versus 0.807) and LOS ⩾7 d (0.386 versus 0.405) (Fig. 4a). Performance increased substantially when the backbone remained frozen and only the task-specific probe was retrained in eICU. Mortality AUPRC increased from 0.372 to 0.686, while LOS ⩾3 d and LOS ⩾7 d increased from 0.792 to 0.881 and from 0.386 to 0.514, respectively. Following local

adaptation, ATLAS-EHR achieved the highest AUPRC and AUROC for both length-of-stay endpoints and approached the performance of a fully retrained XGBoost model for mortality. These results indicate that the transferred backbone retained task-relevant information that could be recovered by locally trained probes (Supplementary Table 4d).

Robustness to random event loss was strongly task dependent (Fig. 4b). At 10% deletion, six of seven tasks retained at least 96% of their unperturbed primary performance, whereas DRG retained 90.4%. At 50% deletion, mortality and LOS ⩾3 d still retained 90.7% and 91.3% of baseline performance, respectively, while ICD retained 71.7% and DRG only 42.6%. The shared backbone therefore remained informative under substantial input loss for several clinical outcomes, although the degree of preservation varied markedly across downstream tasks (Supplementary Table 4e,f).

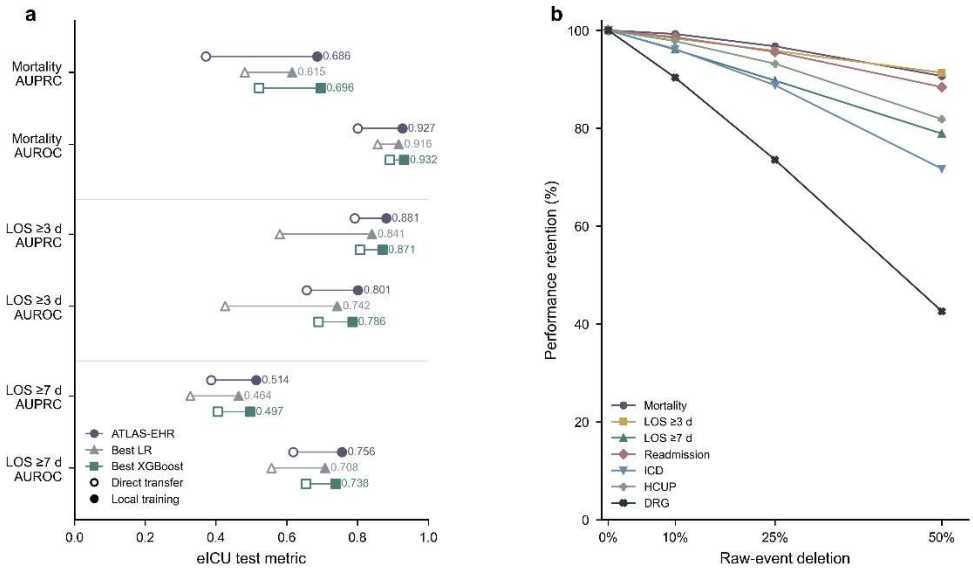


Fig. 4 | Cross-institutional reuse and robustness to event loss. a, Performance after transfer from MIMIC-IV to eICU for mortality and two length-of-stay endpoints. Open symbols denote zero shot transfer and filled symbols local training or adaptation; for ATLAS-EHR, local adaptation retrained only the task-specific probe while keeping the pretrained backbone frozen. b, Performance retention across seven downstream tasks after random deletion of 10%, 25% or 50% of raw EHR events in the MIMIC test set.

2.5 Learned representation space captures physiological and observation-process structure

Clustering of the contextual representations identified 30 recurrent states across 1,556,609 test-set timestamps, with 0.24% assigned as noise. Representative clusters showed coordinated patterns across multiple observed variables (Fig. 5; Supplementary Table 5a). C2 was dominated by elevated creatinine and urea nitrogen, consistent with a renal chemistry-associated state. C8 combined markedly reduced GCS motor response with elevated lactate, reduced arterial base excess and increased oxygen flow, capturing a distinct neurologic–metabolic disturbance. C5 combined elevated AST, CK-MB and troponin with reduced venous pH, whereas C22 was dominated by creatine kinase, AST, ALT and bilirubin. Although both C5 and C22 contained markers of tissue injury, their different combinations separated a cardiac injury and acid–base pattern from a hepatic–muscle enzyme pattern. These states therefore

represented clinically distinguishable constellations of multivariable physiological abnormalities rather than a single axis of laboratory abnormality.

The broader cluster set also contained structure reflecting how the EHR was observed and documented. C25, for example, was characterized by Braden assessments together with skin, mobility and mental-status measurements. The learned space therefore contained both clinically recognizable physiological structure and recurrent patterns of clinical observation.

Cluster membership was associated with heterogeneous clinical outcomes (Supplementary Table 5b). Admission-level mortality prevalence ranged from 1.2% to 15.1% across the 30 clusters, including 5.5–9.5% across the four representative physiological states shown in Fig. 5. The highest prevalence was observed in C27 at 15.1%, which combined older age with elevated bilirubin and abnormalities in nitrogenous waste and coagulation-related measurements.

Mortality associations also varied with the timing of state occupancy during admission. For C21, characterized by marked bilirubin and coagulation abnormalities together with broader organ-dysfunction signals, mortality prevalence increased from 4.8% among admissions occupying the state within the first 24h to 5.0% at 24–48h and 7.6% after 48h. For C1, characterized by anemia-associated profile including lower hemoglobin, hematocrit and red-cell measurements with higher RDW, mortality prevalence decreased from 5.8% to 4.8% and 3.9% across the same intervals. Thus, the relationship between latent-state occupancy and outcome depended on both state identity and when that state occurred during the admission.

