Artificial Intelligence Review (2024) 57:32
https://doi.org/10.1007/s10462-023-10677-z

Transformers in health: a systematic review on architectures for longitudinal data analysis

Clauirton A. Siebra[1,2] · **Mascha Kurpicz-Briki**[3] · **Katarzyna Wac**[1]

Published online: 3 February 2024

Abstract
Transformers are state-of-the-art technology to support diverse Natural Language Processing (NLP) tasks, such as language translation and word/sentence predictions. The main advantage of transformers is their ability to obtain high accuracies when processing long sequences since they avoid the vanishing gradient problem and use the attention mechanism to maintain the focus on the information that matters. These features are fostering the use of transformers in other domains beyond NLP. This paper employs a systematic protocol to identify and analyze studies that propose new transformers' architectures for processing longitudinal health datasets, which are often dense, and specifically focused on physiological, symptoms, functioning, and other daily life data. Our analysis considered 21 of 456 initial papers, collecting evidence to characterize how recent studies modified or extended these architectures to handle longitudinal multifeatured health representations or provide better ways to generate outcomes. Our findings suggest, for example, that the main efforts are focused on methods to integrate multiple vocabularies, encode input data, and represent temporal notions among longitudinal dependencies. We comprehensively discuss these and other findings, addressing major issues that are still open to efficiently deploy transformers architectures for longitudinal multifeatured healthcare data analysis.

Keywords Transformers · Deep learning · Health datasets · Longitudinal analysis

✉ Clauirton A. Siebra
clauirton.dealbuquerque@unige.ch

Mascha Kurpicz-Briki
mascha.kurpicz@bfh.ch

Katarzyna Wac
katarzyna.wac@unige.ch

[1] Quality of Life Technologies Lab, University of Geneva, 1227 Carouge, Switzerland

[2] Informatics Center, Federal University of Paraiba, João Pessoa 58055-000, Brazil

[3] Applied Machine Intelligence Research Group, Bern University of Applied Sciences, Höheweg 80, 2502 Biel/Bienne, Switzerland

1 Introduction

Longitudinal data implies continuous assessments repeated over time. This type of data is common in the health area, and its major advantage is its capacity to separate cohort and temporal effects in the context of the analyses (Diggle et al. 2002). For example, longitudinal data is part of clinical studies that follow a group of patients with diabetes over five years to track changes in their blood sugar levels and complications. Longitudinal data contrast with cross-sectional data in which a single outcome is measured for each individual. An example of cross-sectional study may determine the prevalence of hypertension among adults living in a specific metropolitan area, by collecting data at a single point in time and providing a snapshot of the population's hypertension status, rather than following all the individuals over time. Thus, longitudinal data analysis can generate important conclusions for health personnel from a temporal perspective. For example, the study of Zhao et al. (2019) relies on longitudinal Electronic Health Records (EHR) and genetic data to create a model for 10-years cardiovascular disease event prediction. Similarly, Severson et al. (2021) employed longitudinal data collected for up to seven years to develop a Parkinson's disease progression model for intra-individual and inter-individual variability and medication effects. Perveen et al. (2020) aimed to create models that provide predictions concerning the future condition of pre-diabetic individuals. They exploited sequences of clinical measurements obtained from longitudinal data from a sample of patients. According to all these studies, prognostic modeling techniques are important decision support tools to identify a prior patients' health status and characterize progression patterns. In other words, they support the health personnel by predicting future health conditions that could guide the implementation of preventive and adequate interventions.

These and other recent research efforts have in common the use of machine learning techniques. The main example is the family of deep learning recurrent neural networks (RNNs), specially designed to provide a tractable solution to handle longitudinal data (Mao and Sejdić 2022). RNNs support tasks such as sequence classification, anomaly detection, decision-making, and status prediction. These tasks rely on identifying temporal patterns and modeling nonstationary dynamics of human contexts (e.g., physical, physiological, mental, social, and environmental), providing a way to understand complex time variations and dependencies. The literature brings several derivations of RNNs already employed in the health area. For example, long short-term memory (LSTM) networks are a type of RNN capable of learning order dependence in long sequence prediction problems. Guo et al. (2021) used LSTM models to predict future cardiovascular health levels based on previous measurements from longitudinal electronic health record (EHR) data. Gated Recurrent Units (GRU) are derivations of RNN that use gates to control the flow of information, deciding what information should be passed to the output. GRU was used, for example, for early detection of post-surgical complications using longitudinal EHR data (Chen et al. 2021a). Bidirectional RNNs can analyze longitudinal data in both directions, and they were used, for example, to detect medical events (e.g., adverse drug events) in EHR (Jagannatha and Yu 2016). This type of RNNs offers advantages in the health domain due to their ability to capture data from past and future time intervals. For example, consider the heart rate (HR) anomaly detection task. While unidirectional RNNs only consider data points from the past to detect an eventual problem, bidirectional RNNs consider both past and future heart rate measurements for each data point in an HR sequence. Thus, this approach allows a better understanding of the temporal context and dynamics of the patient's

heart rate. Some studies also use different RNN derivations in the same problem (e.g., longitudinal prediction modeling of Alzheimer's disease) to identify the best strategy in terms of accuracy (Tabarestani et al. 2019).

RNNs have a strong ability to deal with longitudinal data since their recurrent architecture can remember past information and utilize it to make predictions at future time steps. Therefore, RNNs can model the temporal dependencies between parts of the longitudinal data and understand how they evolve over time. However, they cannot keep up with context and content for longer sequences. The main reason is that RNNs suffer from the *vanishing gradient problem* since long-term information must sequentially travel through all RNN units before generating results. Thus, such information will likely vanish by being multiplied many times by small values. RNN-like networks, such as LSTM and GRU, consider this problem, but their more complex architectures still present sequential paths from older past units until the final one. For example, mHealth applications daily assess different multifeature longitudinal data of their users, which generate multi-longitudinal sequences of health data (Wac 2016). In five years, the longitudinal data sequence will have about 1825 timesteps. According to experimental results (Culurciello 2018), RNNs are good for remembering sentences in the order of hundreds but not thousands of timesteps. Moreover, this sequential flow through RNN units also brings performance problems since RNNs must process data sequentially. Thus, they cannot employ parallel computing hardware and graphics processing units (GPU) in training and inference.

Transformers (Vaswani et al. 2017) are a recent type of deep neural network focused on analyzing sequences. Their architecture allows the processing of entire sequences in parallel. Consequently, it is possible to scale the speed and capacity of such processing compared to previous RNN-like approaches. Moreover, transformers introduced the *attention mechanism*, which considers the relationship between attributes, irrespective of where they are placed in a sequence. This mechanism allows tracking of the relations between attributes across long sequences in both forward and reverse directions. While transformers were originally conceived to cover traditional problems of the Natural Language Processing (NLP) area, such as text classification and named entity recognition, their high performance in dealing with sequential data encouraged the adaptation of this architecture to other areas that involve the analysis of longitudinal data, such as digital health.

This paper uses a systematic review protocol to identify and analyze studies that proposed adaptations for transformers' architectures so they can handle longitudinal health data. This protocol contains a set of research questions that guide the analysis of these approaches regarding their architectures, input vocabulary, aims, positional embedding implementations, explainability, and other technical aspects. Moreover, this analysis also allows the identification of trends in the area, main limitations, and opportunities for research directions. Thus, our main contribution is to consolidate a body of knowledge that supports advances in longitudinal health data analysis using transformers.

2 Transformers background

This section summarizes the main concepts of transformers intending to enable a better understanding of this review. These concepts are also used to formulate the research questions of the research protocol (Sect. 3).

2.1 Encode-decode architectures

The original transformer architecture (Fig. 1) is composed of two modules: an encoder (Fig. 1, left) and a decoder (Fig. 1, right). The basic unit of an encoder (Nx) has two sub-modules: a multi-head attention (composed of multiple self-attention layers), followed by a fully connected network (Feed Forward). Normalization layers and residual connections are also used in both sub-modules to stabilize the network over training. The encoder function is to extract features from input sequences. To that end, the complete sequence is parallelly processed at once. Each token of the sequence flows through its own path inside the architecture but maintains dependencies on the paths of the other tokens. This strategy enriches each token with contextual information from the whole sentence. Encoder units (Nx) can be stacked on each other according to the task. For example, Bidirectional

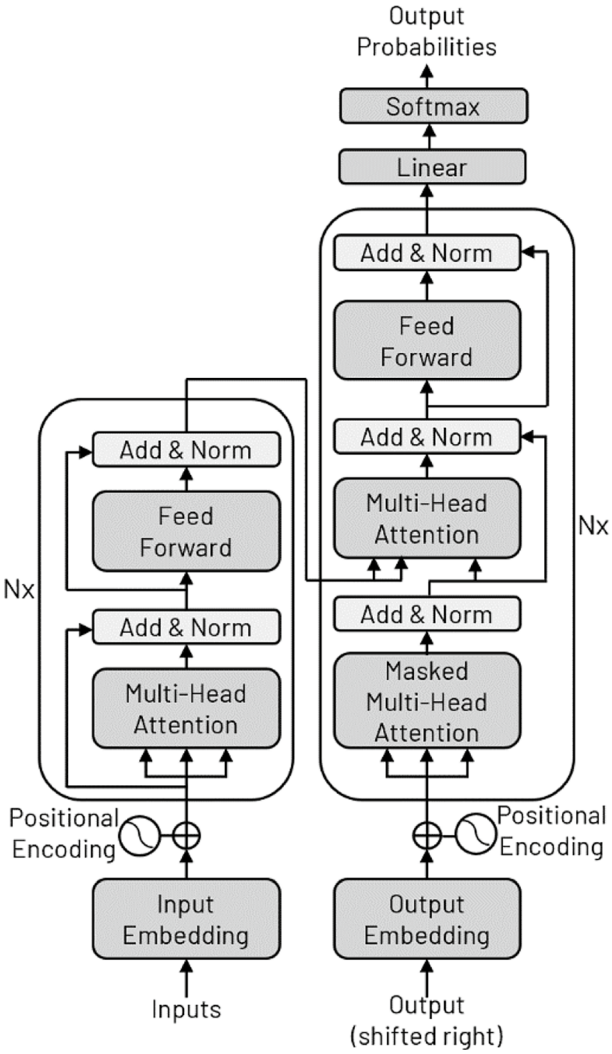

Fig. 1 The original transformer encoder-decoder architecture (Vaswani et al. 2017). The left-side block represents the encoder, while the right-side block represents the decoder

Encoder Representations from Transformers (BERT) uses a stack of 12 encoders in its basic version (Devlin et al. 2018).

The decoder and encoder architectures are very similar, and decoder units (Nx) can also be stacked on each other. However, the decoder outputs one token at a time since each output token becomes part of the next decoder input (auto-regressive process). Over this process, the vector of features from the encoder supports the decoder in focusing on the appropriate positions of the input sequence. We suggest the original paper "Attention is All You Need" (Vaswani et al. 2017) for details about these components.

Transformers can be classified into three types according to the modules that they implement: encoder-only (only the encoder module is implemented), decoder-only (only the decoder module is implemented), and encoder-decoder (the complete architecture is implemented). Apart from this classification, architectural designs can modify one or more layers according to the target task.

2.2 Vocabulary

Transformers were initially designed to support the solution of natural language processing problems. Thus, their inputs and outputs are tokens (words) that compose a vocabulary. For example, the illustrated Softmax layer (Fig. 1) produces probabilities over an output vocabulary during a language translation task. In NLP problems, this vocabulary corresponds to the lexicon of a language such as English or French. Similarly, programming language code generation using transformers (Svyatkovskiy et al. 2020) employs the tokens of the programming language as vocabulary. Observe that vocabularies are neither compulsory nor necessarily composed of textual tokens. For example, the study in Bao et al. (2021) shows that patches of an original image are employed as visual tokens. Therefore, like natural language, images are represented as sequences of discrete tokens obtained by an *image tokenizer*. Transformers that use health images as input are out of the scope of this paper. A review on this subject can be seen in He et al. (2022).

2.3 Input embedding

The input embedding layer is a type of lookup table that contains vectorial representations of input data (e.g., each term of the vocabulary). This layer is essential because transformers process vectors of continuous values like any other machine learning algorithm. There are several proposals for input embeddings, which can be classified into context-independent (traditional) and context-dependent (contextualized) embeddings (Wang et al. 2020). While the former produces unique and distinct representations for each token without considering its context; the latter learns different embeddings for the same token according to its context (Fig. 2). Transformers also allow each sequential input to contain multiple embeddings. This approach is essentially a kind of early fusion (sum or concatenation) of embeddings. The papers discussed in our review bring several examples in such a direction.

