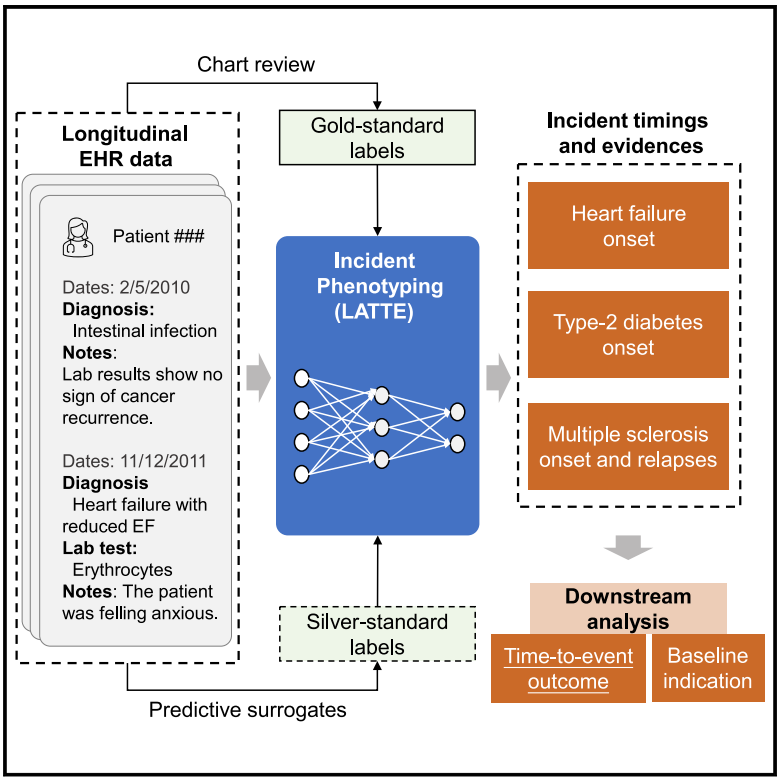
Article

Patterns

LATTE: Label-efficient incident phenotyping from longitudinal electronic health records

Graphical abstract



Highlights

- An incident phenotyping method is proposed to identify timings of clinical events
- We achieve label efficiency by exploiting EHR embeddings and predictive surrogates
- Model is validated on incident type 2 diabetes, heart failure, and multiple sclerosis
- Results facilitate assessing cardiac risks among patients with rheumatoid arthritis

Authors

Jun Wen, Jue Hou, Clara-Lea Bonzel, ..., Junwei Lu, Kelly Cho, Tianxi Cai

Correspondence

tcai@hsph.harvard.edu

In brief

A label-efficient method, LATTE, is proposed to identify the timings of clinical events from longitudinal electronic health records. It achieves significantly improved performance in identifying the onset of type 2 diabetes, heart failure, and relapses of multiple sclerosis. LATTE has strong cross-site portability and is highly interpretable by indicating the important features and visits that drive the predictions.

Wen et al., 2024, Patterns 5, 100906
January 12, 2024
https://doi.org/10.1016/j.patter.2023.100906

CellPress

Patterns

CellPress
OPEN ACCESS

Article

LATTE: Label-efficient incident phenotyping from longitudinal electronic health records

Jun Wen,[1,2] Jue Hou,[3] Clara-Lea Bonzel,[1,2] Yihan Zhao,[4] Victor M. Castro,[5] Vivian S. Gainer,[5] Dana Weisenfeld,[6] Tianrun Cai,[2,5] Yuk-Lam Ho,[2] Vidul A. Panickan,[1,2] Lauren Costa,[2] Chuan Hong,[7] J. Michael Gaziano,[1,2,6] Katherine P. Liao,[1,2,6] Junwei Lu,[2,8] Kelly Cho,[1,2,6] and Tianxi Cai[1,2,8,9,*]

[1]Harvard Medical School, Boston, MA, USA
[2]VA Boston Healthcare System, Boston, MA, USA
[3]University of Minnesota, Minneapolis, MN, USA
[4]Harvard University, Cambridge, MA, USA
[5]Mass General Brigham, Boston, MA, USA
[6]Brigham and Women's Hospital, Boston, MA, USA
[7]Duke University, Durham, NC, USA
[8]Harvard T.H. Chan School of Public Health, Boston, MA, USA
[9]Lead contact
*Correspondence: tcai@hsph.harvard.edu
https://doi.org/10.1016/j.patter.2023.100906

THE BIGGER PICTURE Electronic health record (EHR) data collected during routine clinical care are increasingly used by translational and clinical researchers to address a variety of questions, such as identifying associations between diseases or phenotypes, predicting disease risk or prognosis, or supporting the safety and efficacy of treatments. The feasibility of these studies relies on precisely inferring the timing and ordering of clinical events from EHR data to define baseline eligibility and patient outcomes. Rule-based extraction methods can be inaccurate, and existing machine-learning approaches generally require large-scale labels for training. Better methods for identifying the timing of clinical events could help expand the use of EHR data to address important medical questions and could improve the quality of the resulting analyses.

Proof-of-Concept: Data science output has been formulated, implemented, and tested for one domain/problem

SUMMARY

Electronic health record (EHR) data are increasingly used to support real-world evidence studies but are limited by the lack of precise timings of clinical events. Here, we propose a label-efficient incident phenotyping (LATTE) algorithm to accurately annotate the timing of clinical events from longitudinal EHR data. By leveraging the pre-trained semantic embeddings, LATTE selects predictive features and compresses their information into longitudinal visit embeddings through visit attention learning. LATTE models the sequential dependency between the target event and visit embeddings to derive the timings. To improve label efficiency, LATTE constructs longitudinal silver-standard labels from unlabeled patients to perform semi-supervised training. LATTE is evaluated on the onset of type 2 diabetes, heart failure, and relapses of multiple sclerosis. LATTE consistently achieves substantial improvements over benchmark methods while providing high prediction interpretability. The event timings are shown to help discover risk factors of heart failure among patients with rheumatoid arthritis.

INTRODUCTION

In recent years, electronic health record (EHR) data collected during the routine delivery of care has opened opportunities for discovery and translational research.[1,2] For example, EHR-derived cohorts have led to large-scale clinical studies and phenome-wide association studies.[3,4,5,6] Due to their large size and broad patient population, EHR cohorts are increasingly

used to support real-world evidence (RWE) on the efficacy and safety of therapeutic drugs or intervention procedures.[7,8,9,10] However, the capacity of EHR data for supporting RWE studies is currently limited due to the lack of direct observations on the precise timing of clinical events, such as the onset of heart failure. The timing information plays an important role in RWE studies, including in determining eligibility at baseline or defining time-to-event outcomes.[11] Readily available EHR features, such as the timing of relevant international classification of disease (ICD) codes, are often inaccurate due to either miscoding or ICD codes being assigned to visits that rule out a disease. Additionally, event-time-derived surrogates tend to have systematic biases.[12,13] On the other hand, it is time and resource prohibitive to extract event information via manual chart review. For binary phenotype traits, such as the presence or absence of a condition, a wide range of supervised, unsupervised, and label-efficient semi-supervised machine learning-based phenotyping algorithms have been successfully developed and validated across many disease phenotypes.[14,15,16,17,18] On the contrary, few methods currently exist to accurately and efficiently derive computational event time phenotypes based on longitudinal EHR data.

Existing approaches to deriving computational event time phenotypes can generally be categorized as rule based[13,19] and machine learning based.[20,21] For example, Chubak et al.[19] developed rules to predict breast cancer recurrence based on the earliest observation of expert-specified codes. Uno et al.[13] proposed to alleviate the systematic temporal biases between code timings and phenotype onset by using points of maximal increase in lieu of peak values. Even though rule-based methods can achieve notable performance for some phenotypes, they are limited by the reliance on expert knowledge to curate a small set of predictive surrogate concepts, which prevents their application to diseases without such predictive concepts or the scaling up to data with hundreds of unspecific features. Rule-based methods are tailored to specific applications, such as cancer recurrence per domain knowledge, and are hardly generalizable to other applications.

A more generalizable alternative approach is to employ machine learning to derive computational incident phenotyping algorithms using temporal patterns of EHR data. For example, random forests were investigated for phenotyping opioid overdose events.[22] Due to the stronger learning power, recently, deep learning models have been introduced for phenotyping.[23,24,25] Based on the label availability, these methods can be categorized into unsupervised clustering,[2,26,27] which mostly aims for novel phenotype or subtype discovery; supervised models,[21,28,29] which focus on devising novel network architectures to better model the structures of EHR data; and semi-supervised methods,[30,31,32,32,33] which aim for label efficiency by either leveraging the predictive silver labels or through unsupervised pre-training. For example, Zang et al.[33] propose to generate silver-standard labels to learn the screening of borderline personality disorders based on the model pre-trained using gold-standard labels. As for time-to-event prediction, sequential models are prevalently introduced to model visit temporal dependency. For example, a graph-based framework is proposed for temporal phenotyping by Liu et al.,[34] and recurrent neural networks are employed to mimic physician attention for interpretable phenotyping[21] and to perform outcome-oriented temporal phenotyping.[35] Due to the large-scale parameters, deep learning-based algorithms heavily depend on large-scale labels, which are expensive to obtain and not widely available, especially when the annotations involve event timings. To alleviate this issue, a recent semi-supervised algorithm, semi-supervised adaptive Markov Gaussian embedding process (SAMGEP),[20] is developed by imposing relatively simple linear effects on the concept embedding aggregation, which could lead the visit embedding vectors to be trivial and dominated by non-indicative common concepts, and by modeling disease progression as a Gaussian process emission through a hidden Markov model (HMM), which has limited capability of incorporating future information and capturing complex long-range visit dependency patterns. Meanwhile, self-supervised representation learning methods,[36,37] which aim to capture the concept or visit dependency by predicting the masked parts of longitudinal EHR data, and contrastive representation learning[38,39] are investigated to improve model generalization in scenarios with zero or few labels.

In this study, we propose a semi-supervised label-efficient incident phenotyping (LATTE) algorithm to derive the timing of clinical incidents from longitudinal EHR data. LATTE attains high accuracy with a small label size by effectively leveraging longitudinal silver-standard labels and the prior knowledge from semantic embeddings of EHR concepts to perform unsupervised pre-training and semi-supervised model co-training. Another key advantage of LATTE compared with existing literature lies in its cross-site portability, which is enabled via contrastive representation learning. Efficient and accurate annotations of clinical event times through LATTE strengthen the potential of EHR data for generating RWE.

Method

Overview

Architecture of LATTE. As illustrated in Figure 1, with longitudinal EHR data and a small number of labels on the phenotype status over time as the input, LATTE consists of four key computational components: (1) a concept re-weighting (CR) module that assigns a weight for each input concept or feature; (2) a visit attention network (VAN), which aims to assign higher weights to visits that are more indicative of the incident; (3) a sequential model to capture visit temporal dependency and obtain visit representations; and (4) final incident predictions at each visit by the predictor.

Procedures of LATTE. To alleviate the need for gold-standard labels, the learning of LATTE comprises the following two steps. (1) LATTE constructs longitudinal silver-standard labels for the event status over time based on predictive surrogates to perform unsupervised model pre-training. (2) LATTE is fine tuned jointly by the gold-standard labels and silver-standard labels. The four computational components of LATTE are optimized end to end together at both the pre-training and fine-tuning stages.