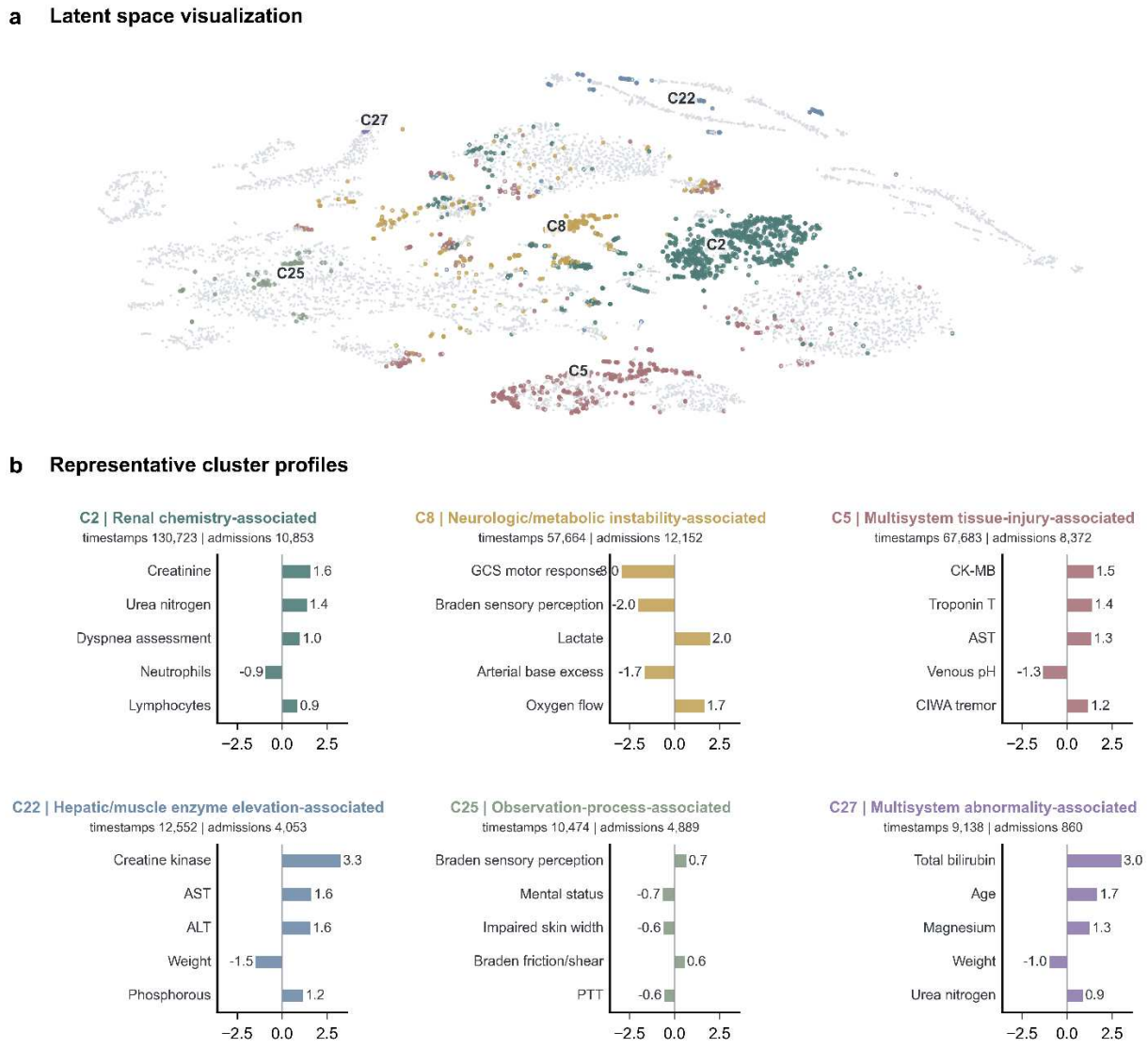


Fig. 5 | Physiological and observation-process organization of the latent space. a) Two-dimensional visualization of contextual timestamp-level representations. b) Standardized observed-variable profiles for six representative clusters, illustrating renal, neurologic–metabolic, tissue-injury, hepatic–muscle, observation-process and multisystem patterns. Cluster labels are descriptive only.

2.6 Phenotype probes show clinically concordant sensitivity to disease-specific evidence

Disease-specific perturbation analyses were performed in 258 test admissions positive for stage IV chronic kidney disease (CKD) and 406 positive for severe sepsis.

For CKD, sensitivity to patient-specific values was strongly concentrated in renal-function measurements (Supplementary Table 6a-c). Neutralizing laboratory creatinine produced the largest single-variable reduction in the CKD logit (median, 1.218), followed by urea nitrogen (0.300) and smaller effects from other renal measurements (Fig. 6a). Variable removal produced a broader and partly bidirectional profile involving renal chemistry, age, hematologic and physiological variables (Fig. 6b). In a complementary

grouped analysis of 255 evaluable admissions, joint neutralization of BMP- and CMP-related chemistry measurements produced median logit reductions of 1.711 and 1.735, respectively, with positive reductions in 99.6% of admissions for both groups. Removal of the corresponding measurement groups produced median reductions of 0.989 and 0.986, respectively.

Severe-sepsis sensitivity was more broadly distributed across physiological systems (Supplementary Table 6d,e). Value neutralization identified contributions from hemodynamic, acid–base, inflammatory and metabolic measurements without the dominant single-variable effect observed for CKD (Fig. 6c). Variable removal showed a similarly distributed profile, with the largest effects for weight and age and additional sensitivity across heart rate, blood pressure, oxygenation, respiratory, metabolic and renal measurements (Fig. 6d).

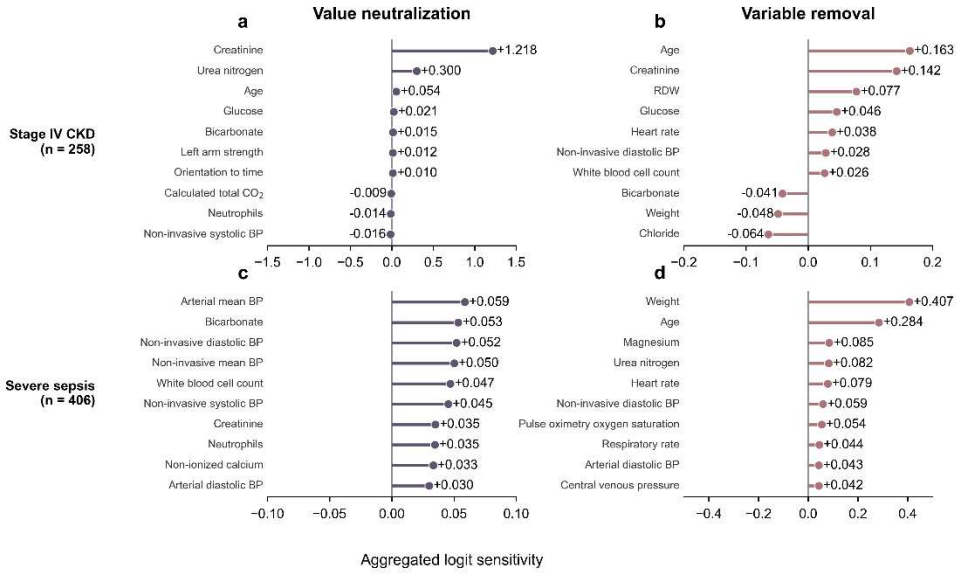


Fig. 6 | Disease-probe sensitivity to quantitative values and measurement context. Perturbation analysis of frozen probes in label-positive admissions with stage IV CKD (n = 258) or severe sepsis (n = 406). a,c) Value neutralization preserves variable occurrence and timing while replacing patient-specific numeric values with their training-set neutral reference. b,d) Variable removal eliminates all occurrences of the selected variable before re-encoding. Positive logit sensitivity indicates a reduction in the disease-specific logit after perturbation.

3. Discussion

In this study, we developed ATLAS-EHR to learn a representation of patient state which implicitly incorporates the asynchronous, high dimensionality, and temporally varying nature of EHR data. ATLAS-EHR imputed masked clinical measurements with informative uncertainty. The learned representation remained accessible to lightweight probes across heterogeneous downstream tasks and reusable in an external population after local adaptation. Its latent space contained clinically recognizable physiological structure alongside clinical workflow fingerprints, while phenotype prediction probes showed high sensitivity to clinically concordant evidence. Together, these findings position ATLAS-EHR as an information-rich representation of longitudinal structured EHRs.