2.4 Positional encoding

The positional encoding layer adds a positional vector to each set of inputs assessed simultaneously. This step is essential since transformers process all the inputs in parallel, unlike RNN or LTSM approaches, where inputs are fed in sequence. While these techniques do

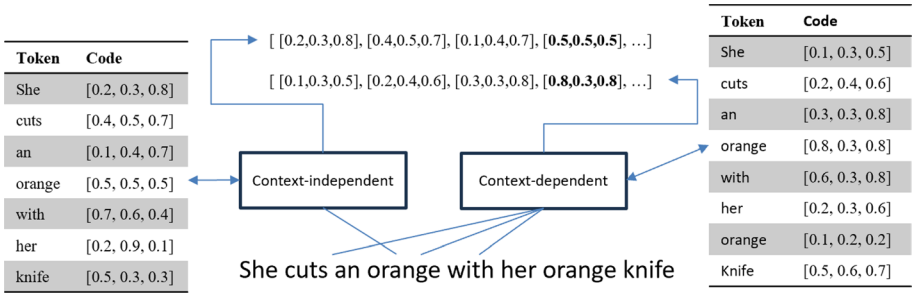

Fig. 2 Difference between context-independent and dependent embeddings. The left case shows that each word has only one entry in the lookup table. On the right, the same token can have multiple entries according to its context (neighborhood tokens indicate if orange is a noun or an adjective)

not require any specific positional strategy because they already understand the sequence of the inputs, transformers need this additional information. The use of sequential indexes is the simplest strategy to generate positional encodings. It works well if the number of sequential inputs is low. However, it becomes problematic as the number of inputs increases since the high values of the latter positional encodings may dominate the initial values, distorting the final results. A simple solution is to convert the encoding values as a fraction of length, i.e., index/w, where w is the number of inputs. However, several applications do not know this w value a priori. Strategies such as frequency-based positional encodings (Vaswani et al. 2017) avoid this issue. The study in Dufter (2021) brings a comprehensive discussion about positional encoding.

2.5 Target task

Transformers are used to different tasks, which usually affect how their architectures are designed. Some tasks, such as question answering and text translation, are specific for natural language processing. Apart from that, tasks that can use longitudinal data are:

- Classification of sequences: This task usually employs architectures that have the softmax as the activation function of their final layer. Such a function returns probabilities over pre-defined classes according to the input sequence (Prakash et al. 2021).
- Language modeling: This is the traditional strategy used to train transformers models. The aim is to fit a model to a corpus, which normally is domain specific (Yang et al. 2022). This strategy is also useful for inputting missing data on health records (Liu et al. 2023).
- Prediction: Given a sequence of tokens, the aim is to predict the next token of this sequence. Thus, architectures are designed to only attend to the left context (tokens on the left of the current position). The prognostic of a disease is an example of this task (Florez et al. 2021).

2.6 Training strategies

The original training process of transformers is conducted in two phases. First, they are pretrained using strategies such as Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) (Devlin 2018). The transformer model is trained to predict masked words

in the former strategy. In the latter strategy, the model must predict whether two sentences follow each other. After the pretraining, the model is fine-tuned to a specific task, replacing, for example, part of the language model with additional linear layers (traditional machine learning training). However, transformer architectures can also be designed to avoid the pretraining phase. In this case, they directly learn using backward propagation, such as the architecture described in Dong et al. (2021) that predicts early signs of generalized anxiety disorder and depression.

2.7 Validation and comparisons

A common approach to validate the results obtained using transformers is to compare such results with outcomes of other machine learning methods. This approach was found in almost all the reviewed papers, as discussed later in this paper (Sect. 4.5). Indeed, longitudinal data analysis in health does not have a "gold standard" machine learning method that could be used as a comparisons baseline. The choice of this depends on the specific research question, the nature of the health data, and the goals of the analysis. A useful way to verify the effectiveness of transformers is by comparing their results with outcomes generated by clinicians. However, only two papers (Li et al. 2020; Rasmy et al. 2021) employed clinicians during the validation process. Section 4.5 discusses this topic in more detail.

3 Research method

A systematic review protocol, as stated by Kitchenham (2004), was used as the research method in this paper. The steps of this method follow the schema illustrated in Fig. 3.

According to Kitchenham, the formulation of research questions is the first and most important activity of the protocol definition. We have defined such questions considering four different perspectives. **Demographical** questions focus on metadata aspects of the paper. These questions are:

- DemRQ1: When was the paper published (year)?
- DemRQ2: Where was the paper published (journal or conference name and impact factor)?

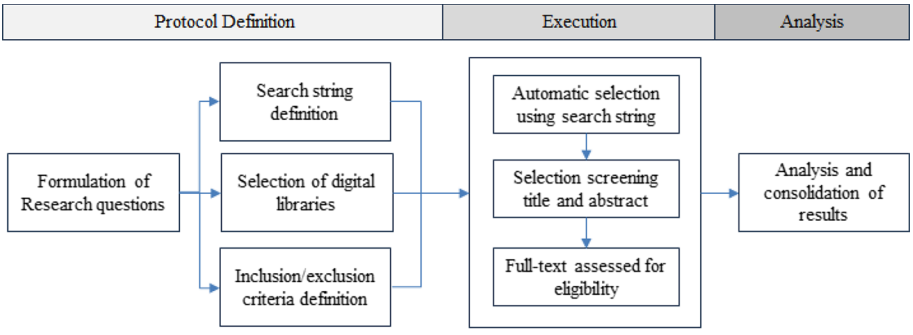

Fig. 3 Systematic review schema applied

- DemRQ3: What is the main research group/university?
- DemRQ4: Which is the objective and scientific contribution?

Input questions focus on the format and organization of the input data. These questions are:

- InpRQ1: Which is the input vocabulary and how was it created?
- InpRQ2: Which is the data unit and frequency (periodic or aperiodic)?
- InpRQ3: Are static attributes also employed?
- InpRQ4: Which is the strategy used to implement the positional encoding?
- InpRQ5: How are these attributes embedded?

Architectural questions characterize the organization of the network layers and possible adaptations regarding the traditional transformer architecture (Fig. 1):

- ArcRQ1: Which is the architectural type (Encoder-only, Decoder-only, Encoder-Decoder)?
- ArcRQ2: How are the Nx modules (see Fig. 1) organized?
- ArcRQ3: How does the proposed architectural design support the learning objective (task)?

Evaluation questions focus on the training process and evaluation aspects of the trained model:

- EvaRQ1: What are the characteristics of the dataset used for evaluation (e.g., size)?
- EvaRQ2: What is the pretraining strategy?
- EvaRQ3: What is the fine-tuning strategy?
- EvaRQ4: Which are the evaluation criteria?
- EvaRQ5: Are there comparisons to other approaches?

Explainability question tries to identify if the approach presents ways to explain its results.

- ExpRQ1: Is the interpretability/explainability of the model discussed?

The next step is the search string definition. We relied on general terms associated with the research questions to compose such a string (Kitchenham 2004). This strategy avoids bias and low covering. After a stage of string calibration, where we tried several terms' combinations, the resultant string was: (transformer) AND ("deep learning" OR "neural network" OR "machine learning") AND (health OR medical OR patient). We defined four datasets to use this string (ScienceDirect, IEEE Xplore Digital Library, ACM Digital Library, and PubMed), and three stages for paper selection, which are:

- Stage 1: Automatic match process using the search string. Only the title and abstract were considered as the search space for this match. Moreover, we only considered full papers written in English and dated from 2018 to 2023.
- Stage 2: Manual title and abstract screening of the papers that resulted from the previous stage. The aim was to identify papers that are related to our research area and

have the potential to answer our research questions. This stage only considered primary studies and dropped out duplicate papers.
- Stage 3: Remaining papers from the previous stage were fully analyzed to extract the data that could answer the research questions. Papers that did not contain enough information were dropped out.

The temporal search range was defined from 2018 to 2023 since the studies about transformers were initiated after the seminal paper of Vaswani et al. (2017), released in December 2017. Two reviews separately conducted the paper selection over stages 1 and 2 to avoid bias. A third reviewer participated in cases of disagreements, looking for a consensus.

4 Results

This section first discusses the selection process conducted in this review. Then, we consolidate the results obtained in the four sets of research questions (demographical, input, architectural, evaluation, and explainability), emphasizing their main remarks.

4.1 Selection process

The selection of papers based on search string (stage 1) returned 456 papers. After analyzing their title and abstract, 59 papers were chosen for a more detailed analysis (stage 2). We selected 21 papers (stage 3) from this set to compose our study. Most of the 38 papers discarded during stage 3 are related to NLP (14 papers) and image-based (6 papers) applications in health. However, their approaches do not process longitudinal multifeature data. The next paragraphs discuss some of the discarded papers, emphasizing the rationale of our decisions and better characterizing the scope of this review.

Papers related to NLP discuss mechanisms to process textual health documents (e.g., EHR notes). As we could expect, since transformers come from the NLP area, the initial use of this technology in health was focused on tasks such as identifying similar sentences (Mahajan et al. 2020), named entity recognition (Falissard et al. 2022), summarization (Chen et al. 2020), and classification based on text analysis. For example, the work in Mahajan et al. (2020) relied on ClinicalBERT (Huang et al.2019), a pre-trained domain-specific transformer-based language model for identifying semantic textual similarity in the clinical domain and redundant information in clinical notes. The work in Falissard et al. (2022) focused on automatically recognizing ICD-10 medical entities from the French natural language using transformers. The work in Chen et al. (2020) relied on a BERT-based structure to build a diagnoses-extractive summarization model for hospital information systems. Regarding classification based on text analysis, the work in Shibly et al. (2020) created a transfer learning system by fine-tuning the BERT language model. The aim was to identify the right doctor, in terms of clinical expertise, who should receive the preliminary textual diagnosis description.

Papers that present image-based approaches investigate the use of transformers for medical image analysis. For example, the work of Liu et al. (2021a) investigates the use of transformer layers to optimize the extracted features of breast cancer tumor images. The work of Fu et al. (2022) uses a transformer-encoded Generative Adversarial Network (transGAN) to reconstruct low-dose PET (L-PET) images into high-quality full-dose PET

(F-PET). W-Transformer (Yao et al. 2022) and xViTCOS (Mondal et al. 2021) are other examples of transformer-based works focused on health image analysis. While the former integrates convolutional and transformers layers to detect spinal deformities in X-ray images, the latter uses a transfer learning technique (pre-trained transformer model) for COVID-19 screening tests that rely on chest radiography.

Apart from papers of these two groups, other papers that use transformers, for example, to analyze types of sounds (e.g., cough sounds for disease detection (Yan et al. 2022)) and define new health corpora to train transformers (Wang et al. 2019), were also discarded since they do not consider multifeature longitudinal health data. This means simultaneous data streams (e.g., physiological, physical, psychological, social, or contextual) that are assessed over long periods in daily life.

4.2 Demographical research questions

The 21 remaining papers are temporally distributed according to their year of publication (**DemRQ1**) as follows: three papers in 2020, fourteen papers in 2021, three papers in 2022, and one in 2023. This temporal distribution emphasizes that no paper was identified in the initial two years of our search (2018 and 2019). This fact suggests that the transformers technology had a maturation time until research groups identified its potential for analyzing other than NLP longitudinal health data.

The quality of the publication vehicles (**DemRQ2**) is a factor that indicates the importance of the technology. Our analysis identified that most of these 21 papers were published in standing journals with high-impact factors, such as the *IEEE Transactions on Neural Networks and Learning Systems* (IF = 14.25) (Rao et al. 2022a, b), *IEEE Journal of Biomedical and Health Informatics* (IF = 7.021) (Rao et al. 2022a, b; Meng et al. 2021; Darabi et al. 2020; Li et al. 2023a, b), *Scientific Reports* (IF = 4.996) (Li et al. 2020; Zeng et al. 2022), *Nature Digital Medicine* (IF = 11.653) (Rasmy et al. 2021), and *Journal of Biomedical Informatics* (IF = 8.000) (Li et al. 2021). Similarly, papers were also published in important conferences such as the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (IF = 7.02) (Ren et al. 2021), *AAAI Conference on Artificial Intelligence* (IF = 3.4) (Prakash et al. 2021), *International Conference on Machine Learning and Applications* (IF = 3.844) (Fouladvand et al. 2021; Dong et al. 2021), and *International Symposium on Computer-Based Medical Systems* (IF = 1.040) (Florez et al. 2021).