Input data structure and notations

EHR data consist of longitudinal patient visits, with each visit recording the observations of both structured EHR concepts (codes for diagnosis, procedures, medication prescriptions, laboratory test orders, and results) and unstructured narrative clinical notes. Raw EHR concepts are rolled up to higher-level

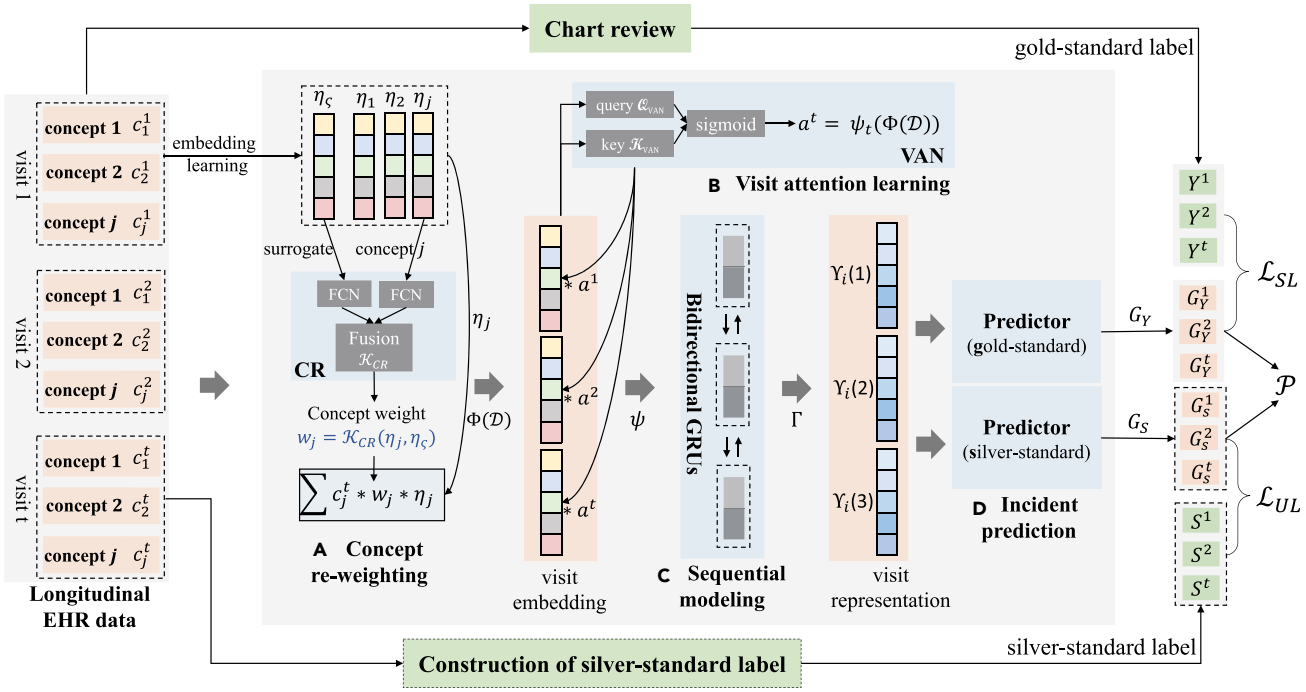


Figure 1. LATTE framework

LATTE is an end-to-end neural network pipeline consisting of four major components: (a) the CR module, which selects important input features based on their semantic relationship to the target phenotype; (b) the VAN, which learns to pay attention to the most incident-indicative visits; (c) BiGRU layers, which model the sequential dependency among visits; and (d) incident predictors, which generate incident predictions at each visit.

concepts according to common ontology as in previous studies[40]: Diagnosis (ICD) codes are grouped into the established phenome-wide association study (PheWAS) catalogs (PheCodes), procedure codes are grouped into categories according to the Clinical Classifications Software for Services and Procedures (CCS), medications are mapped to ingredient-level RxNorm codes, and laboratory tests are mapped to Logical Observation Identifiers Names and Codes (LOINC). From free-text clinical notes, we extract NLP (natural language processing) mentions of clinical terms mapped to the concept unique identifiers (CUIs) in the Unified Medical Language System (UMLS) using the Narrative Information Linear Extraction (NILE) tool.[40,41,42]

We assume that the training data contain a total of N patients, organized in a longitudinal format indexed by t and that without generality, the first M patients are labeled. For patient i and visit at time $t \in \{1,\ldots,T_i\}$, let $Y_i^t \in \{0,1\}$ denote the gold-standard label on the event status, which is annotated by physicians; let $S_i^t \in [0,1]$ be the silver-standard label for incidence status; and let $\mathbf{D}_i^t \in \mathbb{R}^p$ be the p-dimensional vector of features, where T_i is the last follow-up time for patient i, upper bounded by global maximal follow-up time $t_{\max}$. Here, the scalable silver-standard labels are obtained in an unsupervised manner. They are expected to be indicative of gold-standard labels but may be contaminated by bias or noise. For learning the onset time of a phenotype, silver-standard labels are typically derived from the corresponding diagnosis codes or NLP mentions. Hence, the input data consist of $\mathcal{G} = \{(Y_i^t, S_i^t, \mathbf{D}_i^t) : t = 1,\ldots,T_i, i = 1,\ldots,M\}$ for M labeled patients and $\mathcal{U} = \{(S_i^t, \mathbf{D}_i^t) : t = 1,\ldots,T_i, i = M+1,\ldots,N\}$ for the other $N - M$ unlabeled patients. Let $\mathcal{D}_i = \{\mathbf{D}_i^t : t = 1,\ldots,T_i\}$ be the features aggregated over follow-up time. Our LATTE algorithm builds a prediction model for $\mathbb{P}(Y_i^t = 1 | \mathcal{D}_i)$, $t = 1,\ldots,T_i$, the incidence rate over time given the longitudinal features.

Semantic embedding vector of input features. LATTE leverages a q-dimensional semantic embedding vector as the prior knowledge of each element of the p-dimensional feature $\mathbf{D}$ to compress visit information into a q-dimensional visit embedding $\varphi(\mathbf{D}_i^t)$. Let $\boldsymbol{\eta}_1,\ldots,\boldsymbol{\eta}_p \in \mathbb{R}^q$ be the semantic embedding vectors associated with features or concepts in $\mathbf{D}$ and $\boldsymbol{\eta}_\varsigma$ be the semantic embedding vector of the surrogate concept, which can be the target phenotype itself or a concept that is most predictive of it. The embedding vectors are obtained by performing matrix factorization, a variant skip-gram algorithm,[43,44] on a co-occurrence matrix of the EHR concepts aggregated from large-scale, unlabeled, patient-level, longitudinal EHR data.[40] As shown previously,[40,41,42,45] such embedding vectors effectively capture the clinical semantic relationship or similarity of EHR concepts or input features.

Construction of longitudinal silver labels. The silver-standard label S_i^t, which serves as a noisy proxy for Y_i^t, can be designed according to specific applications. When the event information can be well captured by surrogate features, such as a relevant NLP CUI or PheCode, LATTE constructs longitudinal silver-standard labels by leveraging such surrogate concepts, similar to other weakly supervised algorithms, such as PheNorm.[18] Let c_i^t denote the counts of silver-standard label concepts at visit t and U_i^t be a health utilization measure at t which is often needed for normalizing the silver-standard labels.

Table 1. Summary of the used notations

Notation	Description	Notation	Description
T_i	last follow-up time for patient i	$t_{\max}$	global maximal follow-up time
Y_i^t	incidence status for patient i at visit t	S_i^t	silver-standard label for patient i at visit t
Y_i	aggregation of Y_i^t for patient i across visits	S_i	aggregation of S_i^t for patient i across visits
D_i^t	features for patient i at visit t	$\mathcal{D}_i$	aggregation of D_i^t for patient i across visits
F	visit representation module	$\mathcal{V}_i$	visit representation of patient i
c_i^t	count of concept i at visit t	η	semantic concept embedding
$\mathcal{L}$	training loss	$\mathcal{P}$	penalty terms
Γ	Bi-GRU deep learning module	N	number of training patients
CR	concept re-weighting module	VAN	visit attention network
φ	CR transformation	ψ	visit attention transformation
$\mathcal{K}_{\mathrm{CR}}$	parameter of CR	ϑ	parameters of all layers
$\mathcal{Q}_{\mathrm{VAN}}$	query matrix of the VAN	$\mathcal{K}_{\mathrm{VAN}}$	key matrix of the VAN
G_Y	logistic transformation for incidence	G_S	logistic transformation for silver labels
$\tilde{\mathcal{G}}$	generated synthetic data	$\mathcal{A}_{i,k}$	time window of synthetic data

When the outcome of interest is time to the first onset of a condition, with Y_i^t indicating whether the event has occurred by time t, we use the cumulative counts $\sum_{u=0}^{t} c_i^u$ up to visit t,

$$S_{\mathrm{cum},i}^t = \mathrm{expit}\left[\left\{\log\left(1+\sum_{u=0}^{t} c_i^u\right) - \alpha \log\left(1+\sum_{u=0}^{t} U_i^u\right)\right\} \Big/ \tau\right]. \quad \text{(Equation 1)}$$

For prediction of recurrent events, such as relapse status over time, with Y_i^t episodically shifting between 0 and 1, we only use c_i^t at visit t,

$$S_{\mathrm{rec},i}^t = \mathrm{expit}\left[\left\{\log\left(1+c_i^t\right) - \alpha \log\left(1+U_i^t\right)\right\} / \tau\right]. \quad \text{(Equation 2)}$$

The hyperparameter τ denotes the temperature of the $\mathrm{expit}(\cdot)$ and controls the sharpness of the silver-standard label. A small τ would sharpen the silver labels to be in alignment with the gold-standard label, which is 0 for a negative visit and 1 for a positive visit. A controls the influence of U_i^t for constructing silver-standard labels.

Architecture and training of LATTE

We next describe the construction of deep-learning models $G_Y \circ \mathbf{F}(t, \mathcal{D}_i)$ for incidence $\mathbb{P}(Y_i^t = 1 | \mathcal{D}_i)$ and $G_S \circ \mathbf{F}(t, \mathcal{D}_i)$ for silver-standard label $\mathbb{E}(S_i^t | \mathcal{D}_i)$. Inspired by semi-supervised learning with a semi-parametric transformation model,[46] we let the models for gold-standard incidence Y_i^t and silver-standard label S_i^t share the core visit representation learning component $\mathbf{F}(t, \mathcal{D}_i) \in \mathbb{R}^q$ while allowing different prediction functions G_Y and G_S: $\mathbb{R}^q \mapsto [0,1]$. We create in $\mathbf{F}(t, \mathcal{D}_i)$ (1) a CR module to learn incident-indicative input concepts, (2) a VAN to highlight informative visits from background noises among other visits, and (3) the bidirectional gated recurrent unit (Bi-GRU) network for communication over time. Under Loss functions for semi-supervised learning, we describe the loss functions for the training of deep-learning model F. Based on the cross-entropy loss of the binary outcome Y_i^t, we devise (1) a kernel weighting strategy to borrow outcome data near t for learning $G_Y \circ \mathbf{F}(t, \mathcal{D}_i)$, (2) a penalty to regularize $G_Y \circ \mathbf{F}(t, \mathcal{D}_i)$ toward a function increasing/smooth along t and compatible with the context of the outcome, and (3) a semi-supervised training strategy to leverage the informative silver-standard labels S_i^t. We summarize the notations in Table 1.

Construction of deep-learning model. As illustrated by the model architecture in Figure 1, the incident prediction is produced through the following steps.

(1) Input layer: longitudinal $\mathcal{D}_i = \{\mathbf{D}_i^1, \ldots, \mathbf{D}_i^{T_i}\}$.