A central implication is that patient state in structured EHRs is not well represented as a collection of independent variable trajectories[23]. Clinical measurements acquire meaning jointly: a laboratory or

physiological measurement is constrained by other measurements that describe the patient at that moment and by the broader trajectory in which that state occurs. The imputation analyses provide direct evidence that such contextual information was learned during self-supervision pretraining and that the value of contextual modelling is inherently variable specific: simple carry-forward remains an effective inductive bias for slowly changing measurements (e.g. heparin infusion rate), whereas contextual information becomes more useful when a variable's own trajectory provides an incomplete description of its current value (e.g. blood pressure and blood gases). This heterogeneity argues against treating EHR imputation as a uniform problem, as no single imputation strategy is likely to be optimal across variables with markedly different temporal and cross-variable dependencies. It also points to a broader role for representation learning: the shared context can complement information from variable-specific trajectories by capturing the multivariable state that makes different measurements jointly interpretable and predictable.

A distinctive property of ATLAS-EHR is that imputation becomes a query over the learned patient representation rather than a separate variable-specific modelling task. The encoded clinical context can be queried for any supported variable through non-parametric inference, yielding a context-conditioned distribution over plausible values without variable-specific parameter fitting[24] or autoregressive generation[25]. Yet the reliability of imputation necessarily depends on how informative the surrounding context is for the measurement being queried. ATLAS-EHR captures this variability through its uncertainty estimates. Our experiment showed that uncertainty estimates effectively distinguished more accurate from less accurate imputations. This provides a basis for assessing whether an inferred measurement is sufficiently supported for downstream use or whether the available information remains inadequate. For clinical use, a system that can identify when the available record provides insufficient evidence could avoid replacing missing information with unjustifiably confident estimates[26]. Uncertainty may therefore be more valuable to identify the need for updated diagnostics or quality control of risk models than to directly guide clinical decision-making[27,28].

One of the core design choices of ATLAS-EHR is to focus pretraining on observations that primarily reflect the evolving patient state. We purposely excluded diagnostic and procedural codes, medications, free-text clinical documentation, imaging interpretations, care-location and transfer information, and other variables that primarily reflect clinical interpretation, intervention or disposition. Although highly predictive, they may encode information generated downstream of clinical recognition and management, making it difficult to determine whether predictive performance arises from the evolving patient state or from actions already taken in response to it[22,29]. Despite this deliberately restricted evidence base, ATLAS-EHR demonstrated strong predictive power across both standard clinical prognostic tasks (mortality and LOS) and administrative tasks (disease code classification). This suggests greater potential in phenotyping discovery and classification, where established practices often contribute strongly to phenotype definitions[30]. This design may also improve the practical scalability of structured-EHR pretraining. Laboratory measurements, bedside observations, and structured assessments are already generated at scale by routine hospital information systems. A representation learned directly from this common data substrate could be reused across multiple clinical research questions, potentially reducing repeated task-specific feature engineering and limiting the need to continually expand the input space toward increasingly broad multimodal data.

The exclusion of explicit diagnostic and management records does not, however, make the learned representation purely physiological. Structured EHRs capture patient state through a clinical observation

process, in which care providers determine what is measured, when measurements are repeated and how intensively patients are monitored. Without explicitly training signals to separate sources of information, ATLAS-EHR nevertheless learned latent states with distinct dominant signatures: some were characterized primarily by clinically coherent renal, metabolic, neurological or tissue-injury patterns, whereas others reflected nursing assessments, monitoring intensity or measurement practice. This emergent structure makes practice-specific observation signals visible within the learned representation, providing a basis for model interpretation and auditing.

The external-validation findings define an important boundary of representation reuse across institutions. Direct reuse of MIMIC-trained task probes in eICU was substantially less effective than retaining the pretrained backbone and adapting only the task-specific readout to the local population. This suggests that the learned representation may retain cross-institutional portability even when endpoint-specific prediction functions fail to transfer. A realistic deployment architecture may therefore combine a shared representation of routinely collected structured EHRs with locally trained and validated task-specific heads rather than rely on universal zero-shot prediction. This separates reuse of the pretrained representation from the clinically consequential prediction layer, which can remain locally trained, calibrated and validated.

Disease-specific perturbations showed that ATLAS-EHR probes grounded their predictions in clinically concordant quantitative evidence. CKD prediction was organized around a concentrated renal profile[31], whereas severe-sepsis prediction drew on a broader constellation of hemodynamic, acid–base, inflammatory, respiratory and metabolic abnormalities, reflecting the different clinical expression of these conditions[32]. In CKD, creatinine provided the strongest individual signal, but the disease evidence extended across the broader biochemical profile. Joint perturbation of the closely related BMP/CMP measurements produced a larger and strikingly consistent loss of CKD evidence, indicating that clinically relevant renal information was shared across correlated measurements, allowing related observations to provide overlapping evidence of the same underlying state rather than reducing CKD prediction to a single biomarker. Moreover, neutralizing the quantitative values while preserving which measurements were obtained and when they were recorded substantially weakened prediction, arguing against a shortcut based primarily on patterns of medical attention. Complete removal exposed an additional contribution from measurement availability and surrounding context, showing that the observation process remained informative but did not account for the disease signal alone. These properties are particularly desirable for clinical AI deployment, where clinically intelligible model behavior is important for robust validation and trustworthy use[33].

ATLAS-EHR represents longitudinal EHRs as a sequence of timestamp-level latent states, providing a valuable scale for characterizing disease evolution beyond individual physiological trajectories or coarse diagnostic labels[34]. ICD-level diagnoses are too broad to capture important nuances of time-varying diseases, such as variation in severity, organ-system involvement, or phase of illness[35,36]. Patients sharing the same diagnosis may therefore exhibit different multivariable states over the course of an admission. By preserving these intermediate states along the longitudinal record, ATLAS-EHR provides finer temporal and multivariable resolution for capturing clinically important variation that remains obscured within the same coded diagnosis.

Several limitations define the scope of these conclusions. Pretraining used a single institution, external evaluation was restricted to harmonizable endpoints and variables, and local adaptation required labeled

target-institution data. Masked-observation imputation does not establish performance under naturally occurring missingness, while random deletion does not reproduce clinician- or severity-dependent observation patterns. Latent states were descriptive rather than validated phenotypes, and their outcome associations were non-causal; disease perturbations were restricted to label-positive admissions and did not establish diagnostic specificity or causal feature importance. The present study establishes a representation space but does not yet model the dynamics that govern movement through it. Explicitly modelling transitions between timestamp-level latent states could reveal distinct disease trajectories within the same diagnostic category. Incorporating clinical interventions as actions would extend this framework from describing state evolution to examining how interventions are associated with subsequent transitions and, with appropriate causal designs, how treatment effects might be estimated within the learned state space. Such a framework could connect representation learning to trajectory forecasting, treatment-effect modelling and clinical decision support.

4. Methods

This work utilized publicly available data and therefore did not need Cleveland Clinic Institutional Review Board approval.