The identification of several papers from the same research group (**DemRQ3**) shows the ongoing efforts and technological developments rather than a paper resulting from a one-off, isolated study. We found only one case in our analysis. The University of Oxford released the BEHRT model in 2020 (Li et al. 2020) and advanced this model in two works published in 2022 (Rao et al. 2022a, b) and another in 2023 (Li et al. 2023a, b). The Graduate Institute of Biomedical Electronics and Bioinformatics at the National Taiwan University used transformer models for two disease pattern retrieval and classification in 2021 (Boursalie et al. 2021) and prediction of postoperative mortality in 2022 (Chen et al. 2022). However, our analysis did not consider this latter application since it does not use longitudinal data. **DemRQ3** was also useful to attest to the strong interdisciplinarity of this type of research. Many papers involve authors that are from some health-related department/faculty such as Medicine (Li et al 2020; Li et al 2021; Rasmy et al. 2021), Women's and Reproductive Health (Rao et al. 2022a, b), Radiologic Sciences (Meng et al. 2021), and Biomedical Informatics (Boursalie et al. 2021). The most interesting example is the study in Fouladvand et al. (2021), with authors from the Biomedical Informatics, Computer

Science, Internal Medicine, Pharmaceutical Sciences, Biostatistics, Psychiatry, Family and Community Medicine departments. In this case, the involvement of interdisciplinary teams is related to the research complexity that involves aspects of opioid use disorder.

Finally, Table 1 summarizes our findings regarding the learning objective (task) and application domain (**DemRQ4**). This table shows that prediction is the most common learning task, while the scientific contributions are mainly focused on:

- Designing domain-specific vocabularies (e.g., diagnosis codes for EHR) and their embeddings using simple operations such as sum or concatenation (Li et al 2020; Rao et al. 2022a; Florez et al. 2021; Meng et al. 2021; Rasmy et al. 2021).
- Including temporal semantic elements, such as temporal distance between inputs (Boursalie et al. 2021; An et al. 2022), artificial time tokens as part of the vocabulary, (Pang et al. 2021), time reference for an event from defined anchor date (Prakash et al. 2021), and special representations for duration of visits and interval between visits (Peng et al. 2021).
- Conducting a more complex input data preprocessing, using, for example, LSTM (Fouladvand et al. 2021) or a pre-transformer (Chen et al. 2021b) module, including spatial dependences (Li et al. 2021), creating hierarchical input structures (Ye et al. 2020; Li et al. 2023a, b), filtering relevant information (Shome 2021), using self-supervised learning to augment the input value (Dong et al. 2021, and joining categorical and textual information (Darabi et al. 2020).
- Modifying the learning algorithm to simultaneously analyze different data streams (Fouladvand et al. 2021), employing an attention mechanism based on probabilities (Li et al. 2021), or proposing new forms of pre-training (Zeng et al. 2022; Ren et al. 2021).

These scientific contributions are better discussed along this paper.

4.3 Transformers input research questions

Table 2 summarizes the answers to the input research questions. The second column (Longitudinal inputs) indicates the vocabulary (lexicon) used in each approach (**InpRQ1**). Many papers (Li et al. 2020; Rao et al. 2022a, Florez et al. 2021, Pang et al. 2021, Prakash et al. 2021) use standardized categorical codes of diagnosis (e.g., ICD), medications, and other health elements as part of their vocabulary. Some papers mix categorical and continuous data (Li et al. 2023a, b; Rao et al. 2022b). In this case, they apply a categorization process to transform the continuous data into tokens of a vocabulary. Works that assess data from sensors (Shome 2021; Dong et al. 2021; An et al. 2022] do not use the concept of vocabulary since their training strategies directly use the raw data as input. A third type of input representation strategy uses special modules to learn their input (FCNN (Chen et al. 2021b), DTW (Li et al. 2021)).

The third column represents the longitudinal unit, which aggregates the data assessed at each timestep (**InpRQ2**). Many studies use the idea of visits as longitudinal unit, and they are aperiodic (Li et al. 2020; Darabi et al. 2020). For aperiodic units, for example, the work in Boursalie et al. (2021) proposes concatenating the time between assessments (elapsed time) in the unit encode. Proposals based on sensors (Shome 2021; Dong et al. 2021) have each data capture cycle as their longitudinal unit, which is mostly periodic.

Table 1 Learning tasks and main contribution

References	Learning objective	Scientific contribution
Li et al. (2020)	Prediction (Next diagnosis)	Adaptation of BERT (Devlin et al. 2018) to cover EHR elements (diagnosis codes)
Rao et al. (2022a)	Prediction (Heart failure)	Extends the previous model (Li et al. 2020) so it considers further EHR elements
Florez et al. (2021)	Prediction (Prognostic)	Transformer architecture (without pretraining) to cover EHR elements (e.g., diagnosis codes)
Meng et al. (2021)	Prediction (Depression)	Extends the input code embeddings for bidirectional sequential learning
Fouladvand et al. (2021)	Binary classification (Opioid disorder)	Data preprocessing using LSTM models. Simultaneous analysis of multiple types of healthcare data streams
Boursalie et al. (2021)	Prediction (Next diagnosis)	Inclusion of elapsed time between longitudinal units (e.g., medical visits) embeddings
Chen et al. (2021b)	Prediction (Critical care)	Use of a first transformer model for unsupervised embedding of the disease concept into a second transformer
Li et al. (2021)	Prediction (Influenza propagation)	Inclusion of spatial dependency, using curve similarity, as part of the embedding. Adaptation of attention mechanism for a probabilistic mode
Ye et al. (2020)	Prediction (Disease risk)	Additional modules to model the hierarchical structure of EHR data and extract long- (transformer encoder) and short-term dependences (convolutional layer)
Rasmy et al. (2021)	Prediction (Disease risk)	Adaptation of BERT (Devlin et al. 2018) to cover EHR elements (e.g., diagnosis codes)
Zeng et al. (2022)	Prediction (Next diagnosis)	Proposal of unsupervised pretraining that relies on the Next Visit Prediction and Categorical Prediction approaches
Shome (2021)	Classification (Physical activity)	Use of a depthwise separable residual feature extractor network for input preprocessing
Dong et al. (2021)	Prediction (Mental disease)	Graphs as input. Integrates a contrastive self-supervised learning method such that the bag classifier can be learned using semi-supervision
Darabi et al. (2020)	Prediction (Medical events)	Use of two input distributed representation for a patient based on text and medical codes
Li et al. (2023a, b)	Prediction (Risk of health issues)	Use of hierarchical structures to allow extracting associations from longer sequences
Pang et al. (2021)	Prediction (Clinical events)	Incorporating temporal information using artificial time tokens as part of the input vocabulary

Table 1 (continued)

References	Learning objective	Scientific contribution
Ren et al. (2021)	Classification and regression (diabetes and hypertension)	Proposal of new forms of pre-training to handle data insufficiency, data incompleteness and short sequence problems using a time-aware transformer architecture
Prakash et al. (2021)	Classification (Rare disease)	Extension of Med-BERT by including context embedding and temporal reference embedding, together with a novel adaptive loss function to handle the class imbalance
Rao et al. (2022b)	Prediction (Treatment risks)	Use of an unsupervised learning for richer feature extraction, which capture both static and temporal medical history variables
An et al. (2022)	Prediction (Mortality)	Learn the personalized irregular temporal patterns of medical events
Peng et al. (2021)	Prediction (Diagnosis)	Use of neural ordinary differential equation to handle both irregular intervals between a patient's visits with admitted timestamps and length of stay in each visit

Table 2 Summary of input research questions

References	Longitudinal inputs (lexicon = Γ)	Longitudinal unit (t) and frequency	Static attributes	Positional encoding	Input embedding
Li et al. (2020)	Diagnosis code (*D*)	Patient visits *Vp* = list of *D*. Aperiodic	–	$f(t)$ = sinusoidal function. t is the *Vp* sequential value	$E_{Vp} \leftarrow \Gamma_{Vp} \oplus Age_{Vp} \oplus f(t) \oplus$ A/B segment
Rao et al. (2022a)	Diagnosis (*D*) and medication (*M*) codes	Patient visits *Vp* = list of *D* ∪ list of *M*. Aperiodic	–	–	$E_{Vp} \leftarrow \Gamma_{Vp} \oplus$ Age-in-months$_{Vp} \oplus$ Year$_{Vp}$
Florez et al. (2021)	Diagnosis code (*D*)	Patient admission *Ap* = list of *D*. List of fix size |*D*|-dimension. Uses padding to complete. Aperiodic	–	$f(t)$ = sinusoidal function, t is the *Vp* sequential value	$E_{Ap} \leftarrow$ hot_vector$(\Gamma_{Ap}) \oplus f(t)$
Meng et al. (2021)	Diagnosis (*D*), [d]CPT (*C*), medication (*M*), and [a]Topic features (*T*) codes	Patient visits *Vp* = list of *D* ∪ list of *C* ∪ list of *M* ∪ list of *T*. Aperiodic	Gender	$f(t)$ = sinusoidal function, t is the *Vp* sequential value	$E_{Vp} \leftarrow \Gamma_{Vp} \oplus f(t) \oplus$ A/B segment $\oplus$ Age$_{Vp} \oplus$ Gender
Fouladvand et al. (2021)	Diagnosis (*D*) and medication (*M*) codes	Patient visits *Vp* = list of *D* ∪ list of *M*. Periodic (monthly)	[b]Age, Gender	$f(t)$ = month $\in$ [month$_{min}$, month$_{max}$]. t is the *Vp* sequential value	$E_D \leftarrow$ LSTM(D) $\oplus f(t)$ $E_M \leftarrow$ LSTM(M) $\oplus f(t)$ $E_{Vp} \leftarrow E_D \copyright E_M$
Boursalie et al. (2021)	Diagnosis (*D*) and medication (*M*) codes	Patient visits *Vp* = list of *D* ∪ list of *M* ∪ elapsed time. Aperiodic	[c]Age, Gender	$f(t) = t$, where t is the *Vp* sequential value	$E\Delta \leftarrow \Delta t(Vp^t, Vp^{t-1})$ $E_{Vp} \leftarrow (D \copyright M \copyright E\Delta)_{Vp} \oplus f(t)$
Chen et al. (2021b)	FCNN-1 ([e]15 attributes) and FCNN-2 ([d]NL features)	Patient visits *Vp* = FCNN1 ∪ FCNN2. Aperiodic	–	$f(t)$ = sinusoidal function. t is the *Vp* sequential value	$E_{Vp} \leftarrow \Gamma_{Vp} \oplus f(t) \oplus$ A/B segment
Li et al. (2021)	[d]DTW (number of patient)	Patient counts *Cp* = DTW (patient counts). Periodic (weekly)	–	$f(t)$ = sinusoidal function. t is the *Cp* sequential value	$E_{Cp} \leftarrow \Gamma_{Cp} \oplus f(t)$
Ye et al. (2020)	Diagnosis code (*D*)	Patient visits *Vp* = HAM(list of *D*). Aperiodic	–	f_{long} = sinusoidal function, f_{short} = conv1d layer	$E_{Vp} \leftarrow \Gamma_{Vp} \oplus f_{long}(Vp) \oplus f_{short}(Vp)$
Rasmy et al. (2021)	Diagnosis code (*D*)	Patient visits *Vp* = list of *D*. Aperiodic	–	$f(t) = t$, where t is the *Vp* sequential value g = codes priority order (1..n)	$E_{Vp} \leftarrow \Gamma_{Vp} \oplus f(t) \oplus g(\Gamma_{Vp})$

Table 2 (continued)