(2) module: mining the input concepts' semantic relationship to the target phenotype,

$$\varphi : \mathbf{D}_i^t \in \mathbb{R}^p \mapsto \varphi\left(\mathbf{D}_i^t\right) \in \mathbb{R}^q, q \ll p, \mathbf{\Phi}(\mathcal{D}_i) = \left(\varphi\left(\mathbf{D}_i^1\right), \ldots, \varphi\left(\mathbf{D}_i^T\right)\right).$$

3. Visit attention module: obtaining the visit attentions by contrasting $\varphi(\mathbf{D}_i^t)$ along t,

$$\boldsymbol{\psi} : \mathbf{\Phi}(\mathcal{D}_i) \in \mathbb{R}^{q \times T_i} \mapsto \boldsymbol{\psi} \circ \mathbf{\Phi}(\mathcal{D}_i) \in \mathbb{R}^{T_i}, \boldsymbol{\psi} = \left(\psi_1, \ldots, \psi_{T_i}\right).$$

4. Sequential modeling module: communicating the visit embeddings along time through a Bi-GRU layer:

$$\mathbf{\Gamma} : \{\mathbf{\Phi}(\mathcal{D}_i), \boldsymbol{\psi}(\mathcal{D}_i)\} \mapsto \{\mathbf{F}(1, \mathcal{D}_i), \ldots, \mathbf{F}(T_i, \mathcal{D}_i)\}, \mathbf{F}(t, \mathcal{D}_i) \in \mathbb{R}.$$

5. Incident predictor: separating the logistic regression transformations for incidence and silver-standard label models with link $\mathrm{expit}(x) = 1/(1+e^{-x})$,

$$\text{for incidence } G_Y\{\mathbf{F}(t, \mathcal{D}_i)\} = \mathrm{expit}\left\{\beta_{Y,0} + \boldsymbol{\beta}_Y^{\top} \mathbf{F}(t, \mathcal{D}_i)\right\},$$

$$-\text{ standard label } G_S\{\mathbf{F}(t, \mathcal{D}_i)\} = \mathrm{expit}\left\{\beta_{S,0} + \boldsymbol{\beta}_S^{\top} \mathbf{F}(t, \mathcal{D}_i)\right\}.$$

for silver

We detail the design of φ, with parameter $\mathcal{K}_{\mathrm{CR}}$, $\boldsymbol{\psi}$, with parameters $(\mathcal{Q}_{\mathrm{VAN}}, \mathcal{K}_{\mathrm{VAN}})$, and $\mathbf{\Gamma}$ in the following paragraphs. The final

models are determined by the combined parameters of φ, $\boldsymbol{\psi}$, $\boldsymbol{\Gamma}$, G_Y and G_S across all layers; namely,

$$\vartheta = \left(\mathcal{K}_{\mathrm{CR}}, \mathcal{Q}_{\mathrm{VAN}}, \mathcal{K}_{\mathrm{VAN}}, \boldsymbol{\Gamma}, \beta_{Y,0}, \boldsymbol{\beta}_Y, \beta_{S,0}, \boldsymbol{\beta}_S\right).$$

CR: Selecting important features. LATTE learns a CR module to attach a weight to each input concept by mining its semantic relationship to the target phenotype, which reduces colinearity of features and overfitting along irrelevant features. We characterize the relevance of features by a multilayer perception (MLP) network:

$$\mathcal{K}_{\mathrm{CR}} : (\boldsymbol{\eta}_j, \boldsymbol{\eta}_{\mathcal{C}}) \in \mathbb{R}^q \times \mathbb{R}^q \mapsto \mathcal{K}_{\mathrm{CR}}(\boldsymbol{\eta}_j, \boldsymbol{\eta}_{\mathcal{C}}) \in \mathbb{R}.$$

$\mathcal{K}_{\mathrm{CR}}$ has two input branches: one branch receives the embedding vector of the input concept, and the other one receives the embedding vector of the target phenotype. Both branches first use one fully connected layer to learn a low-dimensional representation, which is then fused to output the importance of the input concept, ranging from 0–1. The $\mathcal{K}_{\mathrm{CR}}$ module is shared across all concepts and optimized end to end to learn the weight of each input concept.

Using standardized weights derived from $\mathcal{K}_{\mathrm{CR}}$, we aggregate the concept embedding vectors within each visit to obtain the visit embeddings:

$$\varphi\left(\mathbf{D}_i^t; \mathcal{K}_{\mathrm{CR}}\right) = \frac{1}{p}\sum_{j=1}^{p} D_{ij}^t \frac{\exp\left\{\mathcal{K}_{\mathrm{CR}}(\boldsymbol{\eta}_j, \boldsymbol{\eta}_{\mathcal{C}})\right\}}{\sum_{j'=1}^{p}\exp\left\{\mathcal{K}_{\mathrm{CR}}(\boldsymbol{\eta}_{j'}, \boldsymbol{\eta}_{\mathcal{C}})\right\}}\boldsymbol{\eta}_j. \quad \text{(Equation 3)}$$

Such a competing normalization across all p concepts would induce visit embeddings dominated by the most informative features, thus filtering out the noise from irrelevant features.

Compared with the direct utilization of embedding similarity between $\boldsymbol{\eta}_j$ and $\boldsymbol{\eta}_{\mathcal{C}}$ as the concept importance, our data-driven module $\mathcal{K}_{\mathrm{CR}}$ would adaptively capture the predictability of features even when the semantic similarity aligns poorly with predictability of incidence.

Visit attention: Highlighting informative visits. In the next step, LATTE devises a VAN to highlight informative visits according to the attention value. The VAN module can reduce noise from non-informative visits adaptively for patients with heterogeneous background noises. In the VAN, we employ a self-attention[47] layer to contrast informative visits from the whole visit sequence in the background. Self-attention[47] has been shown to be a successful technique for capturing long-range sequential dependency for various data types, including text[48] and video.[49] The VAN receives sequential visit embedding vectors to obtain corresponding attention values. There are two components in the VAN, shared across all visits:

$$\text{Query } \mathcal{Q}_{\mathrm{VAN}} : \varphi\left(\mathbf{D}_i^t\right) \in \mathbb{R}^q \mapsto \mathcal{Q}_{\mathrm{VAN}} \circ \varphi\left(\mathbf{D}_i^t\right) \in \mathbb{R}^d,$$

$$\text{Key } \mathcal{K}_{\mathrm{VAN}} : \varphi\left(\mathbf{D}_i^t\right) \in \mathbb{R}^q \mapsto \mathcal{K}_{\mathrm{VAN}} \circ \varphi\left(\mathbf{D}_i^t\right) \in \mathbb{R}^d,$$

both of which consist of a linear mapping layer as in Vaswani et al.[47] to map the q-dimensional visit embedding to the d-dimensional query and key vectors, respectively. The attention value for visit t of patients i is derived from averaging the inner products between its query vector and the key vectors across all available T_i visits, including itself,

$$\psi_t(\boldsymbol{\Phi}(\mathcal{D}_i); \mathcal{Q}_{\mathrm{VAN}}, \mathcal{K}_{\mathrm{VAN}}) = \frac{1}{T_i}\sum_{u=1}^{T_i}\mathrm{expit}\Big[\mathcal{Q}_{\mathrm{VAN}}\left\{\varphi\left(\mathbf{D}_i^u\right)\right\}^{\top} \mathcal{K}_{\mathrm{VAN}}\left\{\varphi\left(\mathbf{D}_i^t\right)\right\} \Big/ \sqrt{d}\Big], \quad \text{(Equation 4)}$$

where $\mathrm{expit}(x) = 1/(1+e^{-x})$. We choose the expit function rather than the typical soft-max function[47] because there could be multiple visits or incidents to pay attention to.

Bi-GRU: Information communication along time. Last, LATTE employs a recurrent neural network to model the sequential dependency between visits to learn visit representation, upon which phenotype incidents are predicted at each visit. To enable both prior and future visit information to be utilized for the prediction at the current visit, one Bi-GRUs layer is employed for the sequential modeling. Bi-GRUs receives the patient's longitudinal visit embedding vectors $\varphi(\mathbf{D}_i^t)$ from CR, along with their corresponding attention values $\psi_t \circ \boldsymbol{\Phi}(\mathcal{D}_i)$ learned by VAN, and outputs the representation of each visit

$$\begin{aligned}\boldsymbol{\Gamma}\{\boldsymbol{\Phi}(\mathcal{D}_i), \boldsymbol{\psi}(\mathcal{D}_i)\} &= \mathrm{BiGRU}_{\boldsymbol{\Gamma}}\left\{\varphi\left(\mathbf{D}_i^1\right)\psi_1 \circ \boldsymbol{\Phi}(\mathcal{D}_i), \ldots, \varphi\left(\mathbf{D}_i^{T_i}\right)\psi_{T_i} \circ \boldsymbol{\Phi}(\mathcal{D}_i)\right\} \\ &= \{\mathbf{F}(1, \mathcal{D}_i), \ldots, \mathbf{F}(T_i, \mathcal{D}_i)\} \\ &= \{\mathcal{V}_i(1), \ldots, \mathcal{V}_i(T_i)\},\end{aligned}$$

where $\mathcal{V}_i(t)$ denotes the visit representation of patient i at visit t.

Loss functions for semi-supervised learning. For a binary outcome Y_i^t, the pooled cross-entropy loss across patients and visits would be the typical choice for model training, as done in SAMGEP[20] and RETAIN.[21] However, this loss function has two major limitations for incident phenotyping: (1) the outcomes Y_i^t only contribute to the prediction model at time t, ignoring the longitudinal nature of the outcome-feature pairs, and (2) the fact that there is no guarantee regarding the monotonicity/smoothness of the prediction and the model may compromise its rationality. For example, the cumulative incidence rate may be expected to be non-decreasing over time. Moreover, a large amount of training data is typically required to effectively utilize the capacity of the complex deep learning model. This poses a major concern because it is very costly to generate a large quantity of gold-standard outcomes. To address these challenges, we propose a two-step semi-supervised training strategy combining information from outcomes Y_i^t and silver-standard labels S_i^t through supervised and unsupervised kernel-weighted losses $\mathcal{L}_{\mathrm{SL}}$ and $\mathcal{L}_{\mathrm{UL}}$ with penalty terms $\mathcal{P}_{\mathrm{SL}}$ and $\mathcal{P}_{\mathrm{UL}}$, which encourages the model's monotonicity/smoothness.

Next, we introduce the loss function components $\mathcal{L}_{\mathrm{SL}}$, $\mathcal{L}_{\mathrm{UL}}$, and $\mathcal{P}$, used for the two-step label efficient training strategy described under Label-efficient training of LATTE.

Kernel-weighted losses: Incorporating distance into incidence. The outcomes around a given time t may provide useful modality information for the prediction model at that time t. For cumulative

incidence, the features $\mathbf{D}_i^t$ may follow different patterns depending on the distance between t and the onset time. We translate the distance factor into a kernel weighting, where

$$w_{Y,i}^t = w_{\min} + \exp\left\{-d_{Y,i}(t)^2 \Big/ \left(2h^2\right)\right\}, d_{Y,i}(t) = \min\left\{|u-t| : Y_i^u = 1\right\},$$

$$w_{S,i}^t = w_{\min} + \exp\left\{-d_{S,i}(t)^2 \Big/ \left(2h^2\right)\right\}, d_{S,i}(t) = \min\left\{|u-t| : S_i^u \geq \kappa\right\}.$$

Here, h is a bandwidth hyperparameter, κ is a threshold for silver-standard label S_i^u above which incidence Y_i^u is likely active, and we set $\min\{\varnothing\} = +\infty$. A minimum weight of $w_{\min}$ ensures the stability of the loss. With standardized kernel weighting, we construct the supervised and unsupervised loss functions,

$$\mathcal{L}_{\mathrm{SL}}(\vartheta) = -\frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{t=1}^{T_i} w_{Y,i}^t\left[Y_i^t \log\{G_Y \circ \mathbf{F}(t,\mathcal{D}_i)\} + \left(1-Y_i^t\right)\log\{1-G_Y \circ \mathbf{F}(t,\mathcal{D}_i)\}\right]}{\sum_{t'=1}^{T_i} w_{Y,i}^{t'}}$$

$$\mathcal{L}_{\mathrm{UL}}(\vartheta) = -\frac{1}{N}\sum_{i=1}^{N}\frac{\sum_{t=1}^{T_i} w_{S,i}^t\left[S_i^t \log\{G_S \circ \mathbf{F}(t,\mathcal{D}_i)\} + \left(1-S_i^t\right)\log\{1-G_S \circ \mathbf{F}(t,\mathcal{D}_i)\}\right]}{\sum_{t'=1}^{T_i} w_{S,i}^{t'}}. \quad \text{(Equation 5)}$$

Penalty: Regularization toward monotonicity/smoothness. Depending on the type of incidence studied, we construct two penalty terms. We use the full cohort for the construction of penalties because (1) no outcome information is needed, and (2) the monotonicity/smoothness is expected for prediction over the full cohort. For the prediction of the cumulative incidence rate, with Y_i^t a counting process with at most one jump, we impose a penalty that encourages the longitudinal prediction to be non-decreasing across time,

$$\mathcal{P}_{\mathrm{cum}}(\vartheta) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i-1}\sum_{t=1}^{T_i-1}\max\{G_S \circ \mathbf{F}(t,\mathcal{D}_i) - G_S \circ \mathbf{F}(t+1,\mathcal{D}_i), 0\}. \quad \text{(Equation 6)}$$

For the prediction of recurrent incidence rate, with Y_i^t episodically shifting between 0 and 1, we impose a penalty that regulates the longitudinal prediction to ensure its smoothness over time.

$$\mathcal{P}_{\mathrm{rec}}(\vartheta) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i-1}\sum_{t=1}^{T_i-1}\left\|\mathbf{F}(t,\mathcal{D}_i) - \mathbf{F}(t+1,\mathcal{D}_i)\right\|_2. \quad \text{(Equation 7)}$$

We denote $\mathcal{P}$ the penalty chosen between $\mathcal{P}_{\mathrm{cum}}$ and $\mathcal{P}_{\mathrm{rec}}$, depending on the type of incidence studied.