4.1 Data

4.1.1 MIMIC-IV v3.1

ATLAS-EHR pretraining and internal validation was performed using the MIMIC-IV v3.1 database, a large public dataset comprising de-identified electronic health records for patients admitted to the Beth Israel Deaconess Medical Center between 2008 and 2019 [7]. MIMIC-IV includes comprehensive longitudinal data covering both hospital admissions and intensive care unit (ICU) stays, with detailed information on demographics, laboratory measurements, bedside observations, medications, procedures, and clinical outcomes.

In this study, only time-stamped EHRs that directly describe patient state were used: (1) *hosp.patients*, to obtain admission-time demographics; (2) *hosp.admissions*, to define encounter timing and outcome windows; (3) *hosp.labevents* and (4) *hosp.d_labitems*, for laboratory measurements and item definitions; (5) *icu.chartevents* and (6) *icu.d_items*, for ICU bedside measurements and item definitions; (7) *icu.outputevents*, for output measurements; (8) *icu.inputevents*, used only for body weight; (9) *hosp.diagnoses_icd* and (10) *hosp.drgcodes*, aggregated at admission level as diagnostic/billing label metadata (Supplementary Table 7a).

Interventions, administrative records, delayed-result measurements, socioeconomic factors, post-discharge data, images and free-text reports were excluded from this study, as they are beyond the scope of the current work.

4.1.2 eICU-CRD v2.0

External validation was performed using the eICU Collaborative Research Database (eICU-CRD) v2.0, a large, multicenter database containing de-identified electronic health records from more than 200,000 ICU stays involving over 139,000 patients admitted during 2014–2015[37]. The database comprises data

collected from 335 ICUs across 208 hospitals throughout the United States through the Philips eICU telehealth program.

Specifically, the following tables were used: (1) *patient*, to obtain patient demographics, hospital and ICU encounter timing, and clinical outcomes; (2) *lab* and (3) *customlab*, for laboratory measurements; (4) *intakeoutput*, for recorded fluid-output and other physiological measurements; (5) *nurseassessment*, (6) *nursecharting*, and (7) *physicalexam*, for structured bedside assessments and observations; (8) *respiratorycharting*, for respiratory measurements and assessments; and (9) *vitalaperiodic* and (10) *vitalperiodic*, for intermittently and continuously recorded vital signs, respectively (Supplementary Table 7b).

Multiple ICU stays belonging to the same hospital admission were consolidated into a single hospital-admission-level trajectory, with timestamps expressed relative to hospital admission. Interventions, medications, procedures, administrative records, delayed-result measurements, socioeconomic factors, post-discharge data, images, and free-text documentation were excluded.

4.2 Preprocessing

Admission data were split deterministically at the patient level using a subject-hash 80/10/10 train/validation/test partition to avoid leakage across admissions from the same individual. Preprocessing artifacts, including variable registry construction, filtering thresholds, clipping bounds and standardization statistics were estimated from the training split only.

Candidate variables were first profiled in the training split and retained after frequency filtering. Variable names were then normalized through a mapping pipeline that combined semantic consolidation using pretrained BioClinicalBERT[38] embedding space similarity with source- and polarity-aware disambiguation and unit-conversion. Admissions were filtered for quality, retaining admissions with more than one timestamp and excluding admissions above the 98th percentile of timestamp count. A MIMIC variable registry (Supplementary Table 7c) was then derived: numeric variables were assigned robust clipping bounds based on the 5th and 95th percentiles and standardized using trimmed-distribution means and standard deviations, whereas others were registered as categorical variables. Registry constraints were subsequently applied to remove low-support variables, convert zero-variance numeric variables to categorical items, and assign unique vocabulary identifiers. To ensure compatibility with the MIMIC, only eICU variables that could be mapped to the corresponding MIMIC-IV variable registry were retained (Supplementary Table 7d).

Each event was then mapped to the variable registry and presented as a tuple *(vid, value, time)*. Here, *vid* denotes the variable identifier, *value* is either the standardized value for numeric variables or a presence indicator for categorical variables, and *time* is minutes since admission. Mapped events were assembled in chronological order to create model-ready event streams. Age and sex (demographics) were extracted from their first occurrence within each admission and broadcast to all timestamps.

Post-processed data descriptions are provided in Supplementary Table 7e,f. Full preprocessing pipeline and flowchart can be found in our repository.

4.3 Tokenization

Each event was passed through a type-aware feature encoder. The event representation was jointly defined by the identity and type of the corresponding variable and, for numeric variables, was further modulated by its standardized value. For a numeric variable with standardized value x, the value was expanded into $[x, x^2, sign(x)]$ to provide the unsigned quadratic magnitude and an explicit indicator of direction relative to the reference distribution. The pre-normalization numeric representation was defined as

$$z_i^{num} = e_i + e_{num} + b_i^{num} + \sigma(g_i) \odot MLP(Concat[x, x^2, sign(x)])$$

where e_i denotes the variable identity embedding, e_{num} is the learned numeric-type embedding, b_i^{num} is a variable-specific additive bias in the numeric encoding pathway, g_i is the variable-specific gating vector, σ denotes the sigmoid function, and $\odot$ denotes element-wise multiplication.

For categorical variables, the pre-normalization event representation was formulated as

$$z_i^{cat} = e_i + e_{\text{cat}} + b_i^{pres}$$

where e_{cat} is the learned categorical-type embedding and b_i^{pres} is a variable-specific additive bias applied in the categorical observation pathway. The type-specific representations were subsequently processed by the feature encoder's output normalization and an additional fusion normalization. Representations at padded positions were set to zero and excluded from subsequent within-timestamp set encoding.

4.4 ATLAS-EHR pretraining

ATLAS-EHR adopted a joint-embedding predictive architecture (JEPA) comprising an online encoder[20], an exponential-moving-average (EMA)-updated target pathway, and a latent predictor. During self-supervised pretraining, selected numeric events were removed from the online context pathway before feature encoding, within-timestamp aggregation, and temporal contextualization. Consequently, neither the identity nor the value of a masked event was available to the online context encoder; its identity was introduced only subsequently through the value-blind predictor query. Categorical events remained as visible contexts and were not used as prediction targets. The target encoder was not optimized directly by gradient descent; its parameters were updated after each optimizer step as EMA copies of the corresponding online modules, using a decay coefficient of 0.996.

4.4.1 Online hierarchical encoder

ATLAS-EHR operated at the irregular timestamp level of the processed EHR sequence and treated each retained unique timestamp as a local physiological-state anchor. Events recorded at the same timestamp were represented as an unordered set, thereby enforcing permutation invariance to arbitrary within-timestamp event ordering introduced by database layout, extraction, or serialization. The SetEncoder applied four layers of multi-head self-attention to exchange information among concurrently observed events. The resulting event representations were summarized using both learned gated pooling and masked element-wise maximum pooling. The two pooled representations were concatenated, linearly projected, and normalized to produce a timestamp-level representation, z_{set}.