References	Longitudinal inputs (lexicon = Γ)	Longitudinal unit (t) and frequency	Static attributes	Positional encoding	Input embedding
Zeng et al. (2022)	Diagnosis (*D*), procedure (P), medication (*M*), and visit type (VT) codes	Patient visits *Vp* = list of *D* ∪ list of P list of *M* ∪ VT Aperiodic	[c]Age, Gender	*f*(*t*) = visit date	$E_{Vp} \leftarrow$ Maxpool(Γ_{Vp} © *f*(*t*)) © Age © Gender
Shome (2021)	No lexicon. Raw data from accelerometer and gyroscope sensors	Sensor readings *S*. Periodic (continuous)	–	*f* (*t*) = continuous timestep	$E_t \leftarrow$ Set of Residual blocks (*S*,*f*(*t*))
Dong et al. (2021)	No lexicon. [e]Daily life data using sensors	Sensor readings *S*. Periodic (daily)	–	*f* (*t*) = visit date	$E_t \leftarrow$ GCL(Graph(S © *f*(*t*)))
Darabi et al. (2020)	Diagnosis (*D*), procedure (P), medication (*M*) codes, and medical notes	Patient visits *Vp* = list of *D* ∪ list of P list of *M* ∪ VT ∪ medical notes. Aperiodic	[c]Demographics (age, gender, race, etc.)	*f*(*t*) = sinusoidal function. *t* is the *Vp* sequential value	$E_{Vp} \leftarrow E_{codes}$ © E_{notes} © *Demographics* $E_{codes} \leftarrow$ Encoder($\Gamma_{Vp} \oplus f(t)$) $E_{notes} \leftarrow$ Summarizer(NL$_{Vp}$) $\oplus f(t)$
Li et al. (2023a, b)	Diagnosis (D), Medication (M), Procedure (P), Examinations (E), Blood pressure (BP), Drinking status (DS), Smoking status (SS), and body mass index (BMI)	Patient visits *Vp* = list of patient records (D ∪ M ∪ P ∪ E ∪ BP ∪ DS ∪ SS ∪ BMI). Aperiodic	–	*f*(*t*) = sinusoidal function. *t* is the *Vp* sequential value	$T_{sw\text{-}i} \leftarrow$ Transform ($\Gamma_{Vp} \oplus Age \oplus$ A/B $\oplus f(t))_{\lvert sw \rvert}$, where |sw| = slide windows length $E_{Vp} \leftarrow T_{sl\text{-}1}$ © … © $T_{sl\text{-}n}$
Pang et al. (2021)	Diagnosis (D), Medication (M), and Procedures (P)	Patient visit *Vp* = list of D ∪ list of M ∪ list of P. Aperiodic	–	Redefined. Implementation of Time (T_{emb}) and Age (A_{emb}) embeddings	$E_{Vp} \leftarrow \Gamma_{Vp} \oplus T_{emb} \oplus A_{emb}$
Ren et al. (2021)	Examinations (E)	Patient visit *Vp* = list of E. Aperiodic	–	*f* (week) = sinusoidal function, week index as input	$E_{Vp} \leftarrow \Gamma_{Vp} \oplus f$ (week)
Prakash et al. (2021)	Diagnosis (D), Treatment (T), and Procedures (P)	Patient visit *Vp* = list of D ∪ list of M ∪ list of P. Aperiodic	–	*f* (*t*) = *Vp* time reference regarding an anchor date	$E_{Vp} \leftarrow \Gamma_{Vp} \oplus$ Type(Γ_{Vp}) index (*Vp*) $\oplus f(t)$

Table 2 (continued)

References	Longitudinal inputs (lexicon = Γ)	Longitudinal unit (t) and frequency	Static attributes	Positional encoding	Input embedding
Rao et al. (2022b)	Diagnosis (D), Medication (M), and Blood pressure (BP)	Patient visit Vp = list of D ∪ list of M ∪ BP. Aperiodic	Gender, region, smoking	$f(t) = t$, where t is the Vp sequential value	$E_{Vp} \leftarrow \Gamma_{Vp} \oplus Age_{Vp} \oplus Year_{Vp} \oplus f(t)$ © (gender © region © smoking)
An et al. (2022)	No lexicon. Raw data of vital signals and examinations. Medications are later included	Sensor readings S = vital signal (VS) ∪ examinations (E). Aperiodic	–	f_{vs} = interval between vital signals, f_{exa} = interval between examinations	$E_{vs} \leftarrow$ vital signal $\oplus f_{vs}$ $E_E \leftarrow$ vital signal $\oplus f_{exa}$
Peng et al. (2021)	Diagnosis (D)	Patient visit Vp = list of D. Aperiodic	–	Redefined using $f1 = {}^{f}\text{ODE}(Vp\text{-}start, Vp\text{-}end), ff2 = {}^{f}\text{ODE}(Vp_1, Vp_2)$	MO ← Medical ontology $E_{oe} \leftarrow$ OntologyEncoder (MO, Γ_{Vp}) $E_{Vp} \leftarrow E_{oe} \oplus f1 \oplus f2$

[a]Examples of topic features are words such as "pain", "fracture", or any part of the body

[b]These static features are only used in the final layers, after the transformer module

[c]Attributes only used in the first visit (t = 1)

[d]CPT = Current Procedural Terminology; © = concatenation operation; ⊕ = summing operation; NL = natural language; FCNN = Fully connected neural network; DTW = Dynamic Time Warping algorithm, HAM = Hierarchical Attention Module; GCL = Graph Convolution Layer

[e]Systolic and diastolic blood pressure, heart rate, SpO2, respiratory rate, body temperature, height, weight, pain index, Glasgow coma slace, eye response, verbal response, motor, and motor response

[f]ODE = Ordinary differential equations

The fourth column indicates whether the approaches consider static attributes (**InpRQ3**). The most common are age and gender, used at the beginning of each patient sequence (Boursalie et al. 2021; Rao et al. 2022b) or directly in the last layer of the architecture (Fouladvand et al. 2021). However, some works use the attribute age as a resource to improve the sequential semantic notion (Li et al. 2020; Rao et al. 2022a). Thus, they support the positional encoding task. Our review shows that diverse concepts and union of concepts are used as positional encode (**InpRQ4**). For example, the sinusoidal function (Florez et al. 2021; Li et al. 2021; Darabi et al. 2020), which is used in the original transformer paper (Devlin et al. 2018), simple sequential values (Rasmy et al. 2021; Li et al. 2023a, b; Rao et al. 2022b), and calendar dates (Zeng et al. 2022; Dong et al. 2021). A/B segments are encodings used as an additional semantic layer to distinguish two adjacent longitudinal units (Li et al. 2020; Meng et al. 2021; Chen et al. 2021b). For example, information related to each patient visit alternately receives the segments A and B. However, the advantages of its use to improve the representation of positional encodings are unclear in the literature. There are also slight variations. For example, the work in Ren et al. (2021) uses the number of weeks as the input parameter of the sinusoidal function rather than simple sequential values. On the other hand, some papers completely redefine the idea of positional encoding. The work in Pang et al. (2021) defines two embeddings, one for representing the idea of continuous time using age as basis (A_{emb}) and another for cyclic time using the calendar data as basis (T_{emb}). The work in Peng et al. (2021) uses two ordinary differential equations (ODE) to represent visit durations given their initial and final time and the interval between such visits. According to Peng et al. (2021), ODEs are particularly interesting in handling arbitrary time gaps between observations.

The sixth column indicates how the input information is embedded (**InpRQ5**). As discussed in the previous section, the contribution of most papers focused on these embedding functions. For example, the work in Fouladvand et al. (2021) uses the sum of two LSTM network outputs, which have the diagnosis and medication longitudinal data as input. These inputs are also summed to the months when they were assessed. The work in Boursalie et al. (2021) first calculates the elapsed time between two visits (longitudinal unit) and concatenates this value to the diagnosis and medication codes. The final embedding is summed to the positional encoding. The approach in Shome (2021) uses the results of a set of *depthwise separable residual feature extractor* networks. This set receives the concatenation of the assessed sensorial data and the timestep when such data were assessed. The work presented in Darabi et al. (2020) considers two different input modalities. Firstly, the encoding of the sum of categorial inputs and positional encoding. Secondly, the summarization of medical notes, also with its positional encoding. The final embedding is given by concatenating these two information and demographics data. Hierarchies of transformers (Pang et al. 2021) are also used to create clusters of sequential data according to a sliding window. A pre-transformer handles each of these clusters, and all the results are concatenated and used as the input of the main architectural transformer. Another interesting approach is proposed in Peng et al. (2021), which uses medical ontologies to augment the sequential information of patients. Thus, this approach uses an ontology encoder to map the ontology information to a vector.

4.4 Architectural research questions

Table 3 shows that encode-only is the most common architectural type (**ArcRQ1**) found in our analysis.

Table 3 Architectural type of reviewed papers

Transformer		
Encode-only	Decode-only	Enc/Dec
Li et al. (2020), Rao et al. (2022a), Meng et al. (2021), Fouladvand et al. (2021), Chen et al. (2021b), Ye et al. (2020), Rasmy et al. (2021), Zeng et al. (2022), Shome (2021), Dong et al. (2021), Darabi et al. (2020), Li et al. (2023a, b), Ren et al. (2021), Rao et al. (2022b), Peng et al. (2021)	Florez et al. (2021), Boursalie et al. (2021), Prakash et al. (2021)	Li et al. (2021), Pang et al. (2021), An et al. (2022)

This result may be expected since the best-known transformer example (BERT) also follows the encode-only approach. Thus, the authors were motivated by the results obtained for BERT. However, it is interesting to analyze why other works employed different architectural types rather than the encode-only approach. The proposals in Florez et al. (2021), Prakash et al. (2021), and Boursalie et al. (2021) construct models that do not need pre-training. Moreover, they are interested in producing step-by-step iterative outcomes, where the previous output is used as input for the current process. This feature requires autoregressive architectures, such as the decode-only discussed in Sect. 2.1. The work in Li et al. (2021) employs an enc/dec architecture since its problem requires a pre-analysis of long time-series sequences at once (encode function) and the use of this analysis to support multi-step-ahead time-series predictions in a generative style (decoder function). The proposal in Pang et al. (2021) uses a simple decoder to implement a second learning objective (Visit type Prediction—VTP) to boost further the performance of the encoder learning, which relies on the traditional masked language modelling (MLM) as its learning strategy. In An et al. (2022), the authors use end/dec architectures to derive an aware contextual feature representation of inputs. The result is a list of temporal features generated one by one according to the autoregressive decoder style.

The next step is to understand how the Nx module of each approach differs (**ArcRQ2**) from the original design illustrated in Fig. 1. Part of the works (Li et al. 2020; Rao et al. 2022a; Meng et al. 2021; Rasmy et al. 2021; Zeng et al. 2022; Shome 2021; Darabi et al. 2020; Li et al. 2023a, b; Rao et al. 2022b; Peng et al. 2021) basically follow the encoder stage of the original transformer architecture (Fig. 1), including a final fully connected layer for classification/prediction. This layer uses activation functions such as Sigmoid (Li et al. 2020; Rao et al. 2022a) and Softmax (Florez et al. 2021; Zeng et al. 2022). Other works follow (Florez et al. 2021; Boursalie et al. 2021; Prakash et al. 2021) the decoder stage of this architecture, while the proposal in An et al. (2022) relies on both stages. While these proposals did not present significant changes in the Nx module, the remaining works conducted diverse modifications (Fig. 4). The work in Dong et al. (2021) includes a layer called *Prediction Bag Label*, which is a specific component to the problem of predicting classes for groups of graphs. The work in Chen et al. (2021b) uses two transformers since the first transformer generates an embedding of diseases, which is used as input of the second transformer to augment its classification accuracy. The work in Fouladvand et al. (2021) proposed three main modifications. Firstly, it does not use residual connections. Secondly, the multi-head attention block is modified to generate attention weights between different input streams. As these streams were mixed, the architecture also includes reconstruction layers to redefine the original streams. The work in Li et al. (2021) uses the encoder-decoder transformer architecture.