Label-efficient training of LATTE. LATTE is built on deep neural networks, which, in general, heavily depend on large-scale annotations. To address the possible over-fitting under a small set of gold-standard labels, we rely on scalable silver-standard labels on which LATTE is optimized in two steps: unsupervised pre-training and semi-supervised joint training.

Unsupervised pre-training. In the first step, LATTE is pre-trained using the silver labels. As shown in Figure 1, instead of sharing the same predictor, $G_Y \circ \mathbf{F}$, that predicts gold-standard labels, we attach an additional silver predictor, $G_S \circ \mathbf{F}$, to the visit representation to predict the silver labels. We pre-train all deep learning model parameters ϑ, excluding G_Y, using the unsupervised loss $\mathcal{L}_{\mathrm{UL}}$ and the penalty $\mathcal{P}$. We then denote the resulting loss

$$\mathcal{L}_{\mathrm{PT}}(\vartheta) = \mathcal{L}_{\mathrm{UL}}(\vartheta) + \lambda\mathcal{P}(\vartheta), \quad \text{(Equation 8)}$$

$$\widehat{\vartheta}_{\mathrm{PT}} = \underset{\vartheta}{\operatorname{argmin}}\mathcal{L}_{\mathrm{PT}}(\vartheta), \quad \text{(Equation 9)}$$

where λ is a hyperparameter determining the level of penalization. The CR $\mathcal{K}_{\mathrm{CR}}(\boldsymbol{\eta}_j, \boldsymbol{\eta}_{\mathcal{G}})$ learned in the pre-training can then be used in another round of feature selection because it reflects the relevance of the features $\mathbf{D}_i^t$ to the silver-standard labels S_i^t.

Semi-supervised joint training. In the second step, LATTE performs semi-supervised fine-tuning of the model using both the gold-standard labels and the silver labels. The use of separate predictors for those two types of labels aims to prevent the potential poor quality of silver-standard labels from deteriorating the learning of gold-standard labels. The final training objective then becomes

$$\mathcal{L}_{\mathrm{SSL}}(\triangle\vartheta) = \mathcal{L}_{\mathrm{SL}}(\widehat{\vartheta}_{\mathrm{PT}} + \triangle\vartheta) + \gamma\mathcal{L}_{\mathrm{UL}}(\widehat{\vartheta}_{\mathrm{PT}} + \triangle\vartheta) + \lambda\mathcal{P}(\widehat{\vartheta}_{\mathrm{PT}} + \triangle\vartheta),$$
$$\widehat{\vartheta}_{\mathrm{LATTE}} = \widehat{\vartheta}_{\mathrm{PT}} + \underset{\triangle\vartheta}{\operatorname{argmin}}\mathcal{L}_{\mathrm{SSL}}(\triangle\vartheta), \quad \text{(Equation 10)}$$

where γ balances the contribution of gold-standard labels and silver-standard labels. CR model $\mathcal{K}_{\mathrm{CR}}$, VAN modules $(\mathcal{Q}_{\mathrm{VAN}}, \mathcal{K}_{\mathrm{VAN}})$ and the GRU model $\boldsymbol{\Gamma}$ are jointly optimized according to the combination of the three objective functions. The

output transformations G_Y and G_S are optimized according to their involvements in $\mathcal{L}_{SL}(\vartheta)$ and $\mathcal{L}_{UL}(\vartheta)$, respectively.

Enhancing cross-site portability

We further strengthen LATTE's cross-site portability via contrastive representation learning. We consider three aspects of data shift: (1) concept utilization bias, (2) visit frequency, and (3) visit phase. Utilization bias is the fact that different institutions tend to have different preferences regarding concept utilization, especially for medications. Visit frequency shifts reflect the idea that different patients visit hospitals at different frequencies. Visit phase shift is caused by the fact that patients tend to visit and be discharged from hospitals during different stages of the phenotype, causing longitudinal EHR data to reflect patients at various phenotype phases. For example, some patients may have already developed heart failure by the time of their first EHR visit, while other patients may not yet have signs of the condition.

To enhance the robustness of LATTE toward these data shifts, we construct a robustness measure based on synthetic data that will be incorporated into the loss functions. To mimic the three data shifts, we generate the synthetic data

$$\tilde{\mathcal{G}} = \left\{\left(\tilde{Y}^t_{i,k}, \tilde{\mathbf{D}}^t_{i,k}\right) : t = 1,\ldots,\tilde{T}_{i,k}, i = 1,\ldots,N, k = 1,\ldots,M\right\}$$

from labeled data $\mathcal{G}$ using the bellow strategy.

(1) Placing a random Gaussian noise and a random corruption on concept counts to randomly set some concept counts to zero. We sample $\varepsilon^t_{i,k,j} \sim \mathcal{N}(1, 0.05)$, $\delta^t_{i,k,j} \sim Bern(0.9)$ and add noise to the features,

$$\tilde{D}^t_{i,k,j} = \delta^t_{i,k,j}\left(\overline{D}^t_{i,k} + \varepsilon_{i,k,j}\right).$$

2. Aggregating the visit sequence with varied time windows to mimic varied visit frequency. With an integer hyperparameter a_{mid}, we cycle $\mathcal{A}_{i,k}$ through $a_{\text{mid}} - 1, a_{\text{mid}}, a_{\text{mid}} + 1$. We aggregate the original $T_{i,k}$ visits into $\tilde{T}_{i,k} = \lceil T_{i,k} / \mathcal{A}_{i,k} \rceil$ visits, where $\lceil x \rceil$ denotes the smallest integer bigger than x. A new visit at time t thus combines the previous visits from $\epsilon_t = (t-1)\mathcal{A}_{i,k} + 1$ to $\nu_t = min\{t\mathcal{A}_{i,k}, T_{i,k}\}$ by taking the maximal outcome and summing over features

$$\tilde{Y}^t_{i,k} = \max_{u:\epsilon_t \le u \le \nu_t} Y^u_{i,k}, \overline{\mathbf{D}}^t_{i,k} = \sum_{u=\epsilon_t}^{\nu_t} \mathbf{D}^u_{i,k}, t = 1, 2, \ldots, \tilde{T}_{i,k}.$$

3. Randomly shifting the starting visit before the incidents of the EHR sequences to obtain sub-sequence samples. We randomly sample a truncation time $L_{i,k}$ from $\{0, 1, \ldots, min(T_i, L_{\max})\}$ with equal probability. Here, $L_{\max}$ is an integer hyperparameter for maximal truncation time. We truncate the original data as

$$\check{T}_{i,k} = T_i - L_{i,k}, \left(\check{Y}^t_{i,k}, \check{\mathbf{D}}^t_{i,k}\right) = \left(Y^{t-L_{i,k}}_i, \mathbf{D}^{t-L_{i,k}}_i\right), t = 1, \ldots, \check{T}_{i,k}.$$

We construct a concordance of $\mathbf{F}$ between $\mathcal{G}$ and $\tilde{\mathcal{G}}$ to measure the robustness of the model toward data shifts,

$$\mathcal{L}_{CC}(\vartheta) = \frac{1}{N^2 M}\sum_{i'=1}^{N}\sum_{i=1}^{N}\sum_{k=1}^{M}\sum_{t'=1}^{T_{i'}} \times \sum_{t=1}^{\tilde{T}_{i,k}} \frac{\max\left\{(-1)^{\mathbb{I}\left(Y^{t'}_{i'} = \tilde{Y}^t_{i,k}\right)}(c - \|\mathbf{F}(t', \mathcal{D}_{i'}) - \mathbf{F}(t, \tilde{\mathcal{D}}_{i,k})\|_2), 0\right\}}{T_{i'}\tilde{T}_{i,k}},$$

where c is a tolerance for the contrast between positive and negative visits. $\mathcal{L}_{CC}$ encourages the representation of positive visits to cluster together and keep at least a distance of c from negative visits. Incorporating $\mathcal{L}_{CC}$ with a hyperweighting parameter κ, the cross-site portable semi-supervised loss function becomes

$$\mathcal{L}_{CSP}(\vartheta) = \mathcal{L}_{SL}(\vartheta) + \gamma\mathcal{L}_{UL}(\vartheta) + \lambda\mathcal{P}(\vartheta) + \kappa\mathcal{L}_{CC}(\vartheta). \quad \text{(Equation 11)}$$

RESULTS

We first evaluate LATTE on three representative phenotypes to demonstrate its advantages over existing methods regarding label efficiency and cross-site portability for incident phenotyping. Further, based on phenotype incident predictions on longitudinal EHR data, we identify risk factors of heart failure among patients with rheumatoid arthritis (RA).

Performance of incident phenotyping

Data and settings

Data sources. We first evaluate the performance of LATTE in identifying three temporal events, onset of type 2 diabetes (T2D; PheCode 250.2) and heart failure (HF; PheCode 428) and relapses over time for those with multiple sclerosis (MS; PheCode 335), using EHR data from Mass General Brigham (MGB). Both T2D and HF are chronic diseases for which the first onset probability over time is of primary interest. MS relapse is a relapsing and remitting phenotype for which we aim to predict all recurrent relapses over time. For both T2D and HF, the corresponding diagnostic codes are predictive of the ever/never status, but the dates of the first diagnostic codes often deviate from the true incident times with systematic preceding and lagging biases. For MS relapse, no predictive diagnostic code exists, so complex visit dependency modeling is required to precisely identify incident visits. For HF, we further evaluate the transportability of the algorithm to a Biobank cohort at MGB and to the Million Veteran Project (MVP) cohort at the Veterans Affairs (VA).

For T2D, we assemble EHR data for 10,315 patients who have at least one T2D ICD code from the MGB healthcare system. The T2D status and onset dates for 172 randomly selected patients from this cohort are annotated via chart review. Among the 172 patients, 52.3% develop T2D during the follow-up time, with EHR visits observed at MGB, and 11.0% have already developed T2D before their first clinical visit at MGB. 10-fold cross-validation is used for performance evaluation. For MS relapse, we assemble EHR data for 4,706 patients at MGB, of which 1,435 patients are participants of the Comprehensive Longitudinal Investigation of Multiple Sclerosis at BWH (CLIMB) research registry with relapse status annotated over time. Within the CLIMB cohort, 57.2% of patients have at least one relapse event, with a mean of 2.60 relapses per patient.

For HF, we train the algorithm using EHR data from the MGB RA cohort, in which each patient has at least 1 ICD code for RA. HF status and onset time are annotated for a random subset of 234 patients, among which 60.7% are determined to have developed HF during follow-up. Beyond 10-fold cross-validation, We further evaluate the portability of the HF incident phenotyping algorithm trained in the MGB RA cohort to the MGB Biobank cohort and the MVP cohort at the VA. The MGB Biobank and VA-MVP cohorts consist of 13,597 and 122,035 patients with at least 1 ICD code of HF, of which 94 and 208, respectively, are randomly annotated and used for transportability validation.