The sequence of timestamp-level representations was subsequently processed by a four-layer bidirectional temporal Transformer. Temporal information was incorporated through pairwise signed differences between query and key timestamps rather than through absolute positional embeddings. For a query timestamp t_q and key timestamp t_k, the interval $\Delta_{qk} = t_k - t_q$ was transformed into

$$\tilde{\Delta}_{qk} = sign(\Delta_{qk}) \log(1 + |\Delta_{qk}|)$$

and passed through a learned relative-time embedding network. For each attention head, the transformed interval embedding contributed a query-dependent additive term to the query–key attention logit. Consequently, temporal attention was jointly determined by latent content similarity and the signed magnitude of the interval between timestamps.

The temporal encoder produced a contextual update for each timestamp in the sequence:

$$c_{1:T} = TemporalEncoder(z_{1:T}^{set}, t_{1:T})$$

where c_t denotes the temporal encoder output at timestamp t. This contextual update was combined residually with the corresponding local timestamp representation and normalized:

$$z_t^{context} = LN_{context}(z_t^{set} + c_t)$$

4.4.2 Frequency-guided masking

To reduce domination by frequently observed variables, ATLAS-EHR used a frequency-guided masking policy estimated exclusively from the training split. For each non-static numeric variable v, rarity was calculated from its log-transformed training-set event count c_v:

$$r_v = 1 - \frac{\log(1 + c_v) - min_{j \in V} \log(1 + c_j)}{max_{j \in V} \log(1 + c_j) - min_{j \in V} \log(1 + c_j)}$$

The corresponding event-level masking probability was defined as

$$p_v = clip(\mathrm{s}(1 + \gamma r_v), p_{min}, p_{max}),$$

where $\mathcal{V}$ denotes the set of non-static numeric variables, $\gamma = 0.40$ controls the strength of rarity weighting, s is a shared scaling factor, $p_{\min} = 0.05$ and $p_{\max} = 0.50$. The designated static numeric variable was excluded from rarity-based weighting and assigned a fixed masking probability of 0.15. The shared scaling factor s was selected so that, after accounting for both static and non-static numeric variables, the expected proportion of masked numeric events in the training distribution equaled 15%. Each eligible event was then independently sampled from a Bernoulli distribution with probability p_v.

4.4.3 Masked-event prediction and latent objective

Each masked numeric event was represented by a value-blind query derived solely from the shared variable identity embedding:

$$q_i = e_i + MLP_{query}(e_i)$$

Each query was associated with the timestamp of its masked event, which was subsequently used to compute relative-time biases during cross-attention. Specifically, the variable query q_i was combined with the local timestamp representation z_t^{set}, together with their element-wise interaction and difference:

$$q_{t,i}^{local} = MLP_{local}(Concat\left[q_i, z_t^{set}, q_i \bigodot z_t^{set}, q_i - z_t^{set}\right])$$

Here, q_i specifies the identity of the variable to be predicted, whereas z_t^{set} summarizes the other visible events at the same timestamp. The product and difference terms supplied the local fusion network with explicit element-wise interaction and displacement features between the variable query and the local timestamp representation. The resulting $q_{t,i}^{local}$ represented a variable-specific query conditioned on the available local state. It was subsequently processed by a two-layer relative-time cross-attention predictor over $z_{1:T}^{context}$. Masked-event queries were processed independently, without query-to-query self-attention. The predictor produced one estimated latent representation, $z_{t,i}^{pred}$, for each masked numeric event.

The prediction target was defined at event-level rather than as a contextual representation of the complete unmasked trajectory. Specifically, the variable identity and standardized observed value of each masked event were passed through the EMA target feature projector and target fusion normalization to produce a time-free, pre-SetEncoder target representation $z_{t,i}^{target}$. Training minimized the elementwise L1 distance between each predicted event representation and its stop-gradient EMA target:

$$L = \frac{1}{|\mathcal{M}|D} \sum_{(t,i)\in\mathcal{M}} \left\| z_{t,i}^{pred} - sg(z_{t,i}^{target}) \right\|_1$$

where $\mathcal{M}$ denotes the set of masked numeric events, D is the latent dimension, and $sg(\cdot)$ denotes the stop-gradient operation. During gradient accumulation, the elementwise L1 sums were accumulated across all micro batches in the accumulation cycle and normalized by the corresponding total number of masked latent elements. This avoided weighting micro-batches equally when they contained different numbers of masked targets.

4.4.4 Configuration and optimization

The model used 128-dimensional representations, eight attention heads, four within-timestamp SetEncoder layers, four temporal-context layers, and two predictor cross-attention layers. Training used AdamW with a maximum learning rate of 1×10^{-4}, weight decay of 0.02, 3,000 linear warm-up steps, and cosine learning-rate decay to a minimum of 1×10^{-6}. Models were trained for up to 25 epochs using bfloat16 mixed precision, a microbatch size of 16 admissions, and gradient accumulation over two microbatches, corresponding to a nominal single-process effective batch size of 32 admissions. Gradients were clipped to a global norm of 0.5. Checkpoints were selected according to validation loss, and early stopping was enabled after a minimum of 12 epochs with a patience of eight validation epochs.

4.5 Evaluation design

ATLAS-EHR evaluations used the frozen pretrained backbone. Model selection and hyperparameter selection were performed using the training and validation splits only, and the held-out test split was reserved for final evaluation.

Evaluation was organized across complementary levels of analysis. We first assessed masked numeric imputation and context dependence, followed by downstream clinical prediction tasks, cross-hospital transfer and robustness to reduced observation density. We then characterized latent-state organization and examined disease-specific prediction attribution using targeted perturbation analyses.

4.5.1 Masked observation imputation and uncertainty estimation

Imputation was evaluated on observed numeric events withheld under the masking protocol. The analysis therefore assessed imputation of deliberately masked observations rather than naturally missing clinical measurements. All primary imputation errors were calculated in the standardized preprocessing space.

Numeric values were recovered from z_{pred} using a non-parametric target-manifold readout derived from the EMA teacher. For each numeric variable, the minimum and maximum standardized values observed in the training split were used to define 513 evenly spaced candidate values. Each candidate was passed through the frozen teacher numeric feature pathway to generate a variable-specific trajectory in target-latent space. The manifold bank was therefore determined only by the training data and the pretrained teacher and was fixed before validation and test evaluation. For each masked query, the predicted target latent z_{pred} was compared with the candidate latent points for the corresponding variable using mean absolute latent distance. These distances were converted to SoftMax-normalized weights using a global temperature parameter. The weighted mean of the candidate values was used as the imputed value, and their weighted dispersion was used as the initial uncertainty estimate. A single multiplicative scale factor was subsequently applied to the uncertainty estimates. The temperature was selected by minimizing the macro-average of variable-specific validation MSE among variables with at least ten validation queries, and the uncertainty scale was subsequently selected by minimizing pooled validation Gaussian negative log-likelihood. Both parameters were fixed before test evaluation. No supervised numeric decoder was fitted for the primary imputation analysis.