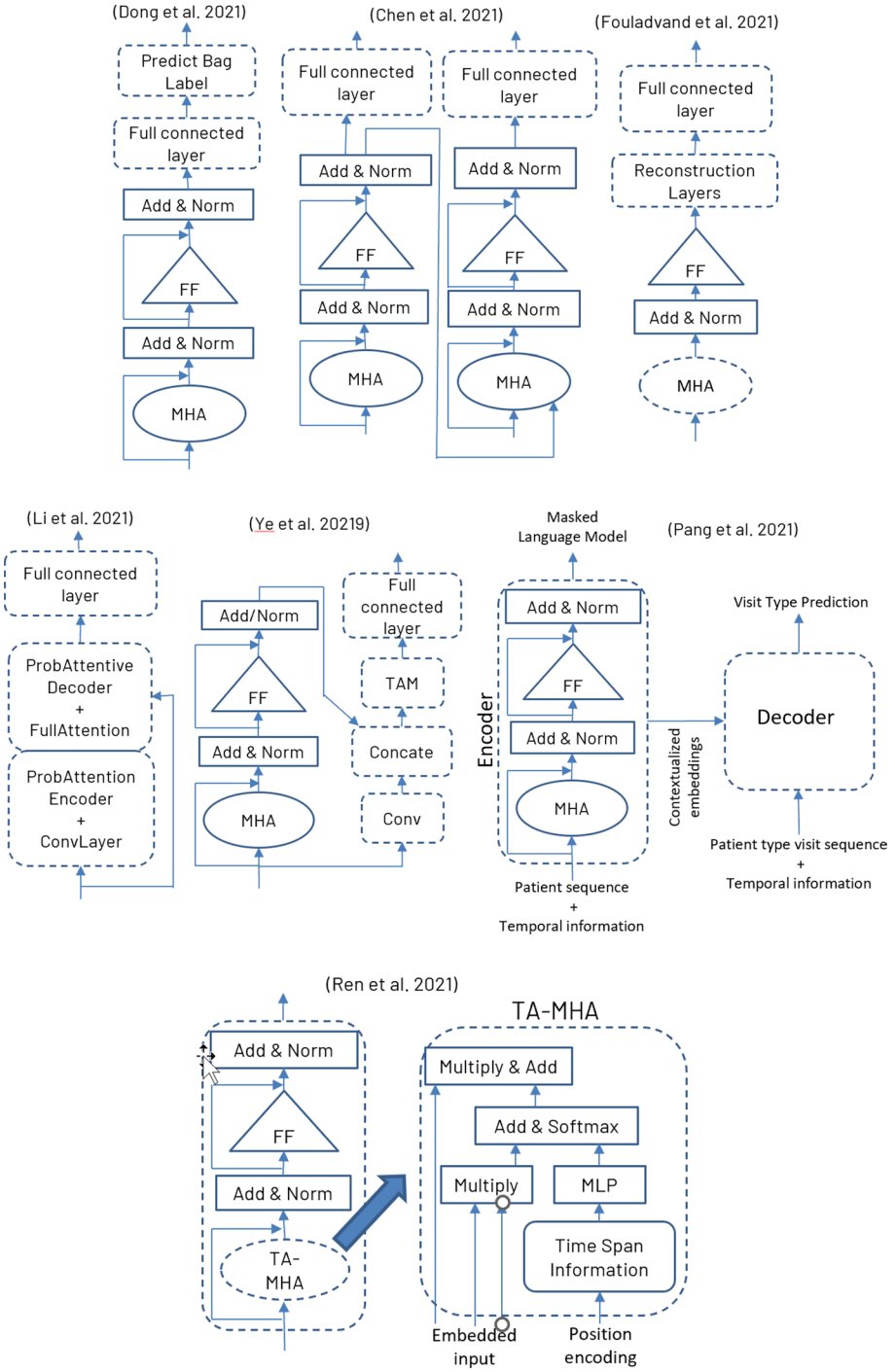

Fig. 4 Schema of the proposed modifications in the Nx module

However, the encoder is modified to use a probabilistic attention mechanism with a convolutional neural network, while the decoder uses this same mechanism with a full attention layer. The work in Ye et al. (2020) concatenates the results of a traditional transformer with the results of a 1-dimensional convolutional layer (conv). The hierarchical attention mechanism (TAM) layer receives this result, generating a dense diagnostic embedding for each longitudinal unit. The work in Pang et al. (2021) uses the contextualized embedding generated by the encoder to support a new learning process conducted by the decoder module, called *visit type prediction*. Finally, the work in Ren et al. (2021) proposes a new multi-head attention module to handle irregular time intervals. This architecture uses a fully connected layer (multilayer perceptron—MLP) to capture the time span information, which relies on position encoding outputs. The other layers account for integrating such temporal information into the health data.

Many works that follow the traditional architecture (Li et al. 2020; Rao et al. 2022a; Florez et al. 2021; Meng et al. 2021; Rasmy et al. 2021; Zeng et al. 2022; Shome 2021; Darabi et al. 2020; Prakash et al. 2021) include a classification/prediction layer as the way of conducting, for example, the fine-tuning process in their specific domains. Meanwhile, the other 14 works presented more complex modifications. Thus, we tried to identify the motivations for their design decisions (**ArcRQ3**). The work in Fouladvand et al. (2021) combines each stream of each type of health data (e.g., diagnosis and medication codes) inside the transformer module, rather than previously concatenating or summing such streams and dealing with them as a unique input stream. According to the authors, this approach facilitates exploring associations within and between these patients' data streams. The work in Chen et al. (2021b) considers the relation between diseases as fundamental to predict care tasks. Thus, the use of its first transformer is justified to extract features of this relation that enhance the predictions of the second transformer. Li et al. (2021) argue that their modifications avoid the quadratic computation of self-attention, its memory bottleneck in stacking layers for long inputs, and the speed plunge in predicting long outputs. The work in Ye et al. (2020) intends to better learn the long and short dependencies among longitudinal data. Thus, it uses two components, which are a traditional transformer and a 1-dimensional convolutional layer. The work in Dong et al. (2021) relies on graphs as input. Thus, the specialization of its final prediction layer is a natural design step for the learning process.

The hierarchical approach in Li et al. (2023a, b) relies on the classical divide-and-conquer metaphor to analyze long sequences. However, transformers do not theoretically have this limitation regarding their sequence size. Thus, the main reason is the hypothesis that hierarchical information clusters improve the learning process and, thus, the accuracy of the model. As the authors emphasize, "medical records naturally have stronger correlation when they are closer in time" (Li et al. 2023a, b). In Pang et al. (2021), the visit type prediction learning strategy takes advantage of the available semantics for augmenting the information gain of the transformer. Indeed, the type of visit is a type of information that is not used in other approaches. At last, the support for temporal notions has motivated some approaches to conducting their modifications, which affect the task processing in the following way:

- The values of intervals between visits are included as tokens. However, they are not part of the vocabulary, and the decoder layers are thus modified to analyze the event-pairs (interval time, visit information) rather than only each token (Boursalie et al. 2021).

- The standard equations of the transformer self-attention are modified to consider the irregular time interval between visits. This information is separately calculated and then integrated into the visit information (Ren et al. 2021).
- Traditional transformers are specifically used to create temporal representations employing time-aware attention weights. An interesting aspect is the use of a transformer for each data type (lab tests and vital signs). A later hierarchical feature fusion model is then used to integrate such temporal data with the prescription patient information (An et al. 2022).
- The approach in Peng et al. (2021) also uses the transformer to learn temporal notions. In this case, the admission intervals and length of sequential visits of a patient. Like the previous case, the transformer is not modified. Rather, the proposal incorporates the information regarding admission intervals and visit lengths to the transformer input to support its learning objective.

Based on this analysis, the transformer literature does not present a concrete strategy to represent this temporal notion. This is one of the main issues of this type of architecture, as detailed later in this paper.

4.5 Evaluation questions

Table 4 summarizes the results regarding the evaluation questions. The data characterization (**EvaRQ1**) column shows that most of the datasets present a high number of samples. This high number is essential for approaches that require a pretraining stage. We also included the data dimension (*dim*) used for each approach between square brackets. For example, the work of Li et al. (2020) has only one dimension represented by diagnosis codes. Approaches that do not require this stage and rely on mobile data (Shome 2021; Dong et al. 2021), for example, present fewer samples. The third column indicates which proposals use pretraining strategies and identifies these strategies (**EvaRQ2**). Masked Language Model (MLM) is the main strategy used for pretraining. However, we also found some variations. For example, the work in Li et al. (2023a, b) implements the *Bootstrap your own latent* (BYOL) strategy with MLM (Li et al. 2023a, b). BYOL trains two networks (online and target), which are augmented separately. The idea is to minimize the mean squared loss between the output of the online predictor and the target projector. In Pang et al. (2021), the *visit time prediction* (VTP) is concurrently conducted with MLM, using a different semantical content that can provide gains to the learning process.

The fourth column characterizes the fine-tuning strategies (**EvaRQ3**) when they are employed. These strategies use traditional layers for prediction or classification, which use sigmoid or softmax as the activation function. A particular strategy (Positive unlabeled (PU)-learning) is proposed in Prakash et al. (2021) aimed at handling the class imbalance. According to this strategy, the training data comprises only positive and unlabeled instances, whereas unlabeled examples include both positive and negative classes. When fine-tuning is not used, the proposals usually employ the traditional backward propagation as the learning mechanism (Florez et al. 2021; Fouladvand et al. 2021; Li et al. 2021; Shome 2021; Dong et al. 2021; An et al. 2022; Peng et al. 2021). The unique exception is the work in Boursalie et al. (2021), which uses predictions for the random subset of elements masked as the final training (Boursalie et al. 2021).

The evaluation criteria (**EvaRQ4**) of the proposals are indicated in the fifth column, which emphasizes the preference of such proposals for the Area Under the Curve-Receiver

Table 4 Summary of the evaluation research questions

Refs	Dataset characterization, Dimensionality	Pretraining strategy	Fine-tuning strategy	[b]Evaluation Criteria	[c,d]Comparison to other approaches
Li et al. (2020)	Clinical Practice Research Datalink (CPRD), EHR, 1.6 million patients [dim = 1]	Masked language model (MLM)	Disease prediction in the next visit, next 6 months visit and next 12 months visit	Average precision score (APS), AUC-ROC	–
Rao et al. (2022a)	Clinical Practice Research Datalink (CPRD), EHR, 100 thousand patients [dim = 2]	Masked language model (MLM)	Prediction of heart failure using specific dataset in this domain	AUC-ROC AUPRC	RETAINEX
Florez et al. (2021)	Medical Information Mart for Intensive Care III (MIMIC-III), 9,537 patients, and InCor from Heart Institute, University of Sao Paulo, 89,000 patients [dim = 1]	–	–	Recall@k, Precision@k, AUC-ROC	LIG-Doctor, LSTM, GRU, DoctorAI
Meng et al. (2021)	EHR, 43,967 patients [dim = 4]	Masked language model (MLM)	Depression prediction in windows of two weeks and one year	AUC-ROC, AUPRC	Dipole, MiME*, HCET, and BERHT
Fouladvand et al. (2021)	IBM MarketScan Commercial Claims32 database, 392,492 patients [dim = 2]	–	–	Precision, F1-score, AUC-ROCss	Original transformer model, LSTM, LR, RF, and SVM
Boursalie et al. (2021)	Dataset from four Canadian hospitals, 66,906 patients [dim = 2]	–	–	Precision, Recall	Med-BERT
Chen et al. (2021b)	a1,019,437 visits from Taiwan hospitals, EHR [dim = N/A]	Pretraining BERT using free text medical records	Prediction of any critical event within three days after an emergency	AUC-ROC	Specific prediction models. See references in Chen et al. (2021b)

Table 4 (continued)

Refs	Dataset characterization, Dimensionality	Pretraining strategy	Fine-tuning strategy	[b]Evaluation Criteria	[c,d]Comparison to other approaches
Li et al. (2021)	Epidemiologic datasets (Japan-prefectures), US-HHS and US-census [dim = N/A]	–	–	RMSE, MAE, PCC	AR, GAR (GAR), ARIMA, RNN, DA-RNN
Ye et al. (2020)	Three real world EHR datasets (12,320, 11,240, and 9540 patients) [dim = 1]	–	–	AUC-ROC, Precision, Recall, F1Score	SVN, LR, RF, LSTM, GRU, Dipole-, Dipole, Retain, RetainEx, t-LSTM, Timeline
Rasmy et al. (2021)	Cerner and Truven EHR dataset, 28 million patients [dim = 1]	MLM + prediction of prolonged length of stay in hospital	Prediction of heart failure among patients with diabetes, and pancreatic cancer onset prediction	AUC-ROC	GRU, Bi-GRU, RETAIN, LR, RF
Zeng et al. (2022)	Partner for Kids (PFK) claims data (model pre-training) with 1,881,020; suicide claims (suicide prediction) with 79,350 patients; asthma claims (asthma exacerbation prediction) with 22,862 patients; PFK-2013 with 160,339; and MIMIC-3 (validation of knowledge transfer) with 7537 patients [dim = 4]	Next Visit Prediction (NVP) objective and Categorial Prediction (CP) objective functions	Suicide and asthma exacerbation predictions using specific datasets for each case	AUC-ROC	LR Sparse, LR Dense, DoctorAI, Dipole, TransE
(Shome 2021)	UCI HAR (30 users) and WISDM datasets (51 users) [dim = N/A]	–	–	Confusion matrices	–

Table 4 (continued)

Refs	Dataset characterization, Dimensionality	Pretraining strategy	Fine-tuning strategy	[b]Evaluation Criteria	[c,d]Comparison to other approaches
Dong et al. (2021)	Mobile sensing dataset collected from around 1,300 participants in the wild [dim = 4]	–	–	F1-score AUC-ROC AUPRC	GCN, GIN, GRU, Deep Set, Set Tfmer
Darabi et al. (2020)	MIMIC-III ICU dataset, 38,597 patients [dim = 4]	Pretrained BERT model initialized from BioBERT	Binary task for readmission (in 30 days) and mortality predictions	AUPRC AUC-ROC	Several methods, such as ClinicalBert and MCE (see [44] for complete list)
(Li et al. (2023a, b)	Clinical Practice Research Datalink (CPRD). 4,063,811 patients [dim = 8]	MLM and BYOL (Bootstrap your own latent)	Predict the 5-year risk of heart failure, diabetes, chronic kidney disease	AUROC AUPRC	BEHRT and Med-BERT
Pang et al. (2021)	Columbia University Irving Medical Center-New York Presbyterian Hospital, 2.5 M [dim = 3]	MLM and VTP (Visit Type Prediction	Prediction of hospitalization, death, new heart failure (HF) diagnosis, and HF readmission	AUC-ROC AUPRC Ablation	BEHRT and Med-BERT
Ren et al. (2021)	Hospital in Beijing, pregnant women with hypertension (2,872) and diabetes (10,080) [dim = 1]	Similarity prediction, masked prediction, and reasonability check	Sigmoid classifier for gestational diabetes and hypertension prediction, pregnancy outcome, and risk period	Accuracy, precision, recall, F1-score, and AUC. Ablation	LSTM, Transformer, RETAIN, T-LSTM, Dipole, HiTANet
Prakash et al. (2021)	Proprietary Symphony Health's IDV® dataset with 3,670 XLHe and 263,187 unlabeled patients [dim = 3]	MLM-based approach using	XLH diagnosis Positive unlabeled (PU)-learning based approach	AUPRC	LSTM and Med-BERT
Rao et al. (2022b)	6,777,845 CPRD patients. [dim = 3]	Masked EHR modeling, with variants for temporal and static features	Risk of treatment	Sum absolute error (SAE)	LR, TMLE, BART, TARNET, Dragonnet