We bin the longitudinal EHR data into consecutive, non-overlapping, 3-month time windows,[20] from which event rates, commonly in years,[9] as clinical outcomes can be easily derived. However, we also clarify that the choice of the time window may depend on whether the event is highly acute, in which case a short window can be considered. For T2D and HF, we use the codified features only. As a chronic condition, management of T2D will generate consistent patterns of codified EHR data, such as routine clinic visits with T2D ICD codes, lab tests for monitoring glycemic control, and prescription of anti-diabetic agents. For HF onset, we additionally include HF-relevant NLP CUIs extracted from medical notes. This is because, as an acute event, key codified information identifying HF, concentrated in a short time window, is subject to data leakage if the patient is admitted into out-of-network hospitals. We leverage documentation of HF events in narrative notes from subsequent in-network encounters. In accordance with our reasoning, we observe that the information from medical notes substantially boosts the performance only for HF but not for T2D. For both phenotypes, the codified features are selected based on medical knowledge graphs extracted via knowledge extraction via sparse embedding regression (KESER), which connects each phenotype to the relevant medical concepts. The concept section by KESER is shown to work effectively for phenotyping.[40] The NLP CUIs are selected via the CUI search tool (http://app.parse-health.org/CUISearch/), which combines the medical knowledge graph of KESER and knowledge from multiple sources. For MS relapse, 155 EHR features are manually selected by a domain expert as in a previous study.[20]

Compared methods. We consider five benchmark methods: (1) long short-term memory recurrent neural network (LSTM); (2) RETAIN,[21] which is trained with longitudinal raw EHR features and without effective utilization of pre-trained concept embedding vectors, and (3) SAMGEP,[20] which aggregates patient visit embedding as a pre-processing step without learning to distinguish the importance of concepts/visits and mines only linear and feedforward dependencies between visits, lacking the capacity to model the complex and long-range temporal relationship between incidents and patient visits. For both methods, the training of incidence phenotyping models depends exclusively on the labeled data with gold-standard outcomes, overlooking the information from the predictive surrogates of the vast unlabeled data.

We also compare methods focusing on representation learning on EHR data, including (4) SCEHR,[39] which proposes contrastive cross-entropy loss and a contrastive regularizer; (5) OTCSurv,[50] which is equipped with similar concepts and visit attention modules and an additional weighted contrastive learning strategy; and (6) BEHRT,[36] which performs self-supervised representation learning on the longitudinal EHR data based on the transformer model.[47] In addition, we include a deep language model (large language model [LLM]), Bio-GPT,[51] in which we concatenate the text strings of observed concepts as the input of each visit. As a baseline, we include predictions based only on the closest PheCodes (T2D, 250.2; HF, 428; MS relapse, 355) and PheNorm with temporally cumulative counts,[18] denoted as PheNorm(acc), which is shown to perform effectively on ever/never phenotyping.

Evaluation measures. To evaluate the methods' performance, we sample varied sizes of patients from the gold-standard labeled set to train the model and evaluate each trained model on the rest of the labels. For T2D and HF, the visits before the first onset are treated as negatives, and the visits after it are treated as negatives. For MS relapse, the visits with relapses are treated as positives and otherwise as negatives. To quantify the accuracy of the methods' predictions, we compute (1) AUC (area under the receiver operating characteristic curve) and (2) F1 score with a cutoff value that achieves 95% specificity. The two values measure the incident identification error by treating each visit independently. We also report the methods' longitudinal phenotype predictions; namely, the area between the label curve Y_i^t and predicted cumulative probability $\hat{Y}_i^t = 1 - \prod_{k=0}^{k \le t}(1 - p_i^k)$, denoted by ABC_{cdf}, where p_i^k denotes patient i's prediction at time k. ABC_{cdf} effectively evaluates the mean absolute difference between true and predicted incident times but would scale up when patients are observed with longer EHR observations. (3) Therefore, we compute the normalized version, ABC_{gain}, which is the methods' percent decrease over a *null* model that sets the probability at each visit to the prevalence of the phenotypes;[20] namely $ABC_{gain} = (ABC_{cdf,null} - ABC_{cdf,method})/ABC_{cdf,null}$.

Result analysis

The phenotyping results are shown in Figure 2 in which the error bars indicate 95% confidence interval. For better visualization, we truncate ABC_{gain} values to be above 0. Different sizes of gold-standard label sets are evaluated, and the "0" labels, visualized as dashed lines, denote the unsupervised scenario where no gold-standard label is used.

T2D. The results of classifying T2D first onset time in the MGB cohort are shown in Figure 2A. LATTE shows significant advantages over the compared methods in terms of ABC_{gain}, F1 score, and AUC. The unsupervised approaches, PheCode only, PheNorm(ACC), and LATTE(silver), achieve comparable performance in terms of AUC and F1 score compared with the supervised models, LSTM, RETAIN, SAMGEP, and LATTE. This indicates that the main PheCode itself is able to distinguish the case visits from control visits. In addition to that, the PheCode and PheNorm(acc) are significantly outperformed in ABC_{gain} by the supervised models, which indicates that PheCode only fails to precisely localize the incidents. On the other hand, the performance of LATTE(silver), which is trained only using silver-standard labels, is comparable to LSTM, RETAIN, and SAMGEP, which are trained by 100 gold-standard labels. The results show the promise of unsupervised representation learning upon the predictive surrogate concepts. Of note, although LATTE(silver) exploits the main PheCode as a strong surrogate, which is the same as in PheNorm, it optimizes the sequential representation learned by GRUs to output

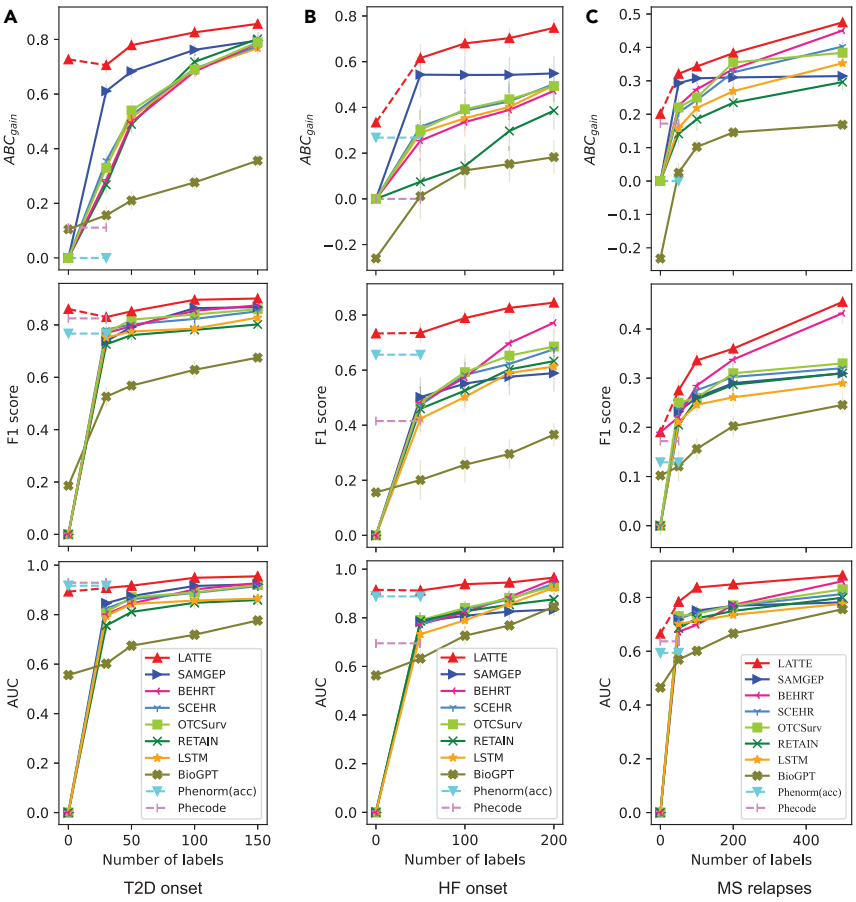


Figure 2. Numerical results of incident phenotyping with varied gold-standard label sizes

We evaluate LATTE on the onset of both type 2 diabetes (T2D) and heart failure (HF) and the onset and relapse of multiple sclerosis (MS). Error bars indicate 95% confidence intervals.

non-decreasing incident risks across time. As a result, the predictions obtained using LATTE(silver) are guided to be more sensitive to the onset of T2D. When only small sets of labels are available (for example, 30 or 50), the deep learning models LSTM and RETAIN suffer severe overfitting and are outperformed by SAMGEP in ABC_{gain}, while the performance gaps become smaller as more gold-standard labels are available.

HF. The results of predicting HF first onset on the MGB RA cohort are provided in Figure 2B. LATTE(silver) consistently outperforms the LSTM and RETAIN when the label set size is small. SAMGEP achieves performance comparable with RETAIN and LSTM in terms of AUC and F1 scores but shows significant improvements in ABC_{gain} that become minor as more labels are available. With increasing label set sizes, the performance improvements are limited for SAMGEP but are substantial for LATTE, LSTM, and RETAIN. And the performance gaps between LATTE and SAMGEP grow, which justifies the stronger learning capability of LATTE compared with SAMGEP.

MS relapse. The results of localizing MS's relapses are provided in Figure 2C. Identifying MS relapse incidents is more challenging than the first onset of T2D and HF because MS relapse can not be well captured by a simple code or NLP concept and thus does not have a highly predictive surrogate concept. We therefore further increase the size label set size to 500 for comprehensive evaluation. LATTE(silver) is only comparable with the PheCodes, and both methods are significantly outperformed by the supervised models with 100 labels or more, which indicates that the PheCodes are non-predictive of the incidents. When available labels are less than 200, SAMGEP outperforms the LSTM and RETAIN, and when the label size grows to 500, SAMGEP is narrowly outperformed by the LSTM in ABC_{gain}, and the performance advantages of LATTE over LSTM, RETAIN, and SAMGEP become more distinct, especially in ABC_{gain}. The results on the three phenotypes also show that relying solely on the direct deployment of language models to medical notes, such as Bio-GPT, may not achieve state-of-the-art performance in incident phenotyping. Instead, unsupervised representation learning on large-scale longitudinal EHR data, such as the BEHRT,[36] are promising.

The results of evaluating model transportability are provided in Table 2. The models are trained on MGB RA cohort to identify HF onsets and evaluated on both MGB Biobank and VA-MVP. The LSTM and RETAIN methods outperform the PheCode:428 in terms of ABC_{gain}, although the PheCode only is distinctive of case/control visits and has high AUC and F1 scores on the Biobank-HF. LATTE suffers an average performance drop in ABC_{gain} by 11.9% on Biobank-HF and VA-MVP, which is significantly better than the 66.1% of LSTM and 25.6% of RETAIN.

Discrimination between case/control visits

We provide visualization of embedding vectors of case/control visits from patients with T2D or HF to show the benefits of the proposed CR and visit attention, and the necessity of sequential modeling. As shown in Figure 3, using t-distributed stochastic neighbor embedding (t-SNE),[52] we visualize the raw visit embeddings as in SAMGEP,[20] visit embedding with CR, and visit embedding with simultaneous CR and visit attention. The case visits (namely, after phenotype onset), and control visits (namely, before phenotype onset) entangle on the raw embedding space for both phenotypes. Through CR enhancements, the case/control visits become more distinguishable in the embedding space. By further incorporating the visit attention, the embedding vectors of case visits become further distinguished from control visits. The case/control visits of T2D are well separated in the embedding space even without the additional sequential modeling. It indicates that developing T2D tends to distinctly reshape the patient's clinical status. In this case, a simple classifier, such as logistic regression, would be able to well identify the onset timings. For HF, although becoming more separated via visit attention, case/control visits are still entangled by large. This indicates the necessity of

Table 2. Cross-site validation of incident phenotyping on the onset of HF (from the MGB RA cohort to Biobank-HF and VA-MVP)

Settings	Method	AUC	F1	ABC_{gain}
MGB-RA → Biobank-HF	PheCode	0.915	0.776	−4.83
	LSTM	0.928 → 0.689	0.669 → 0.588	0.635 → 0.105
	RETAIN	0.889 → 0.743	0.695 → 0.640	0.445 → 0.243
	LATTE	0.969 → 0.879	0.850 → 0.790	0.752 → 0.675
MGB-RA → VA-MVP	PheCode	0.663	0.478	0.289
	LSTM	0.928 → 0.745	0.669 → 0.638	0.635 → 0.325
	RETAIN	0.889 → 0.785	0.695 → 0.667	0.445 → 0.419
	LATTE	0.969 → 0.892	0.850 → 0.765	0.752 → 0.650

modeling visit sequential dependency to distinguish the cases from controls for precise incident localization.