The training-mean reference assigned each query the mean value of the corresponding variable estimated from the training split. Two definitions of last observation carried forward (LOCF) were distinguished: strict LOCF used the most recent observation of the same variable from a strictly earlier timestamp within the same admission; LOCF with training-mean fallback used the strict LOCF value when available and otherwise assigned the variable-specific training mean. Accordingly, two imputation cohorts were analyzed and reported: the all-query cohort contained all masked numeric test queries and was used for overall ATLAS-EHR performance, training-mean comparison, LOCF-with-fallback comparison and uncertainty analysis; the strict-LOCF cohort contained only queries with an available strictly earlier same-variable measurement. Direct comparison between ATLAS-EHR and strict LOCF was restricted to this identical paired subset. The test split contained 1,704,552 all-query observations, of which 1,377,384 had an available strict LOCF value. Imputation performance was summarized using MSE, RMSE, MAE and R^2. For paired comparisons, differences were reported as reference MSE minus ATLAS-EHR MSE. Paired MSE differences were summarized with 95% confidence intervals estimated from 1,000 admission-clustered bootstrap replicates, preserving dependence among multiple masked observations from the same admission. Uncertainty calibration was evaluated on the all-query cohort using Gaussian negative log-likelihood and empirical coverage of nominal prediction intervals. To assess error discrimination, masked queries were ranked by predicted uncertainty, divided into ten equally sized groups, and RMSE was calculated within each group.

4.5.2 Variable-level imputation and context dependence

Variable-level comparisons were performed on the strict-LOCF cohort. Variables were eligible when at least 100 paired queries from at least 20 admissions were available. ATLAS-EHR and strict LOCF were evaluated on the same observations for each eligible variable, with MSE as the principal comparison measure. Confidence intervals were estimated using 1,000 admission-clustered bootstrap replicates. Statistical evidence for paired differences was assessed using 10,000 admission-level sign-flip permutations, with Benjamini–Hochberg correction across eligible variables. A variable was classified as favoring ATLAS-EHR when the false-discovery-rate-adjusted -value was at most 0.05 and the bootstrap confidence interval for the paired MSE difference was entirely positive. The converse criterion defined variables favoring LOCF; all remaining variables were classified as inconclusive.

Temporal persistence was estimated from successive within-admission observations of the same variable after within-admission centering. Adjacent observation pairs were stratified by elapsed time into ≤6h, 6–24h and >24h intervals, with the ≤6h estimate used for the primary analysis. Within each lag stratum, an admission contributed only when at least two qualifying pairs were available, and variable-level estimates required at least 20 pairs from at least five admissions. Confidence intervals for persistence estimates were obtained from 1,000 admission-level bootstrap replicates. The association between short-lag persistence and the LOCF–ATLAS MSE difference was summarized using Spearman correlation. For the descriptive quadrant analysis, variables with short-lag persistence below 0.5 were classified as low-persistence and those with persistence of at least 0.5 as high-persistence.

A factorial context replay was used to examine the information contributing to masked imputation. Three parts of the admission sequence were independently retained or removed: historical context (H), non-target measurements recorded at the query timestamp (C), and future context (F). The masked target remained unavailable in all conditions. All eight combinations of H, C and F were evaluated on the same capped set of 50,000 test queries using the frozen ATLAS-EHR backbone and the same target-manifold readout. Contributions of the three context sources were summarized using the exact three-factor Shapley decomposition of squared-error reduction. Paired condition differences and mean Shapley contributions were summarized with 95% confidence intervals from 1,000 admission-clustered bootstrap replicates. Future context was used only in this retrospective masked imputation analysis and was not available to prospective cutoff-based prediction tasks.

4.5.3 Downstream clinical tasks

Frozen ATLAS-EHR representations z_{set} and z_{context} were evaluated on seven admission-level prediction tasks: in-hospital mortality, LOS ≥ 3 days, LOS ≥ 7 days, 30-day readmission, ICD code prediction, HCUP phenotype-group prediction and DRG assignment. Mortality used observations within the first 48h after admission, both LOS tasks used the first 24h, and readmission, ICD, HCUP and DRG used the full visible admission trajectory. Task-specific cutoffs were applied before representation extraction and probe training. For cutoff-based tasks, contextual representations were recomputed using only timestamps within the corresponding observation window.

Task-specific probes were trained separately for z_{set} and z_{context}. The z_{set} probe applied a single relative-time Transformer layer before admission-level pooling, whereas the z_{context} probe used a single standard Transformer layer without additional timestamp encoding. Both probes used 256-dimensional

hidden representations and four attention heads, followed by learned multi-query attention pooling. Four pooling queries were used for binary tasks and eight for ICD, HCUP and DRG. The pooled representations were passed through a single hidden projection and a task-specific output head; DRG used separate label and cardinality heads for competitive decoding.

ICD and HCUP were formulated as multilabel tasks. The ICD vocabulary was derived from the training split, required at least 20 positive training admissions per retained code and was capped at 2,048 labels. For HCUP prediction, a 26-label target space was derived from ICD9/10-to-HCUP mappings[39,40], including the 25 phenotype groups used in established clinical time-series benchmarks[41] and an additional heart-failure phenotype (Supplementary Table 8). DRG used a training-derived 512-code vocabulary and competitive unordered-set objective. The prespecified primary metric was AUPRC for each binary outcome, micro-AUPRC for ICD, macro-AUPRC for HCUP and micro-F1 for DRG. Secondary measures included AUROC, Brier score and Expected Calibration Error (ECE) for binary outcomes; additional micro- and macro-averaged discrimination and retrieval metrics for ICD and HCUP; and macro-F1, top-1 hit and top-2 hit for DRG. Probe checkpoints were selected using the corresponding primary validation metric.

Primary comparisons used logistic regression and XGBoost as in-house benchmarks, trained using the same train/validation/test splits, state-variable vocabulary and task-specific observation windows as ATLAS-EHR. Six longitudinal summary strategies were evaluated independently: last observed value, mean, maximum, presence, count and inter-measurement interval. The value-based summaries were calculated from the standardized measurement values. Summary entries were initialized to zero for variables not observed within the applicable window; presence and count therefore explicitly encoded observation status, and the interval was also zero when fewer than two measurements were available. Logistic-regression inputs were subsequently standardized using feature means and standard deviations estimated from the training split, whereas XGBoost used the constructed summary features without this additional standardization. For each model and task, the summary representation was selected using the primary validation metric.

4.5.4 External validation and robustness

External validation was performed in eICU at the hospital-admission level. Multiple ICU stays belonging to the same hospitalization were merged chronologically, and event times were expressed relative to hospital admission. To preserve compatibility with the pretrained ATLAS-EHR variable encoder, eICU observations were restricted to variables that could be successfully mapped to the MIMIC-IV variable vocabulary. Only observations with a valid MIMIC variable identifier and successful value conversion were retained, and processed eICU records were required to contain no variable identifiers outside the MIMIC schema.

Endpoints were hospital mortality and LOS, from which LOS ≥3 days and LOS ≥7 days were derived. The same task-specific observation windows used in MIMIC-IV were applied in eICU: 48h for mortality and 24h for both LOS endpoints. Two transfer settings were evaluated: in zero-shot transfer, the MIMIC-trained ATLAS-EHR backbone and downstream probe were applied to eICU without parameter updating; in local adaptation, the MIMIC-trained backbone remained frozen and only a new downstream probe was fitted on the eICU training split and selected on the eICU validation split. The probe architecture and

representation-specific temporal processing were identical to those used for internal evaluation, as described in Section 5.5.3.