Table 4 (continued)

Refs	Dataset characterization, Dimensionality	Pretraining strategy	Fine-tuning strategy	[b]Evaluation Criteria	[c,d]Comparison to other approaches
An et al. (2022)	Subset of 10,000 MIMIC-III and MIMIC-IV patients [dim = N/A]	–	–	Precision, recall, F1-score, and AUC	Several methods, such as TimeLinet and GRUD. See An et al. (2022) for complete list
Peng et al. (2021)	MIC II (7,499) and MIMIC-III (73,452) [dim = 1]	–	–	Accuracy@k	RETAIN, Dipole, GRAM, KAME, MMORE

[a]The number of patients is not clear

[b]Area Under the Precision-Recall Curve (AUPRC), Area Under the Curve-Receiver Operating Characteristic (AUC-ROC), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Pearsons Correlation (PCC)

[c]Linear Regression (LR), Random Forest (RF), support vector machine (SVM), Autoregressive (AR), Global Autoregression (GAR), Autoregressive Integrated Moving Average (ARIMA), Recurrent Neural Network (RNN), Graph Convolution Network (GCN), Graph Isomorphic Network (GIN), gated recurrent neural network (GRU)

[d]References for the methods indicated in this column can be found in the papers that used such methods (references in the first column)

[e]XLH means Xlinked hypophosphatemia, a rare disease that affects bones, muscles, and teeth due to the excessive loss of phosphate

Operating Characteristic (AUC-ROC). Indeed, this performance measurement is interesting because it visually tells how much the model can distinguish between classes (degree or measure of separability) at various threshold settings. Finally, the sixth column shows that comparative analysis is a frequent way to validate the approaches (**EvaRQ5**). However, different from the AUC-ROC, which is almost a default performance measurement, the approaches use diverse techniques for this analysis.

4.6 Explainability question

While the explainability/interpretability of results is desirable for inductive systems, such a feature is compulsory for health support systems. Indeed, previous works (Shortliffe and Sepúlveda 2018; Amann et al. 2020) show that this lack of explainability, which causes legal and ethical uncertainties, is one of the main barriers that impede the advance of machine learning techniques in the health domain. In this context, we could expect discussions about the implementation and evaluation (**EvaRQ6**) of explanainability in the reviewed works. The following schema (Table 5) summarizes our findings regarding the use of explainability in the reviewed papers.

We defined four groups. The first group lists references that do not consider explainability in their research. The second group contains references that agree about the importance of explainability. Such a topic is indicated as one of their future research directions (Fouladvand et al. 2021; Li et al. 2021; Zeng et al. 2022; An et al. 2022). The third group lists references that exclusively rely on the attention weights to associate input features with results. For example, the works (Meng et al. 2021) and (Ye et al. 2020) employ attention weights to relate symptoms to disease codes. Thus, they evaluate the explanations using a quantitative analysis of such weights. The works (Li et al. 2020) and (Rasmy et al. 2021) also rely on attention weights, but they use a visualization tool (Vig 2019) that supports and augments the analysis of such weights. For example, this tool allows detecting model bias, locating relevant attention heads, and linking neurons to model behavior.

The last group also relies on attention weights but uses additional techniques. For example, the work in Rao et al. (2022a) employs perturbation-based techniques (Ivanovs et al. 2021) to show the importance of different contexts in prediction. These techniques are model-agnostic and not exclusive to transformers since they only perturb the input and observe changes in the output. Unlike model-agnostic methods, model-specific strategies, which take advantage of the particularities of neural network architectures, were not identified in our review. This last group brings two works (Dong et al. 2021; Peng et al. 2021) that show an interesting trend of combining inductive architectures with symbolic approaches to augment the explainability power. The work in Dong et al. (2021) also uses attention weights to show the importance of the input elements. However, such inputs are given in the form of graphs that relate concepts of the domain. Thus, this approach augments the explainability of the attention mechanism since it relies on the attention weights assigned for each graph instance rather than only weights between inputs and outcomes. The work in Peng et al. (2021) shows that its model learns interpretable representations according to the structure of an ontology given as input. Thus, it is possible to derive more interpretable embeddings of medical codes. However, capturing all relevant knowledge within an ontology can be difficult. For example, some nuances, tacit knowledge, or rapidly evolving information may be challenging to be represented accurately. Moreover, if the ontology presents biased data, inconsistent or incorrect definitions, relationships, or concepts, it can lead to errors in the interpretable representations.

Table 5 Use of explainability in the reviewed papers

Do not consider explainability	Consider explainability as research directions	Unique use of attention weights	Use of additional techniques
Florez et al. (2021), Boursalie et al. (2021), Chen et al. (2021b), Shome 2021), Darabi et al. (2020), Li et al. (2023a, b), Pang et al. (2021), Ren et al. (2021), Prakash et al. (2021), Rao et al. (2022b)	Fouladvand et al. (2021), Li et al. (2021), Zeng et al. (2022), An et al. (2022)	Li et al. (2020), Meng et al. (2021), Ye et al. (2020), Rasmy et al. (2021)	Rao et al. (2022a), Dong et al. (2021), Peng et al. (2021)

5 Discussion

This section relies on the results of our review to identify important issues that limit the ability of transformers in handling longitudinal health data. Therefore, we summarize such issues and discuss initial efforts that we have identified to address them.

5.1 Data diversity

This review shows that clinical-based sparse electronic health records are the main multi-featured longitudinal data source for transformers in the health domain (Sect. 4.3). However, only a few classes of features (e.g., medication and diagnosis codes) have been considered at the same time thus far. An interesting exception is the work in Li et al. (2023a, b), which uses several health data types (diagnosis, medication, procedure, examinations, blood pressure, drinking status, smoking status, and body mass index). However, such data are part of the same vocabulary, and thus, the learning process cannot explore the particular semantics of each. In other words, the model does not know if a token represents, for example, a medication or a diagnosis. A different approach explicitly indicates this type, such as in Prakash et al. (2021). This additional semantics can buster the efficiency and effectiveness of the learning process, similar to natural language processing, when the model knows that a token represents a noun or an adjective.

In this context, the taxonomy for health data proposed by Mayo et al. (2017) shows that health records are mainly composed of Clinician-reported (ClinRO—evaluations from a trained professional) and performance-reported outcomes (PerfRO—e.g., tests of walking, dexterity, and cognition). However, these outcomes are usually complemented with other types of daily-life assessments using technology-reported outcomes (TechRO, e.g., wearables). Thus, longitudinal health data can be modelled with a set of vocabularies beyond the simple use of diagnoses and medication (Li et al. 2020; Rao et al. 2022a; Fouladvand et al. 2021; Boursalie et al. 2021). Approaches that employ multiple vocabularies (Meng et al. 2021) usually sum or concatenate their inputs to generate a unique data stream (Li et al. 2020; Rao et al. 2022a). A different approach is proposed in Fouladvand et al. (2021), which considers distinct feature streams and their combination is conducted during the calculation of the attention weights. This approach seems to better explore the relationship between the feature types. However, its complexity is exponential regarding the number of streams since the model needs to conduct more operations and maintain the results of these operations in memory for the next steps.

We are currently investigating alternatives to address this problem. According to the representation proposed by many of the approaches of our review, the history of visits of each patient (Vp) is represented as in Eq. (1), where CLS and SEP are the *start* and *separate* special words, v_p^i represents each visit i of patient p, and n is the total number of visits of p.

$$V_p = \{CLS, v_p^1, SEP, v_p^2, SEP, \ldots, v_p^n, SEP\} \tag{1}$$

$$v_p^i = \{d_1, d_2, \ldots, d_m\}, \text{where } [d_1..d_m] \in D \tag{2}$$

In Eq. (2), visits are composed of m words (diagnosis codes d_i) that are part of a unique vocabulary D. Then, the challenge is to extend the representation in (2) towards

multifeature inputs from different vocabularies (e.g., clinical examinations (E), diagnoses (D), and treatments (T), such as in (3):

$$v_p^j = \{e_1, e_2, \ldots, e_h, \text{SEP}', d_1, d_2, \ldots, d_m, \text{SEP}', t_1, t_2, \ldots, t_w, \text{SEP}'\}, \quad \text{where } [e_1..e_h] \in \text{E}, [d_1..d_m] \in \text{D}, [t_1..t_w] \in \text{T} \tag{3}$$

This representation uses a new special word SEP' to set different positions inside visits for words of different vocabularies. However, this approach brings implications to the architecture since the word/sentence/document concept is broken. Thus, this approach may also have implications for the model's accuracy. Another possible strategy to overcome the multiple vocabulary representation is to use a similar idea than segment embeddings to distinguish elements of different vocabularies. This approach seems more natural, and a similar strategy is already explored in Prakash et al. (2021). The following schema (Fig. 5) illustrates this idea using an example.

According to this strategy, a further embedding called "Vocab_type" could be added to the other inputs, including the semantics of each vocabulary to the final embedding. However, this approach generates an overload of information in the final embedding. Thus, we must evaluate the real predictive value of the embeddings to eliminate or represent them differently.

5.2 Temporal handling

Table 2 shows that most approaches use the same positional encoding principles to provide the notion of position (or order) to input tokens. While such encoding works well for textual data, since a text is just a homogeneous sequence of sentences (or words), they represent a limitation to modeling clinical data. By depicting diagnoses as words, each visit as a sentence, and a patient's entire medical history as a document, this representation does not allow to include the notion of temporal (variable and unpredicted) distance between visits and, consequently, diagnoses or any other concept represented inside the visits. Some papers of our review (Rao et al. 2022a; Boursalie et al. 2021), for example, define the position as a continuous crescent numeric stamp for each visit. However, this simple strategy becomes problematic for long periods since the high values of latter positional embeddings

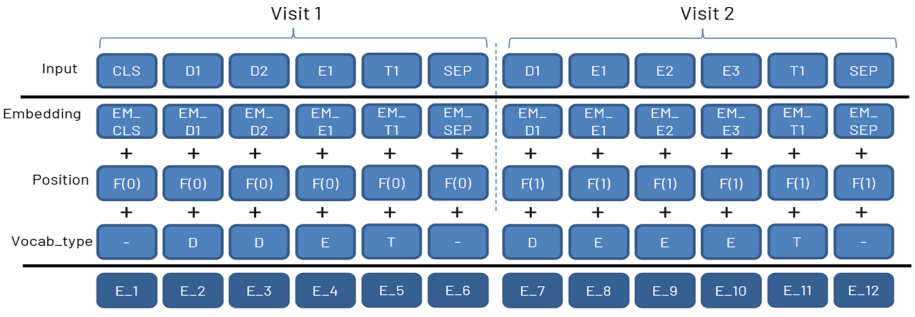

Fig. 5 Possible representation for different input vocabularies (e.g., diagnoses D, examinations E, treatments T)

may dominate the initial values, distorting final results. Moreover, it does not capture the notion of time distance between clinical events.

In this context, we have observed a trend in incorporating temporal notions, such as the inclusion of temporal distances as part of the input (Boursalie et al. 2021; An et al. 2022). While the approach in An et al. (2022) only sums the temporal and input representations, the approach in Boursalie et al. (2021) includes this distance as a token that is not part of a vocabulary. The work in Pang et al. (2021) handles this limitation by creating an artificial vocabulary for temporal distances (e.g., token LT to temporal distance longer than 365 days). However, this approach may lose precision. The approach in Peng et al. (2021) augments this temporal notion since it also considers the temporal length of visits rather than uniquely the distance between visits. Therefore, these two pieces of information (length of visit and distance between current and previous visits) are added to the input content representation (numeric medical codes). Another alternative is to explicitly adapt the positional encoding for a date/time encoding. For example, the work in Ren et al. (2021) uses a continuous value of weeks as input of the positional encoding.

Apart from strategies representing the temporal notion, longitudinal health data present issues such as sparse and irregular time assessment intervals. According to Li et al. (2023a, b) and corroborated with our review, while self-attention is a flexible and efficient learning mechanism for longitudinal analysis, its ability to interpret temporal distance between sparse, irregular data has yet to be investigated. Thus, time representation and modifications in the attention mechanisms may need to be conducted jointly.