Evidence for prediction interpretation

LATTE provides incident prediction with high model interpretability by indicating the healthcare concepts and visits that contribute to the predictions. As shown in Figure 4, we visualize the "evidence" at each visit that drives the onset prediction of HF. Here, the longitudinal "evidence" value is the multiplication of longitudinal concept observations, concept weights, and visit attention values. Each patient could have different longitudinal evidence, and the top 20 concepts are visualized. In Figure 4A, the patient's first HF onset is recorded by the EHR data, and the provided evidence is highly indicative of HF onset. Specifically, LATTE localizes HF onset at the fifth visit, and the top three shreds of evidence provided are "congestive HF nos" (PheCode:428.1), "congestive cardiac fail" (C0018802), and "furosemide" (C0016860), a medication used to treat fluid buildup due to HF. In Figure 4B, the patient is not observed to have HF onset over the available visits, and the top selected concepts are mostly irrelevant to HF. For example, the top three pieces of evidence provided by LATTE are "hydrops"(C0013604), "wheezing" (C0043144), and "anti-arrhythmic agent"(C0003195).

Incident phenotyping for identifying risk factors of heart failure

HF is a significant cause of morbidity and mortality in patients with RA.[53] We revisit the HF study among the MGB RA cohort described under Data and settings on the risk factors associated with HF among patients with RA.[11] The covariates considered include demographics, recent lab test results, prior medications, calendar time of RA diagnosis, comorbidities, and cardiovascular disease history before RA diagnosis (details regarding the definition and extraction of the covariates provided in the original paper[11]). Among the 9,087 RA patients in the study, 1,219 (13.2%) have at least one HF diagnosis code in their EHR after RA diagnosis. The presence of the HF ICD code is very sensitive but not quite specific to actual HF events. Exact HF status and timings for 102 patients sampled from the subset with HF diagnosis code are annotated from chart review, among which 33 (32.3%) have evidence of HF within 10 years from RA diagnosis. The 10-year prevalence of HF is thus 4.5%. In the original analysis by Huang et al.,[11] the time of the first HF code after RA diagnosis is used to define the HF time. Patients with no HF code are marked as censored at the last EHR encounter date. Four sets of analyses are employed to assess the risk factors. Two maximal follow-up times, 10 years or 5 years after RA diagnosis, are considered. Patients at risk at the maximal follow-up, defined as not having the HF code nor reaching the last EHR encounter, are censored at the maximal follow-up. Two regression methods are considered: (1) a univariate Cox model that separately regresses the risk factors with the time to HF code and (2) a multivariate Cox model that regresses all risk factors with the time to HF code.

Using the LATTE prediction on longitudinal HF incidence rate, we substitute the time to HF code with the LATTE-derived HF

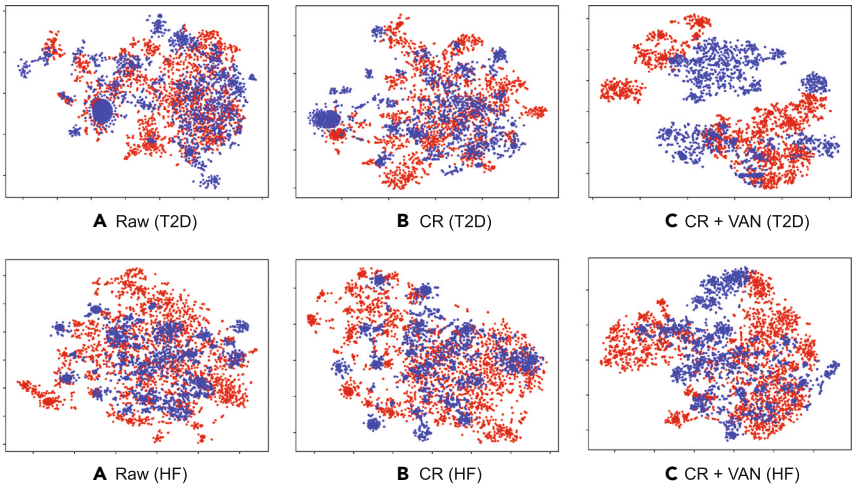


Figure 3. t-SNE visualization of visit embedding vectors on T2D and HF

(A) The visit embedding vectors aggregated based on the observed counts.

(B) The visit embedding vectors aggregated with the concept re-weighting (CR).

(C) The visit embedding vectors aggregated with both CR and visit attention network (VAN).

Blue dots denote the visits before the target incidents, and red dots denote those after the incidents.

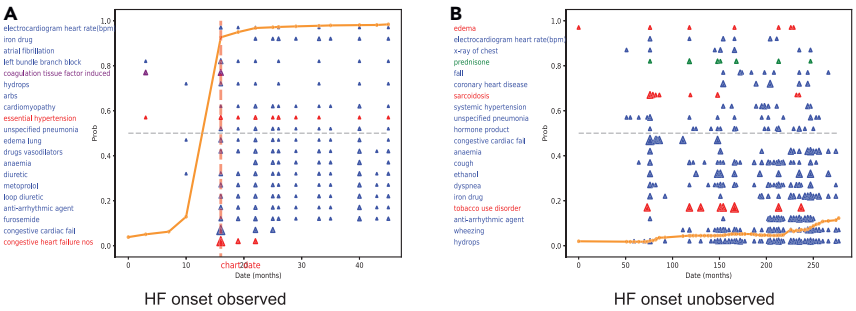


Figure 4. Prediction interpretability of LATTE

We visualize the prediction curve (orange) and longitudinal evidence that drive LATTE's incident predictions. The concepts are ranked from bottom to top by the learned importance. Red, diagnosis code; green, medication code; purple, lab test code; blue, UMLS CUIs extracted from medical notes. The chart date is the annotated date of HF onset.

onset time. According to a threshold to be described later, we define the LATTE-derived HF onset time as the first time when the longitudinal HF incidence rate from LATTE exceeds the chosen threshold. We select the threshold according to the false positive rate (FPR) for HF status at the last EHR encounter and evaluate all patients without HF code and 102 labeled patients up-weighted by the inverse labeling probability among patients with the HF code (1/0.084). To investigate the impact of the threshold, we consider 2 thresholds targeting 0.05 and 0.01 FPR.

In Table 3, we present the risk factor detection results using the time to HF code outcome and LATTE-derived HF time with 2 thresholds. The numbers in the table are the counts of risk factors whose 95% confidence intervals of relative risks do not contain one. By tightening the tolerance of FPR from 0.05 to 0.01, analysis with LATTE-derived HF time starts to detect more potential risk factors. Considering that a lower FPR means a larger threshold and, subsequently, a smaller number of derived HF events, we suggest that LATTE at 0.01 FPR might have filtered out many spurious HF events and, hence, eliminate the bias toward null induced by them. In Figure 5, we present the point estimates along with 95% confidence intervals up to 5-year follow-up. In Figure 6, we present the estimated relative efficiency of coefficient estimation from LATTE-derived HF times in comparison with that from the time to HF code outcome up to 5-year follow-up, which demonstrates a systematic advantage for LATTE-derived HF times. We present the results for the analyses up to 10-year follow-up in Figures S2 and S3 of Note S2, which similarly shows the advantages of LATTE. Another notable discovery from the analyses with LATTE-derived HF time is the decreasing trend of HF risk over calendar time. In the LATTE at 0.01 FPR analysis, patients diagnosed with RA after 2000 are associated with 55% risk reduction in univariate analysis and 61% risk reduction in the multivariate analysis compared with patients diagnosed before 2000, and patients diagnosed with RA after 2010 are associated with 85% risk reduction in both analyses compared with patients diagnosed before 2000. The finding may suggest the progress in managing cardiovascular health among patients with RA during the past decades. Such temporal trending on health outcomes has also been reported for other EHR-based studies with long observation windows.[10]

Table 3. Numbers of identified risk factors of HF among patients with rheumatoid arthritis (RA) from the MGB RA cohort

	Univariate		Multivariate	
Method	10 years	5 years	10 years	5 years
HF Code	16	18	19	19
LATTE FPR 0.05	17	18	20	19
LATTE FPR 0.01	18	19	20	20

We compare LATTE with the different false positive rates (FPRs) with rule-based methods. Univariate and multivariate Cox model analyses have been considered for data up to 5 years and 10 years after RA diagnosis.

DISCUSSION

This study proposes a computational framework based on semi-supervised learning for label-efficient incident phenotyping from longitudinal EHR data. Specifically, we develop and validate the proposed architecture, named LATTE, that identifies phenotype incident timings by learning to focus on the incident-indicative input features and patient visits and modeling the sequential dependency among visits. LATTE does not assume intensive inputs from experts for feature engineering and can perform feature selection from large-scale EHR features in a data-driven manner. It aims to be robust and scalable to all phenotypes with high clinical interpretation and directly portable across multiple clinical sites. Experimental results on the three representative phenotypes show that LATTE identifies phenotype incident timings with high label efficiency. Particularly, LATTE consistently shows significant advantages over existing methods when the annotation size is small or the visit dependency is complicated, as in MS relapse.

While the identification of binary phenotypes using EHR data has been well studied in the literature, few studies have taken it one step further to ascertain the longitudinal status of phenotype evolving over time and derive the incidence timings. Incident timings are essential for the design of many clinical studies, and effective methods to extract incident timings, like LATTE, will enable the use of EHR data to support clinical research. Examples include supporting RWE on the efficacy or safety of therapeutic drugs or intervention procedures, in which incident timings are needed to determine baseline eligibility and define time-to-event outcomes. Further, phenotype incident timings pave the way to mining temporal dependencies among progression of multiple diseases and interventions to uncover the optimal overall disease management plan.

LATTE has major clinical advantages over existing methods. (1) It's highly label-efficient by leveraging information from unlabeled data. Our highly scalable LATTE can curate accurate event times for clinical research using large EHRs with a small cost in label annotation. (2) Instead of training simultaneously on

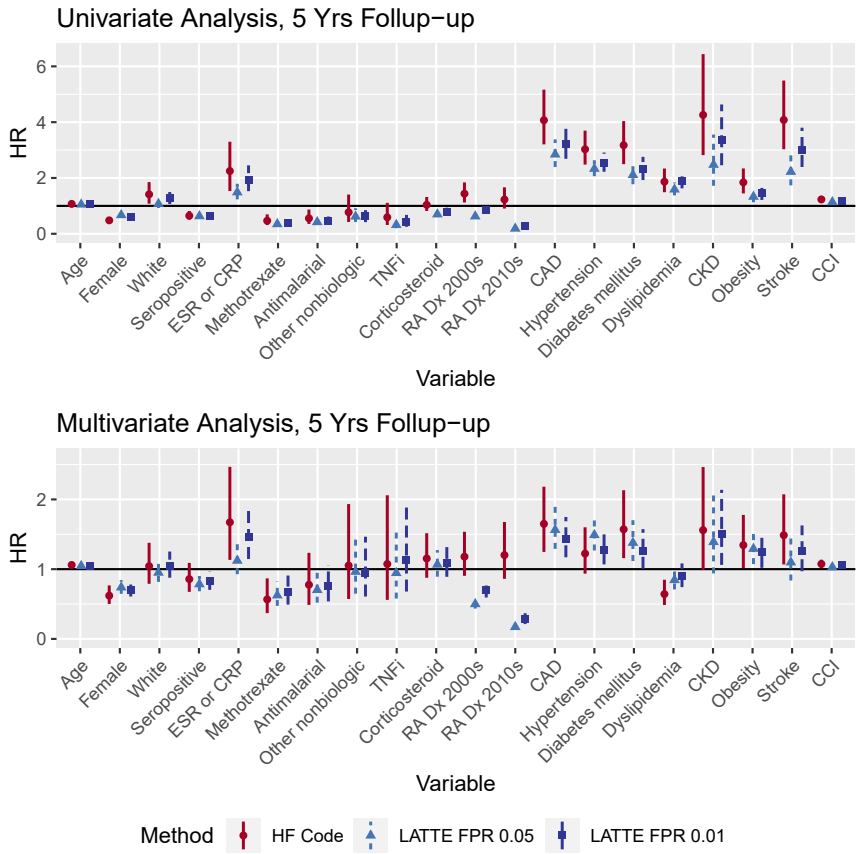


Figure 5. Hazard ratio (HR) for HF risk prediction on patients with rheumatoid arthritis (RA)

We provide the estimated hazard ratios with 95% confidence intervals for risk prediction among RA patients up to 5-year follow-up.

multiple sites, the LATTE model trained in one site enjoys stronger portability and is ready for direct application in other clinical sites, which prevents potential privacy leakage in data sharing. Event times curated by LATTE at one research center can adaptively support clinical research using EHRs at many collaborating sites. (2) By indicating which visits and concept observation drive the incident, LATTE's predictions can be easily used to facilitate incident annotations. Specifically, the LATTE model trained using only the silver-standard labels achieves performance comparable with supervised models with hundreds of labels on HF and T2D, and leveraging LATTE predictions would substantially reduce the annotation expenses. The architecture of LATTE can also guide feature engineering in future clinical studies using EHRs.