Logistic regression and XGBoost used the same MIMIC-derived feature contract, including variable identities, feature ordering and longitudinal summary definitions. In zero-shot transfer, the MIMIC-trained models and the summary representation selected on the MIMIC validation split were applied directly to eICU. In the locally adapted setting, logistic regression and XGBoost were fully retrained on eICU, including model fitting and validation-based selection of the longitudinal summary representation. Local adaptation therefore denotes probe-only training for ATLAS-EHR but complete task-specific model retraining for the tabular benchmarks.

Robustness to reduced observation density was assessed separately on the MIMIC-IV test split. For each repetition, raw state events were assigned deterministic pseudo-random deletion scores, with thresholds of 10%, 25% and 50% defining nested deletion masks. If deletion would otherwise remove all events from a non-empty evaluation window, one event was retained to preserve an evaluable sequence. Event deletion was applied before latent extraction; the perturbed sequences were then re-encoded using the frozen backbone and evaluated with the previously trained frozen probes. Performance at each deletion level was expressed relative to the corresponding unperturbed task-specific primary metric. For each task and deletion rate, performance was summarized across the three deterministic repetitions using the mean primary metric and its between-repetition standard deviation; relative retention was calculated from the repeat-mean performance.

4.5.5 Latent space characterization

Latent space characterization was performed on timestamp-level representations from the held-out test set with an available in-hospital mortality label. The primary analysis used z_{context}, whereas z_{set} was examined as a complementary representation. Clustering was performed in the original high-dimensional latent space. To reduce computational cost, HDBSCAN was applied recursively to a deterministic subset of up to 8,000 timestamp representations, and assignments were propagated to the remaining representations using five-nearest-neighbor classification in the same latent space. For visualization, up to 10,000 timestamp representations were sampled uniformly without replacement and projected using t-SNE. The two-dimensional projection was used only for visualization.

Cluster fingerprints were constructed from observed variables associated with each cluster. Numeric variables were eligible when observed in at least 20 admissions and in at least 1% of admissions represented in that cluster. For manuscript visualization, source-specific variable encodings verified to represent the same clinical measurement were mapped to a common clinical concept. Harmonization was applied after clustering and did not alter latent representations or cluster assignments. For concepts represented by multiple source-specific VIDs, the displayed fingerprint value was calculated as the mean of their existing within-cluster mean standardized values; single-VID concepts were unchanged. Harmonized concepts were then ranked by the absolute magnitude of the resulting standardized value, with the signed mean retained for presentation.

Mortality was analyzed after cluster assignment and was not used for clustering or fingerprint construction. For each cluster, admissions with at least one assigned timestamp were deduplicated before mortality prevalence was calculated. Temporal analyses repeated this procedure for cluster occurrences

within ≤24h, 24–48h and >48h after admission. An admission could contribute to multiple cluster or time-window summaries if its trajectory occupied those states at different times.

4.5.6 Disease-specific prediction attribution

To examine which components of the longitudinal record contributed to disease-specific predictions, attribution analyses were performed for stage IV chronic kidney disease (CKD) and severe sepsis. These diagnoses were selected as complementary case studies with distinct clinical signal structures: CKD is closely linked to a limited set of quantitative renal-function measurements, whereas severe sepsis is a multisystem syndrome represented by distributed hemodynamic, metabolic and inflammatory abnormalities. Analyses were restricted to admissions positive for the corresponding ICD label.

Perturbations were applied at the variable level across the complete admission before re-encoding. In value neutralization, each occurrence of the selected numeric variable was replaced with its variable-specific median estimated from the training split, while preserving variable identity, measurement occurrence and timestamp. This perturbation retained the observation pattern while removing patient-specific quantitative information. In variable removal, all occurrences of the selected numeric variable were removed from the event sequence before re-encoding. Timestamps containing no remaining observations after removal were subsequently omitted. Variable removal therefore altered both the selected quantitative information and its contribution to the surrounding measurement context.

Each perturbed admission was re-encoded with the frozen ATLAS-EHR backbone and evaluated using the frozen ICD probe. Attribution was quantified as the baseline target-disease logit minus the perturbed logit, such that positive values indicated a reduction in model evidence following perturbation. Variable-level summaries included admission-level prevalence, the number of evaluable admissions, median and interquartile range of the logit change, and the proportion of admissions with a positive logit decrease. For manuscript reporting, source-specific VIDs verified to represent the same clinical measurement were mapped to a common clinical concept. Perturbations themselves remained VID-specific. Within each admission, logit changes from member VIDs belonging to the same clinical concept were summed, and the concept-level effect was summarized as the median of these admission-level aggregated sensitivities among evaluable admissions in which the concept was present. These values therefore represent aggregates of independently evaluated VID-level marginal perturbations rather than the effect of simultaneously perturbing all member VIDs.

For the primary figure, harmonized clinical concepts were ranked separately within each disease and perturbation type according to the absolute magnitude of the median admission-level aggregated logit sensitivity, and the leading concepts were displayed. A within-admission same-variable value-shuffling analysis was retained as a supplementary sensitivity analysis and was not included in the primary figure.

Data Availability

The data supporting the findings from this study are available within the manuscript and its supplementary information. Public data from MIMIC is available from PhysioNet at https://doi.org/10.13026/kpb9-mt58.

Code Availability

Open-source code is available on GitHub (https://github.com/sxtforreal/ATLAS-EHR). The pre-trained model weights are available on Hugging Face (https://huggingface.co/sunx999/ATLAS-EHR)

Ethics Declarations

This study was performed using publicly available data and thus did not need Cleveland Clinic Institutional Review Board review.

Acknowledgements

This work was funded in part by National Heart Lung and Blood Institute award R01HL180504 (DC, CN). We would further like to acknowledge the Charles and Loraine Moore Endowed Chair in Cardiovascular Imaging for funding the computational environment which enabled this work.

Author Contributions

X.S., contributed to conceptualization, development of the model, analyzed, and validated the results. W.T., C.N., P.H.C., and X.W. contributed to validation, resources, and review and editing. D.K., and D.C. contributed to supervision, validation, funding, review and editing. X.S. and D.C. wrote and edited the manuscript.

Competing Interests

X.W. is on the Statistical Advisory Board for Nature Medicine. The authors declare no other competing interests.