5.3 Continuous numerical data

While several works follow the original ideas of transformers (vocabulary definition, pretraining, and fine-tuning) using categorical data as input, other proposals use numeric values as input. The need to handle such data is mainly derived from the current trend of using mobile health technology—mHealth (e.g., wearables) to assess multifeature longitudinal health data. For example, the works (Shome 2021) and (Dong et al. 2021) are examples in such a direction, while the taxonomy of Mayo et al. (2017) for health data already considers such a technology (TechROs) as a source of health information.

Current approaches that use continuous numerical data do not take advantage of the original transformers' training methods since they use the traditional three steps (forward propagation, calculation of the loss function, backward propagation) neural network learning process. A research line is the creation of vocabularies that relies on such continuous data. For example, the work in Li et al. (2023a, b) mixes standard vocabularies to represent tokens of diagnosis (ICD-10) and medications (British National Formulary—BNF) with defined vocabularies for continuous values. For example: "For continuous values, we included systolic pressure, diastolic pressure, and BMI within range 80 to 200 mmHg, 50 to 140 mmHg, and 16 to 50 kg/m^2, respectively. Afterwards, we categorized systolic and diastolic pressure into bins with a 5-mmHg step size (e.g., 80–85 mmHg). BMI was processed the same way with a step size 1 kg/m^2." (Li et al. 2023a, b). According to Siebra et al. (2022), the number of categories affects the accuracy of the predictions and can create imbalanced clusters. Thus, the specification of the intervals must consider the clusters balance and use of standard feature-specific categories. For example, the heart rate feature is usually categorized into five zones regarding the level of physical activities: recovery/easy, aerobic/base, tempo, lactate threshold, and anaerobic (Shortliffe and Sepúlveda

2018). If such a categorization creates imbalanced clusters, such clusters could be merged or divided to avoid overfitting issues in the learning process.

5.4 Validation

We observed that most contributions are associated with defining ways to encode the input data rather than significant redefinitions of the transformer architecture. In other words, the proposals are focused on how to encode multifeatured data so the transformers' equations can process such data. This focus on input embeddings becomes evident when we analyze the last column (Input Embedding) of Table 2, which shows diverse combinations and forms to embed the input data. However, some design decisions are not clear in the surveyed works. For example, proposals such as (Li et al. 2020) and (Chen et al. 2021b) use A/B segment embeddings to provide extra information to differentiate codes in adjacent longitudinal units. The validation of this strategy is hard without the execution of experimental analysis. Some works (Pang et al. 2021; Ren et al. 2021; Li et al. 2023a, b; Pang et al. 2021; Ren et al. 2021; Prakash et al. 2021; Rao et al. 2022b; An et al. 2022; Peng et al. 2021) use ablation analysis to demonstrate the influence of specific architectural elements on task accuracy. This analysis is very interesting since it supports, for example, simplifying the approach when we identify elements (e.g., A/B segment or the use of SEP special token) that are not significantly contributing to the learning process.

In general, transformers suffer from a lack of benchmarks. Each work uses its own datasets, which are mostly not public. Even the use of the AUC-ROC measurement, which is a trend among the works, presents some problems. According to the authors of Meng et al. (2021), for example, the AUC-ROC definition in Li et al. (2020) was nonstandard, making it difficult to compare their results with other studies. In common, works that have compared their approaches with traditional deep learning techniques (Florez et al. 2021; Fouladvand et al. 2021; Ye et al. 2020; Rasmy et al. 2021) agree on the superior performance of transformer-based approaches. Only two papers (Li et al. 2020; Rasmy et al. 2021) involved clinicians during the validation process. Li et al. (2020) identified a rate of 76% of overlapping between clinical researchers and automatic decisions regarding the top 10 closest diseases to each of the 87 most common diseases. In the second paper (Rasmy et al. 2021), clinical experts were involved to verify the reliability of the semantic relations between diseases identified by the models. Apart from these simple cases, none of the approaches were adopted or presented follow-up papers that show the application or validations of the proposals in real scenarios. Indeed, Table 4 suggests that performance comparisons are only conducted using other computational approaches. This inexistence of such advanced evaluation processes may be associated with the need for regulations, as discussed in the previous comments. In other words, the technology transition from academia to the market requires well-defined regulatory compliances to guide such a process.

5.5 Explainability and ethics

Our review shows that several transformer architectures provide and evaluate approaches for explainability, aiming to engender trust with the healthcare professionals and provide transparency to the decision-making process. However, such approaches still need to evolve to be used in clinical practice. Moreover, in the mid-term, AI regulations make explainability of the decision models a requirement for so-called *high-risk AI applications/*

domains, such as health. An example of such regulations is the European Union Artificial Intelligence Act (EU AI Act), which will regulate the use of AI in the EU and enforce the use of explainable models (Panigutti et al. 2023). However, as identified in our review, the current approaches for explainability mostly rely on identifying the importance of input features to the models' outcomes (predictions or classifications). Such approaches are still raising several ethical considerations (Ghassemi et al. 2021) since they cannot, for example, identify when AI models inherit biases in the data used for their training or assist in mitigating and ensuring that the generated models provide equitable and fair healthcare outcomes. Moreover, explainability strategies must also be extended to characterize better liability regarding potential errors, which is a critical aspect for legal and ethical reasons in clinical practice (Naik et al. 2022).

Approaches that address these and other ethical questions may employ neuro-symbolic strategies, which integrate symbolic knowledge (e.g., ontologies, knowledge graphs) into the inductive reasoning process. These strategies can be designed to incorporate explicit ethical rules and principles, as well as be programmed with ethical guidelines to guide their decision-making and prevent them from making unethical choices. The study of Dong et al. (2021), for example, represents the first step in this direction as it relies on the weights assigned for each graph. In other words, this process indicates the concepts and relations (part of a graph) that are important to the modeling of outcomes. However, research advances are still required in this area since ethical considerations regarding the use of AI in health must imply multiple, complex and evolving requirements, including data privacy, transparency, accountability, fairness, and more, and addressing these concerns requires a multi-faceted approach that extends beyond the choice of the transformer architectures, or an AI model or a strategy of applying it.

5.6 A guideline for pragmatic development

We relied on the knowledge obtained in this review to create a guideline that considers pragmatic requirements (e.g., data diversity, temporal handling, continuous values, and explainability generation) that represent strong barriers to using the transformer-based approaches in real scenarios. Then, we map proposals identified in this review to handle each requirement (Fig. 6), discussing their weaknesses and benefits.

The first requirement is the inclusion of data diversity (R1). Our review showed that several proposals are based on a unique vocabulary, such as for diagnosis codes (Li et al. 2020; Florez et al. 2021). There are two main directions to increase this diversity:

- At the level of vocabulary and embeddings (R1a): This approach defines different vocabularies and one or more embeddings to indicate the vocabulary of each token. The use of different vocabularies impacts the size of the unit of analysis (e.g., visit), which tends to contain several tokens of different vocabularies. The work of Li et al. (2023a, b) proposed the use of a hierarchy of transformers to handle these long sequences. While this "divide-and-conquer" approach can cover the complexity of long sequences, it increases the computational complexity given the number of transformers used to handle each part of this hierarchy.
- At the level of the internal transformer architecture (R1b): In this case, the architecture formulation is modified to receive different data streams (Fouladvand et al. 2021). The main advantage is that outputs capture the associations between input streams, represented as attention weights between different tokens across such streams. This approach

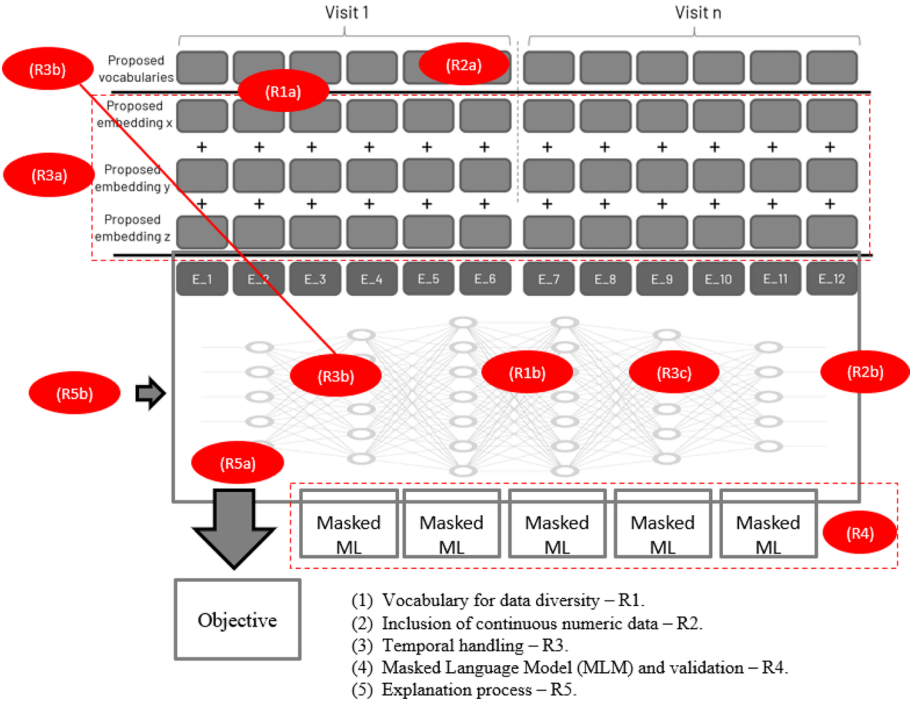

Fig. 6 Visual schema emphasizing the architectural components affected by the requirements implementation according to different approaches

preserves the original semantics of each stream while the network units handle them. However, the weakness of this approach is the exponential complexity regarding the number of streams since each stream must interact with all other streams. As the work of Fouladvand et al. (2021) only demonstrated this approach using two data streams, we cannot attest its adequacy when more streams are used.

The second requirement is associated with the use of continuous data since transformers are based on discrete vocabularies (R2). Two main approaches can be used in this situation:

- At the vocabulary level (R2a): This case relies on categorizing continuous values to create a vocabulary for each of their types. As discussed by Siebra et al. (2022), while this process is simple, it presents the loss of granularity as a drawback.
- At the level of input of one of the transformer's layers (R2b): In this case, pre-processing strategies can create a feature vector that is integrated into one of the layers of the transformers or even in its input. Shome (2021) follows this approach using a set of CNN-based blocks to create a feature vector as input of a transformer architecture. This approach has as the main advantage the use of CNN as a feature extractor module since this CNN ability is already recognized in the machine learning area. Several CNN-based blocks can be part of this architecture, and skipping connections enable the network to learn negligible weights for the layers that are not relevant and do not contribute to overall accuracy. Thus, the number of blocks is not an essential hyperparam-

eter to tune. However, this architecture was exclusively designed to handle continuous values. The inclusion of no continuous data can generate a problem of temporal scale since the architecture must deal with data from different temporal granularities.

The third requirement is associated with temporal handling since transformers only represent the notion of sequence (R3). We identified three different approaches to implement such a requirement:

- At the level of embeddings (R3a): This is the most common approach, and it is implemented using adaptations of the positional encoding (Florez et al. 2021; Fouladvand et al. 2021). While implementing this approach is simple, with minimal changes in the original transformer's architecture, its expressiveness is poor since it cannot explicitly represent temporal distances between clinical events.
- At the vocabulary level (R3b): As proposed by Boursalie et al. (2021), the temporal distance between patients' visits could be part of the input data. As an advantage, the input representation is simple since temporal distances between two patient visits are considered as input tokens. However, temporal distances are infinite numeric values rather than categorical values. This representation must be modified since distances must be employed as tokens that are part of a discrete vocabulary. Moreover, the architecture must also be modified, as exemplified by Boursalie et al. (2021), so it can analyze the main content (patient visit information) together with its distance to other visits. This approach increases the complexity of the training process, requiring a high amount and diversity of data.
- At the level of internal transformer architecture (R3c): This approach considers that temporal information is directly included in one of the layers of the architecture. The work of Peng et al. (2021) is an example of such a direction. This approach increases the expressiveness of representations since both the distance between visits and the length of each visit can be integrated into the learning process. However, the form as this information is integrated into the architecture can affect accuracy. For example, Peng et al. (2021) implemented this integration using an "Add & Normalize" layer. Thus, the semantics are mixed before reaching the core transformer components.

The fourth requirement concerns using the Masked Language Model (MLM) as the main strategy for pretraining (R4). The changes to handle the three previous requirements may invalidate the standard MLM process and parameters. Thus, other strategies or modifications should be considered. For example:

- The literature usually employs a 15% probability of masking in the MLM process. While this value works as a default value, it is not an absolute consensus (Wettig 2022). Moreover, using more than one vocabulary can create other needs that affect this probability.
- The representation of temporal notions and data diversity has the advantage of increasing the expressiveness of the models. On the other hand, such expressiveness also increases the complexity of the learning process. Therefore, this scenario requires further strategies to buster this process. One of these strategies is to modify a percentage of the words with randomly chosen words from the vocabulary. This action is similar to inserting noise into the learning process, which creates more robust results (Li et al. 2020).