LLMs are gaining increasing attention in EHR studies. Existing LLMs for EHR are mostly focused on medical notes,[51,54] with a few exceptions for codified data, such as BEHRT[36] and MedBERT.[37] These models are typically trained in a phenotype-agnostic manner and have limited ability to accurately extract event timing information from a large number of clinical notes. First, clinical event information is embedded in potentially a large number of longitudinally recorded notes, which is a much larger context than what typical LLMs can do (in terms of how many tokens the LLMs can analyze). Second, existing LLMs are not yet well trained to leverage temporality information across a potentially long time horizon. In addition, codified EHR data provide valuable information but cannot be effectively leveraged by LLMs that are trained only based on narrative text data. In the experimental parts, we show that two LLM-based approaches (BioGPT and BEHRT) are consistently outperformed by LATTE. After all, ascertaining clinical event time is a challenging task.

Because LATTE aggregates medical information from various sources, the successful application of LATTE is limited by the availability and quality of these resources. First, LATTE requires pre-learned concept embedding vectors for EHR features to screen predictive features and construct the visit embeddings. The training of such embedding vectors from co-occurrence data is usually time- and resource-consuming and, thus, must be conducted separately. While we proposed to utilize published concept embeddings, caution should be taken regarding the potential transferable issues due to heterogeneity in encoding and documentation patterns across healthcare systems and calendar time. Moreover, some phenotypes, like rare or novel diseases, can be poorly represented or even absent in the global concept embeddings dominated by common phenotypes. To generalize LATTE for rare or novel phenotypes, concept embeddings based on LLMs that are trained based on expert-curated knowledge[55,56] can be considered. Second, effective pre-training of LATTE relies on one reasonably indicative silver standard surrogate for the target phenotype. Identification of a silver standard surrogate is straightforward for diseases with corresponding diagnosis codes but challenging for any phenotypes poorly structured in EHRs; e.g., novel diseases like long coronavirus disease 2019 (COVID-19) or assessments not regularly performed in practice, like RECIST.[57] Powered by the advancement in language models, specialized NLP tools can be used to classify event status from narrative notes and thus provide silver-standard surrogates for LATTE to calibrate against gold-standard labels. Third, LATTE uses patient-level healthcare data within a single healthcare institute. For large-scale studies conducted by multiple institutions, regulations limit the share of patient-level data across institutions to protect patient privacy. Development of a federated learning strategy for LATTE will be necessary. LATTE can also be extended toward broader types of input data and outcome variables. Fourth, LATTE currently only accepts the counts of the observed concepts and is not examined on other types of multimodal EHR data, such as numeric test values or images. The codes or mentions of relevant lab tests or diagnostic images may adequately indicate broad phenotypes, but identification of sub-phenotypes may require specific numerical or image data.

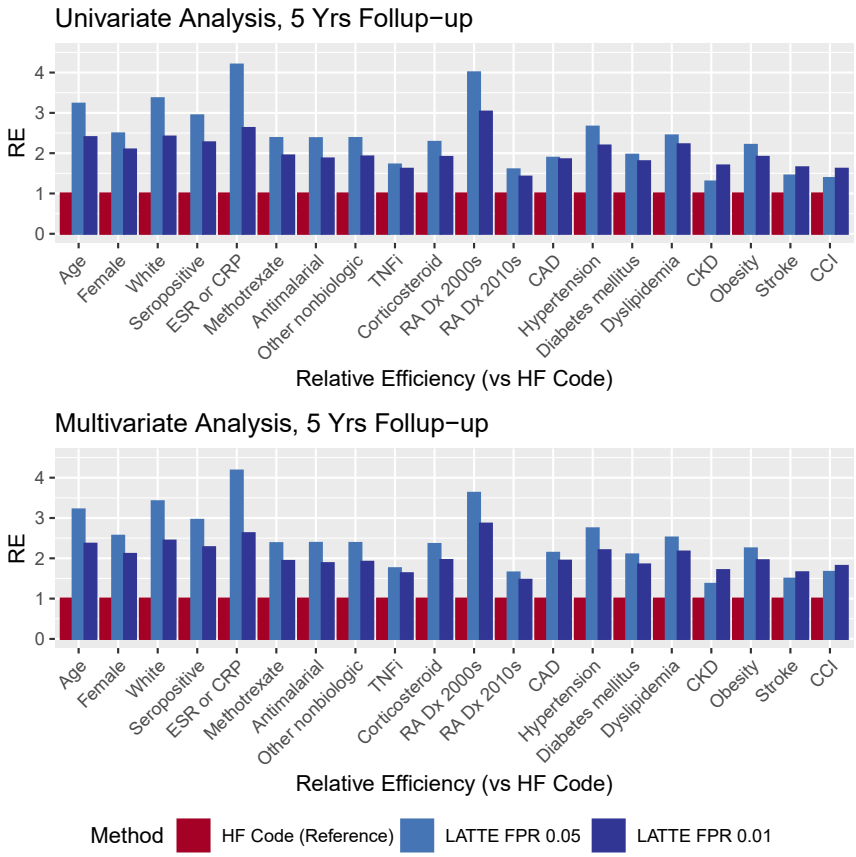


Figure 6. Estimated relative efficiency of coefficient estimation

We provide the estimated relative efficiency of analyses with LATTE-derived HF timings versus analysis with HF diagnosis code-derived outcomes in HR risk prediction among RA patients up to 5-year follow-up. LATTE-derived outcomes achieve systematically improved efficiency.

To incorporate the additional data types, multiple modules for profiling longitudinal test values and images can be created in parallel with the CR module, whose outputs would be concatenated with the visit embedding to comprehensively describe each visit. Fifth, LATTE focuses on the binary events but not other types of clinical data. As a related but distinct task, ascertaining the longitudinal profile of key markers, such as DAS28-CRP, measuring disease severity of RA, is fundamental for evaluating patient outcomes. Combining the architecture of LATTE with the appropriate loss function according to suitable statistical models (e.g., marked point process) may redirect LATTE for a wide variety of clinical variables.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Tianxi Cai (tcai@hsph.harvard.edu).

Materials availability

This study did not generate any physical materials.

Data and code availability

The clinical data reported in this study cannot be deposited in a public repository due to regulations on protected health information (PHI). All source codes with simulated example data have been published on Zenodo[58] (also at https://github.com/celehs/LATTE/). We provide the pre-trained incident phenotyping models for the 3 representative phenotypes and the embedding vectors and weights of selected EHR concepts upon request. Any information required to reanalyze the data reported in this paper is available from the lead contact upon reasonable request.

Implementation details

The concept reweighting module consists of three layers, with the first layer containing two branches, each with shared 32 units, and the fusion layer with 16 units, followed by the output layer with one unit. The VAN uses 32-dimensional query and key vectors. For T2D and HF, we use one GRU layer containing 32 units and for MS relapse two layers. For HF and T2D, the model has 51,000 parameters and 44 million flops and for MS 56,000 parameters and 47 million flops. For the silver-standard label construction defined in Eqn Equation 1 and Eqn Equation 2, we set $\tau = 0.1$ and $\alpha = 0.2$. For improved cross-site portability, we set the distance tolerance $c = 10$ and visit shifts $L_{\max} = 7$. In the final objective for cross-site training, we set $\gamma = 0.1$, $\lambda = 0.5$, and $\kappa = 0.1$ to balance those objectives. We provide their sensitivity analysis in Figure S1 and Note S1. The whole model is optimized end to end using the Adam optimizer with a learning rate of 0.001.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2023.100906.

ACKNOWLEDGMENTS

This research is supported by U.S. Food and Drug Administration under grant U01 FD007929, U.S. National Institutes of Health under grant P30 ARAR72577, and the Million Veteran Program (#MVP000), Office of Research and Development, U.S. Veterans Health Administration. This publication does not represent the views of the U.S. Department of Veterans Affairs or the U.S. Government.

AUTHOR CONTRIBUTIONS

Conceptualization, Tianxi Cai; methodology, J.W., Tianxi Cai, and J.H.; data processing and analysis, J.W., J.H., C.-L.B., V.M.C., V.A.P., D.W., Tianrun Cai, C.H., Y.-L.H., V.S.G., J.M.G., and L.C.; project administration, K.C.; writing, J.W., Tianxi Cai, J.H., C.-L.B., Y.Z., K.P.L., J.L., and Y.-L.H.; guarantor, Tianxi Cai; approval of final manuscript, all authors.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: April 27, 2023
Revised: September 6, 2023
Accepted: December 1, 2023
Published: December 27, 2023