5. References

1. Murdoch, T. B. & Detsky, A. S. The Inevitable Application of Big Data to Health Care. *JAMA* **309**, 1351–1352 (2013).
2. Chen, Z. *et al.* Harnessing the power of clinical decision support systems: challenges and opportunities. *Open Heart* **10**, e002432 (2023).
3. McDermott, M., Nestor, B., Argaw, P. & Kohane, I. S. Event Stream GPT: a data pre-processing and modeling library for generative, pre-trained transformers over continuous-time sequences of complex events. *Adv. Neural Inf. Process. Syst.* **36**, 24322–24334 (2023).
4. Hripcsak, G. & Albers, D. J. Correlating electronic health record concepts with healthcare process events. *J. Am. Med. Inform. Assoc.* **20**, e311–e318 (2013).
5. Cugola, G. & Margara, A. Processing flows of information: From data stream to complex event processing. *ACM Comput. Surv. CSUR* **44**, 15:1-15:62 (2012).
6. Karami, H., Atienza, D. & Ionescu, A. TEE4EHR: Transformer event encoder for better representation learning in electronic health records. *Artif. Intell. Med.* **154**, 102903 (2024).
7. Ambler, G., Omar, R. Z. & Royston, P. A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome. *Stat. Methods Med. Res.* **16**, 277–298 (2007).
8. Pungitore, S. & Subbian, V. Assessment of Prediction Tasks and Time Window Selection in Temporal Modeling of Electronic Health Record Data: a Systematic Review. *J. Healthc. Inform. Res.* **7**, 313–331 (2023).
9. Chen, D. *et al.* Deep Learning and Alternative Learning Strategies for Retrospective Real-World Clinical Data. *Npj Digit. Med.* **2**, 43 (2019).
10. Wu, S. *et al.* Modeling asynchronous event sequences with RNNs. *J Biomed Inf.* **83**, 167–177 (2018).
11. Che, Z., Purushotham, S., Cho, K., Sontag, D. & Liu, Y. Recurrent Neural Networks for Multivariate Time Series with Missing Values. *Sci. Rep.* **8**, 6085 (2018).
12. Chen, D. *et al.* Early Detection of Post-Surgical Complications using Time-series Electronic Health Records. *AMIA Jt. Summits Transl. Sci. Proc. AMIA Jt. Summits Transl. Sci.* **2021**, 152–160 (2021).
13. Labach, A. *et al.* Duett: dual event time transformer for electronic health records. in *Machine learning for healthcare conference* 403–422 (PMLR, 2023).
14. Fallahpour, A. *et al.* EHRMamba: Towards Generalizable and Scalable Foundation Models for Electronic Health Records. Preprint at https://doi.org/10.48550/arXiv.2405.14567 (2024).

15. Restrepo, D. *et al.* Representation Learning of Lab Values via Masked AutoEncoders. Preprint at https://doi.org/10.48550/arXiv.2501.02648 (2025).
16. Xiang, X., Restrepo, D., Jeong, H., Jia, Y. & Celi, L. A. Learning Representations from Incomplete EHR Data with Dual-Masked Autoencoding. Preprint at https://doi.org/10.48550/arXiv.2602.15159 (2026).
17. Steinberg, E., Fries, J., Xu, Y. & Shah, N. MOTOR: A Time-To-Event Foundation Model For Structured Medical Records. Preprint at https://doi.org/10.48550/arXiv.2301.03150 (2023).
18. Renc, P. *et al.* Zero shot health trajectory prediction using transformer. *Npj Digit. Med.* **7**, 1–10 (2024).
19. Renc, P. *et al.* Foundation model of electronic medical records for adaptive risk estimation. *GigaScience* **14**, giaf107 (2025).
20. Assran, M. *et al.* V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning. Preprint at https://doi.org/10.48550/arXiv.2506.09985 (2025).
21. Arena, P. J. *et al.* Information Bias in Electronic Health Records and Administrative Claims Data: A Targeted Review of the Recent Literature. *Clin. Pharmacol. Ther.* **119**, 362–367 (2026).
22. Doutreligne, M. *et al.* Step-by-step causal analysis of EHRs to ground decision-making. *PLOS Digit. Health* **4**, e0000721 (2025).
23. Edelman, B., Kim, H. & Skolnick, J. Scalable discovery and validation of order-specific electronic health record event trajectories for interpretable adverse-outcome risk estimation. *J. Biomed. Inform.* **177**, 105008 (2026).
24. Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O. & Hemingway, H. Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE: A CALIBER Study. *Am. J. Epidemiol.* **179**, 764–774 (2014).
25. Bashir, F. & Wei, H.-L. Handling missing data in multivariate time series using a vector autoregressive model-imputation (VAR-IM) algorithm. *Mach. Learn. Data Min. Tech. Med. Complex Data Anal.* **276**, 23–30 (2018).
26. A new analytical framework for missing data imputation and classification with uncertainty: Missing data imputation and heart failure readmission prediction | PLOS One. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0237724.
27. Li, J. *et al.* Imputation of missing values for electronic health record laboratory data. *Npj Digit. Med.* **4**, 147 (2021).
28. Haitjema, S. *et al.* The use of imputation in clinical decision support systems: a cardiovascular risk management pilot vignette study among clinicians. *Eur. Heart J. - Digit. Health* **5**, 572–581 (2024).

29. Tang, A. S. *et al.* Harnessing EHR data for health research. *Nat. Med.* **30**, 1847–1855 (2024).
30. Ostropolets, A. *et al.* Adapting electronic health records-derived phenotypes to claims data: Lessons learned in using limited clinical data for phenotyping. *J. Biomed. Inform.* **102**, 103363 (2020).
31. Liu, P. *et al.* Predicting the risks of kidney failure and death in adults with moderate to severe chronic kidney disease: multinational, longitudinal, population based, cohort study. *BMJ* **385**, e078063 (2024).
32. Desautels, T. *et al.* Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach. *JMIR Med Inf.* **4**, e28 (2016).
33. Albahri, A. S. *et al.* A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Inf. Fusion* **96**, 156–191 (2023).
34. Zhang, J., Zheng, S., Cao, W., Bian, J. & Li, J. Warpformer: A Multi-scale Modeling Approach for Irregular Clinical Time Series. in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* 3273–3285 (Association for Computing Machinery, New York, NY, USA, 2023). doi:10.1145/3580305.3599543.
35. Paik, J. M. *et al.* Accuracy of identifying diagnosis of moderate to severe chronic kidney disease in administrative claims data. *Pharmacoepidemiol. Drug Saf.* **31**, 467–475 (2022).
36. Birman-Deych, E. *et al.* Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Med. Care* **43**, 480–485 (2005).
37. Pollard, T. J. *et al.* The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Sci. Data* **5**, 180178 (2018).
38. Alsentzer, E. *et al.* Publicly available clinical BERT embeddings. *ArXiv Prepr. ArXiv190403323* (2019).
39. Agency for Healthcare Research Quality. HCUP Databases. Healthcare Cost and Utilization Project (HCUP). (2017).
40. Clinical Classifications Software Refined (CCSR)for ICD-10-CM Diagnoses. Agency for Healthcare Research and Quality.
41. Harutyunyan, H., Khachatrian, H., Kale, D. C., Ver Steeg, G. & Galstyan, A. Multitask learning and benchmarking with clinical time series data. *Sci. Data* **6**, 96 (2019).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTable1.xlsx
- SupplementaryTable5.xlsx
- SupplementaryTable7.xlsx
- SupplementaryTable3.xlsx
- SupplementaryTable8.xlsx
- SupplementaryTable4.xlsx
- SupplementaryTable6.xlsx
- SupplementaryTable2.xlsx