REFERENCES

1. Kohane, I.S., Churchill, S.E., and Murphy, S.N. (2012). A translational engine at the national scale: informatics for integrating biology and the bedside. J. Am. Med. Inf. Assoc. *19*, 181–185.
2. Miotto, R., Li, L., Kidd, B.A., and Dudley, J.T. (2016). Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Sci. Rep. *6*, 26094–26110.
3. Ananthakrishnan, A.N., Cai, T., Savova, G., Cheng, S.C., Chen, P., Perez, R.G., Gainer, V.S., Murphy, S.N., Szolovits, P., Xia, Z., et al. (2013). Improving case definition of crohn's disease and ulcerative colitis in electronic medical records using natural language processing: a novel informatics approach. Inflamm. Bowel Dis. *19*, 1411–1420.
4. Liao, K.P., Cai, T., Gainer, V., Goryachev, S., Zeng-treitler, Q., Raychaudhuri, S., Szolovits, P., Churchill, S., Murphy, S., Kohane, I., et al. (2010). Electronic medical records for discovery research in rheumatoid arthritis. Arthritis Care Res. *62*, 1120–1127.
5. Murphy, S.N., et al. (2006). Integration of clinical and genetic data in the i2b2 architecture. AMIA Annual Symposium Proceedings, *2006* (American Medical Informatics Association), p. 1040.
6. Roden, D.M., Pulley, J.M., Basford, M.A., Bernard, G.R., Clayton, E.W., Balser, J.R., and Masys, D.R. (2008). Development of a large-scale de-identified dna biobank to enable personalized medicine. Clin. Pharmacol. Ther. *84*, 362–369.
7. Gamerman, V., Cai, T., and Elsäßer, A. (2019). Pragmatic randomized clinical trials: best practices and statistical guidance. Health Serv. Outcome Res. Methodol. *19*, 23–35.
8. Hernandez-Boussard, T., Monda, K.L., Crespo, B.C., and Riskin, D. (2019). Real world evidence in cardiovascular medicine: ensuring data validity in electronic health record-based studies. J. Am. Med. Inf. Assoc. *26*, 1189–1194.
9. Hou, J., Kim, N., Cai, T., Dahal, K., Weiner, H., Chitnis, T., Cai, T., and Xia, Z. (2021). Comparison of dimethyl fumarate vs fingolimod and rituximab vs natalizumab for treatment of multiple sclerosis. JAMA Netw. Open *4*, e2134627.
10. Hou, J., Zhao, R., Cai, T., Beaulieu-Jones, B., Seyok, T., Dahal, K., Yuan, Q., Xiong, X., Bonzel, C.L., Fox, C., et al. (2022). Temporal trends in clinical evidence of 5-year survival within electronic health records among patients with early-stage colon cancer managed with laparoscopy-assisted colectomy vs open colectomy. JAMA Netw. Open *5*, e2218371.
11. Huang, S., Cai, T., Weber, B.N., He, Z., Dahal, K.P., Hong, C., Hou, J., Seyok, T., Cagan, A., DiCarli, M.F., et al. (2023). Association between inflammation, incident heart failure, and heart failure subtypes in patients with rheumatoid arthritis. Arthritis Care Res. *75*, 1036–1045.
12. Hassett, M.J., Uno, H., Cronin, A.M., Carroll, N.M., Hornbrook, M.C., and Ritzwoller, D. (2017). Detecting lung and colorectal cancer recurrence using structured clinical/administrative data to enable outcomes research and population health management. Med. Care *55*, e88–e98.
13. Uno, H., Ritzwoller, D.P., Cronin, A.M., Carroll, N.M., Hornbrook, M.C., and Hassett, M.J. (2018). Determining the time of cancer recurrence using claims or electronic medical record data. JCO Clin. Cancer Inform. *2*, 1–10.
14. Ahuja, Y., et al. (2020). surelda: A multidisease automated phenotyping method for the electronic health record. J. Am. Med. Inf. Assoc. *27*, 1235–1243.
15. Kirby, J.C., Speltz, P., Rasmussen, L.V., Basford, M., Gottesman, O., Peissig, P.L., Pacheco, J.A., Tromp, G., Pathak, J., Carrell, D.S., et al. (2016). Phekb: a catalog and workflow for creating electronic phenotype algorithms for transportability. J. Am. Med. Inf. Assoc. *23*, 1046–1052.
16. Liao, K.P., Sun, J., Cai, T.A., Link, N., Hong, C., Huang, J., Huffman, J.E., Gronsbell, J., Zhang, Y., Ho, Y.L., et al. (2019). High-throughput multimodal automated phenotyping (map) with application to phewas. J. Am. Med. Inf. Assoc. *26*, 1255–1262.
17. Newton, K.M., Peissig, P.L., Kho, A.N., Bielinski, S.J., Berg, R.L., Choudhary, V., Basford, M., Chute, C.G., Kullo, I.J., Li, R., et al. (2013). Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the emerge network. J. Am. Med. Inf. Assoc. *20*, e147–e154.
18. Yu, S., Ma, Y., Gronsbell, J., Cai, T., Ananthakrishnan, A.N., Gainer, V.S., Churchill, S.E., Szolovits, P., Murphy, S.N., Kohane, I.S., et al. (2018). Enabling phenotypic big data with phenorm. J. Am. Med. Inf. Assoc. *25*, 54–60.
19. Chubak, J., Yu, O., Pocobelli, G., Lamerato, L., Webster, J., Prout, M.N., Ulcickas Yood, M., Barlow, W.E., and Buist, D.S.M. (2012). Administrative data algorithms to identify second breast cancer events following early-stage invasive breast cancer. J. Natl. Cancer Inst. *104*, 931–940.
20. Ahuja, Y., et al. (2022). A semi-supervised adaptive markov gaussian embedding process (samgep) for prediction of phenotype event times using the electronic health record. Sci. Rep. *12*, 1–12.
21. Choi, E., et al. (2016). Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. Adv. Neural Inf. Process. Syst. *29*, 1–9.
22. Badger, J., LaRose, E., Mayer, J., Bashiri, F., Page, D., and Peissig, P. (2019). Machine learning for phenotyping opioid overdose events. J. Biomed. Inform. *94*, 103185.
23. Shickel, B., Tighe, P.J., Bihorac, A., and Rashidi, P. (2018). Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. IEEE J. Biomed. Health Inform. *22*, 1589–1604.
24. Si, Y., Du, J., Li, Z., Jiang, X., Miller, T., Wang, F., Jim Zheng, W., and Roberts, K. (2021). Deep representation learning of patient data from electronic health records (ehr): A systematic review. J. Biomed. Inform. *115*, 103671.
25. Yang, S., Varghese, P., Stephenson, E., Tu, K., and Gronsbell, J. (2023). Machine learning approaches for electronic health records phenotyping: a methodical review. J. Am. Med. Inf. Assoc. *30*, 367–381.
26. Chang, M., Womer, F.Y., Gong, X., Chen, X., Tang, L., Feng, R., Dong, S., Duan, J., Chen, Y., Zhang, R., et al. (2021). Identifying and validating subtypes within major psychiatric disorders based on frontal–posterior functional imbalance via deep learning. Mol. Psychiatr. *26*, 2991–3002.
27. Lee, C., and Van Der Schaar, M. (2020). Temporal phenotyping using deep predictive clustering of disease progression. In International Conference on Machine Learning (PMLR), pp. 5767–5777.
28. Choi, E., Xu, Z., Li, Y., Dusenberry, M., Flores, G., Xue, E., and Dai, A. (2020). Learning the graphical structure of electronic health records with graph convolutional transformer. Proc. AAAI Conf. Artif. Intell. *34*, 606–613.
29. Ayala Solares, J.R., Diletta Raimondi, F.E., Zhu, Y., Rahimian, F., Canoy, D., Tran, J., Pinho Gomes, A.C., Payberah, A.H., Zottoli, M., Nazarzadeh, M., et al. (2020). Deep learning for electronic health records: A comparative review of multiple deep neural architectures. J. Biomed. Inform. *101*, 103337.

30. Beaulieu-Jones, B.K., and Greene, C.S.; Pooled Resource Open-Access ALS Clinical Trials Consortium (2016). Semi-supervised learning of the electronic health record for phenotype stratification. J. Biomed. Inform. *64*, 168–178.

31. Nogues, I.-E., Wen, J., Lin, Y., Liu, M., Tedeschi, S.K., Geva, A., Cai, T., and Hong, C. (2022). Weakly semi-supervised phenotyping using electronic health records. J. Biomed. Inform. *134*, 104175.

32. Poulain, R., et al. (2022). Few-shot learning with semi-supervised transformers for electronic health records. In Machine Learning for Healthcare Conference (PMLR), pp. 853–873.

33. Zang, C., Goodman, M., Zhu, Z., Yang, L., Yin, Z., Tamas, Z., Sharma, V.M., Wang, F., and Shao, N. (2022). Development of a screening algorithm for borderline personality disorder using electronic health records. Sci. Rep. *12*, 11976.

34. Liu, C., et al. (2015). Temporal phenotyping from longitudinal electronic health records: A graph based framework. In Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, pp. 705–714.

35. Lee, C., Rashbass, J., and van der Schaar, M. (2021). Outcome-oriented deep temporal phenotyping of disease progression. IEEE Trans. Biomed. Eng. *68*, 2423–2434.

36. Li, Y., Rao, S., Solares, J.R.A., Hassaine, A., Ramakrishnan, R., Canoy, D., Zhu, Y., Rahimi, K., and Salimi-Khorshidi, G. (2020). Behrt: transformer for electronic health records. Sci. Rep. *10*, 7155.

37. Rasmy, L., Xiang, Y., Xie, Z., Tao, C., and Zhi, D. (2021). Med-bert: pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ Digit. Med. *4*, 86.

38. Wanyan, T., Honarvar, H., Jaladanki, S.K., Zang, C., Naik, N., Somani, S., De Freitas, J.K., Paranjpe, I., Vaid, A., Zhang, J., et al. (2021). Contrastive learning improves critical event prediction in covid-19 patients. Patterns *2*, 100389.

39. Zang, C., and Wang, F. (2021). Scehr: Supervised contrastive learning for clinical risk prediction using electronic health records. Proceedings. IEEE International Conference on Data Mining, 857. NIH Public Access.

40. Hong, C., Rush, E., Liu, M., Zhou, D., Sun, J., Sonabend, A., Castro, V.M., Schubert, P., Panickan, V.A., Cai, T., et al. (2021). Clinical knowledge extraction via sparse embedding regression (keser) with multi-center large scale electronic health record data. NPJ Digit. Med. *4*, 151–211.

41. Wen, J., Zhang, X., Rush, E., Panickan, V.A., Li, X., Cai, T., Zhou, D., Ho, Y.L., Costa, L., Begoli, E., et al. (2023b). Multimodal representation learning for predicting molecule–disease relations. Bioinformatics *39*, btad085.

42. Zhou, D., Gan, Z., Shi, X., Patwari, A., Rush, E., Bonzel, C.L., Panickan, V.A., Hong, C., Ho, Y.L., Cai, T., et al. (2022). Multiview incomplete knowledge graph integration with application to cross-institutional ehr data harmonization. J. Biomed. Inform. *133*, 104147.

43. Levy, O., and Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. Adv. Neural Inf. Process. Syst. *27*.

44. Mikolov, T., et al. (2013). Distributed representations of words and phrases and their compositionality. Adv. Neural Inf. Process. Syst. *26*, 1–11.

45. Beam, A.L., et al. (2019). Clinical concept embeddings learned from massive sources of multimodal medical data. In PACIFIC SYMPOSIUM ON BIOCOMPUTING 2020 (World Scientific), pp. 295–306.

46. Hou, J., Chan, S.F., Wang, X., and Cai, T. (2023). Risk prediction with imperfect survival outcome information from electronic health records. Biometrics *79*, 190–202.

47. Vaswani, A., et al. (2016). Attention is all you need. Adv. Neural Inf. Process. Syst. *30*, 1–11.

48. Kenton, J.D.M.-W.C., and Toutanova, L.K. (2019). Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding (Proceedings of NAACL-HLT), pp. 4171–4186.

49. Arnab, A., et al. (2021). Vivit: A video vision transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6836–6846.

50. Nayebi Kerdabadi, M., et al. (2023). Contrastive learning of temporal distinctiveness for survival analysis in electronic health records. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, pp. 1897–1906.

51. Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., and Liu, T.Y. (2022). Biogpt: generative pre-trained transformer for biomedical text generation and mining. Briefings Bioinf. *23*, bbac409.

52. Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-sne. J. Mach. Learn. Res. *9*.

53. Nicola, P.J., Maradit-Kremers, H., Roger, V.L., Jacobsen, S.J., Crowson, C.S., Ballman, K.V., and Gabriel, S.E. (2005). The risk of congestive heart failure in rheumatoid arthritis: a population-based study over 46 years. Arthritis Rheum. *52*, 412–420.

54. Alsentzer, E., et al. (2019). Publicly available clinical bert embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop (Association for Computational Linguistics), pp. 72–78.

55. Beltagy, I., et al. (2019). Scibert: A pretrained language model for scientific text. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)), pp. 3615–3620.

56. Yuan, Z., Zhao, Z., Sun, H., Li, J., Wang, F., and Yu, S. (2022). Coder: Knowledge-infused cross-lingual medical term embedding for term normalization. J. Biomed. Inform. *126*, 103983.

57. Eisenhauer, E.A., Therasse, P., Bogaerts, J., Schwartz, L.H., Sargent, D., Ford, R., Dancey, J., Arbuck, S., Gwyther, S., Mooney, M., et al. (2009). New response evaluation criteria in solid tumours: revised recist guideline (version 1.1). Eur. J. Cancer *45*, 228–247.

58. Wen, J., et al. (2023a). jungel2star/latte: Latte: Label-Efficient Incident Phenotyping from Longitudinal Electronic Health Records (Phenotyping). in.