The fifth requirement is related to the generation of explanations (R5). The approaches found for explainability can be implemented at two levels:

- At the level of weights (R5a): This simple method creates visual descriptions regarding the weight values of the network. While simple, it is also restricted since the main information is only the strongness of the relationships between tokens.
- At the level of external knowledge (R5b): This approach is more complex, and current proposals only implement simple versions of this solution to integrate symbolic knowledge into the learning process (Dong et al 2021; Peng et al. 2021). These solutions represent inputs in the form of knowledge graphs or ontologies. While such methods are well-established in the artificial intelligence community, they present limitations for longitudinal representations (Siebra and Wac 2022).

A final remark is about the integrity of the architecture. The efficacy of transformers comes from their architecture. Thus, adaptations should minimally affect its structure or be very well justified and validated. Otherwise, such adaptations may raise several side effects that are hard to understand.

6 Conclusion

There is a huge amount of knowledge codified in the health datasets (e.g., EHRs), derived from the experience of a large number of experts for several years. As we show in this paper, current transformer models rely on such knowledge to make conclusions that are impossible or very hard to derive by humans due to the amount and complexity of relations involved. The approaches discussed in this review try to demonstrate a future where this ability can be leveraged accurately as a decision-support tool for healthcare experts. As the use of transformers to analyze multifeatured longitudinal health data is recent, we have not identified a convergence regarding aspects such as positional encoding, input embedding, or training strategies using categorical or numerical values. Differently, we have identified studies for temporal handling and explanation as the two main research trends in this area. Temporal handling is a compulsory requirement for the health domain and the inexistence of such ability is a barrier for the use of the transformer technology in real applications. Similarly, explainability is also becoming a compulsory requirement for deep learning models, according to the upcoming AI regulations. Indeed, the explainability for transformers models and their results are in the initial stage, and this area requires strategies beyond the simple analysis of attention weights. These open questions are opportunities for research directions, which must mainly consider replicable forms to compare and justify their designs. To the best of our knowledge, this is the first review that analyzes proposals that adapt the transformers technology for longitudinal health data. Other reviews focused on general aspects of transformers (Lin et al. 2022), time series (Wen et al. 2022) and computational vision (Khan et al. 2022; Liu et al. 2021b). We have also used comprehensive language, avoiding as much as possible the use of complex equations, such as in Lin et al. (2022), so this text could be an important reference for research groups that work within the boundaries of interdisciplinary health informatics research.

Author contributions CS: conceptualization, methodology, investigation (execution of review) and writing (original draft). MK: investigation (execution of review) and writing (review & editing), KW: supervision, investigation (execution of review) and writing (review & editing).

Funding Open access funding provided by University of Geneva. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grand agreement H2020-MSCA-IF-2020- 101024693.

Declarations

Competing interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Amann J et al (2020) Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 20(310):1–9

An Y, Liu Y, Chen X, Sheng Y, Hošovský A (2022) TERTIAN: clinical endpoint prediction in ICU via time-aware transformer-based hierarchical attention network. Comput Intell Neurosci 4207940:1–13

Bao H, Dong L, Wei F (2021) Beit: Bert pre-training of image transformers, arXiv preprint arXiv:2106.08254

Boursalie O, Samavi R, Doyle TE (2021) Decoder transformer for temporally-embedded health outcome predictions. In: 20th IEEE International conference on machine learning and applications (ICMLA), pp 1461–1467

Chen YP, Chen YY, Lin JJ, Huang CH, Lai F (2020) Modified bidirectional encoder representations from transformers extractive summarization model for hospital information systems based on character-level tokens (AlphaBERT): development and performance evaluation. JMIR Med Inform 8(4):e17787

Chen D et al. (2021a) Early detection of post-surgical complications using time-series electronic health records. In: AMIA summits on translational science proceedings, pp 152–160

Chen YP, Lo YH, Lai F, Huang CH (2021b) Disease concept-embedding based on the self-supervised method for medical information extraction from electronic health records and disease retrieval: algorithm development and validation study. J Med Internet Res 23(1):e25113

Chen PF et al (2022) Predicting postoperative mortality with deep neural networks and natural language processing: model development and validation. JMIR Med Inform 10(5):e38241

Culurciello E (2018) The fall of RNN/LSTM, towards data science. https://towardsdatascience.com/the-fall-of-rnn-lstm-2d1594c74ce0. Accessed 26 July 2023

Darabi S, Kachuee M, Fazeli S, Sarrafzadeh M (2020) Taper: time-aware patient ehr representation. IEEE J Biomed Health Inform 24(11):3268–3275

Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805

Diggle P, Diggle PJ, Heagerty P, Liang KY, Zeger S (2002) Analysis of longitudinal data. Oxford University Press, Oxford

Dong G, Tang M, Cai L, Barnes LE, Boukhechba M (2021) Semi-supervised graph instance transformer for mental health inference. In: 20th IEEE International conference on machine learning and applications (ICMLA), pp 1221–1228

Dufter P, Schmitt M, Schutze H (2021) Position information in transformers: an overview, arXiv preprint arXiv:2102.11090.

Falissard L, Morgand C, Ghosn W, Imbaud C, Bounebache K, Rey G (2022) Neural translation and automated recognition of ICD-10 medical entities from natural language: model development and performance assessment. JMIR Med Inform 10(4):e26353
Florez AY, Scabora L, Eler DM, Rodrigues JF (2021) APEHR: automated prognosis in electronic health records using multi-head self-attention. In: IEEE 34th international symposium on computer-based medical systems (CBMS), pp 277–282
Fouladvand S et al (2021) Identifying opioid use disorder from longitudinal healthcare data using a multi-stream transformer. In: AMIA annual symposium proceedings. American Medical Informatics Association, pp 476–485
Fu Y et al (2022) A resource-efficient deep learning framework for low-dose brain PET image reconstruction and analysis. In: IEEE 19th International symposium on biomedical imaging (ISBI), pp 1–5
Ghassemi M, Oakden-Rayner L, Beam AL (2021) The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health 3(11):e745–e750
Guo A, Beheshti R, Khan YM, Langabeer JR, Foraker RE (2021) Predicting cardiovascular health trajectories in time-series electronic health records with LSTM models. BMC Med Inform Decis Mak 21(1):1–10
He K et al (2022) Transformers in medical image analysis: a review, arXiv preprint arXiv:2202.12165
Huang K, Altosaar J, Ranganath R (2019) ClinicalBERT: modeling clinical notes and predicting hospital readmission. arXiv preprint arXiv:1904.05342
Ivanovs M, Kadikis R, Ozols K (2021) Perturbation-based methods for explaining deep neural networks: a survey. Pattern Recogn Lett 150:228–234
Jagannatha AN, Yu H (2016) Bidirectional RNN for medical event detection in electronic health records. In: Proceedings of the conference. Association for Computational Linguistics, North American Chapter. Meeting, vol 2016, pp 473–482
Khan S, Naseer M, Hayat M, Zamir SW, Khan FS, Shah M (2022) Transformers in vision: a survey. ACM Comput Surv (CSUR) 54(10s):1–41
Kitchenham B (2004) Procedures for performing systematic reviews. Keele University, Keele, vol 33, pp 1–26
Li Y et al (2020) BEHRT: transformer for electronic health records. Sci Rep 10(1):1–12
Li L, Jiang Y, Huang B (2021) Long-term prediction for temporal propagation of seasonal influenza using Transformer-based model. J Biomed Inform 122:103894
Li Y et al (2023a) Hi-BEHRT: hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records. IEEE J Biomed Health Inform 27(2):1106–1117
Li T et al (2023b) Time-distance vision transformers in lung cancer diagnosis from longitudinal computed tomography. Med Imaging 12464:221–230
Lin T, Wang Y, Liu X, Qiu X (2022) A survey of transformers AI Open (In press)
Liu Y, Yang Y, Jiang W, Wang T, Lei B (2021a) 3d deep attentive u-net with transformer for breast tumor segmentation from automated breast volume scanner. In: 43rd Annual international conference of the IEEE Engineering in Medicine & Biology Society, pp 4011–4014
Liu Y et al (2021b) A survey of visual transformers, arXiv preprint arXiv:2111.06091
Liu L, Liu S, Zhang L, To XV, Nasrallah F, Chandra SS (2023) Cascaded multi-modal mixing transformers for alzheimer's disease classification with incomplete data. Neuroimage 277:120267
Mahajan D et al (2020) Identification of semantically similar sentences in clinical notes: Iterative intermediate training using multi-task learning. JMIR Med Inform 8(11):e22508
Mao S, Sejdić E (2022) A review of recurrent neural network-based methods in computational physiology. In: IEEE transactions on neural networks and learning systems
Mayo NE, Figueiredo S, Ahmed S, Bartlett SJ (2017) Montreal accord on patient-reported outcomes (pros) use series–paper 2: terminology proposed to measure what matters in health. J Clin Epidemiol 89:119–124
Meng Y, Speier W, Ong MK, Arnold CW (2021) Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression. IEEE J Biomed Health Inform 25(8):3121–3129
Mondal AK, Bhattacharjee A, Singla P, Prathosh AP (2021) xViTCOS: explainable vision transformer based COVID-19 screening using radiography. IEEE J Transl Eng Health Med 10:1–10
Naik N, Hameed BM, Shetty DK, Swain D, Shah M, Paul R et al (2022) Legal and ethical consideration in artificial intelligence in healthcare: who takes responsibility? Front Surg 9:266
Pang C, Jiang X, Kalluri KS, Spotnitz M, Chen R, Perotte A, Natarajan K (2021) CEHR-BERT: incorporating temporal information from structured EHR data to improve prediction tasks. In: Proceedings of machine learning for health, pp 239–260

Panigutti C, Hamon R, Hupont I, Fernandez Llorca D, Fano Yela D, Junklewitz H et al (2023). The role of explainable AI in the context of the AI Act. In: Proceedings of the 2023 ACM conference on fairness, accountability, and transparency, pp 1139–1150

Peng X, Long G, Shen T, Wang S, Jiang J (2021) Sequential diagnosis prediction with transformer and ontological representation. In: Proceedings of the IEEE International conference on data mining, pp 489–498

Perveen S, Shahbaz M, Saba T, Keshavjee K, Rehman A, Guergachi A (2020) Handling irregularly sampled longitudinal data and prognostic modeling of diabetes using machine learning technique. IEEE Access 8:21875–21885

Prakash PKS, Chilukuri S, Ranade N, Viswanathan S (2021) RareBERT: transformer architecture for rare disease patient identification using administrative claims. Proc AAAI Conf Artif Intell 35(1):453–460

Rao S et al (2022a) An explainable transformer-based deep learning model for the prediction of incident heart failure. IEEE J Biomed Health Inform 26(7):3362–3372

Rao S et al (2022b) Targeted-BEHRT: deep learning for observational causal inference on longitudinal electronic health records. IEEE Trans Neural Netw Learn Syst. https://doi.org/10.1109/TNNLS.2022.3183864

Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D (2021) Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ Digit Med 4(1):1–13

Ren H, Wang J, Zhao WX, Wu N (2021) Rapt: pre-training of time-aware transformer for learning robust healthcare representation. In: Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining, pp. 3503–3511

Severson K et al (2021) Discovery of Parkinson's disease states and disease progression modelling: a longitudinal data study using machine learning. Lancet Digital Health 3(9):e555–e564

Shibly MMA, Tisha TA, Islam MK, Uddin MM (2020) Transfer learning in classifying prescriptions and keyword-based medical notes. In: the 22nd International conference on information integration and web-based applications & services, pp. 82–90

Shome D (2021) RestHAR: residual feature learning transformer for human activity recognition from multi-sensor data. In: 8th International conference on soft computing & machine intelligence (ISCMI), pp. 181–185

Shortliffe EH, Sepúlveda MJ (2018) Clinical decision support in the era of artificial intelligence. J Am Med Assoc 320:2199–2200

Siebra C, Matias I, Wac K (2022) Behavioral data categorization for transformers-based models in digital health. In: 2022 IEEE-EMBS International conference on biomedical and health informatics (BHI), Ioannina, Greece, pp 01–04

Svyatkovskiy A, Deng SK, Fu S, Sundaresan N (2020) Intellicode compose: code generation using transformer. In: the 28th ACM joint meeting on European software engineering conference and symposium on the foundations of software engineering, pp 1433–1443

Tabarestani S et al (2019) Longitudinal prediction modeling of Alzheimer disease using recurrent neural networks. In: 2019 IEEE EMBS international. Conference on biomedical & health informatics (BHI), pp 1–4

Vaswani A et al (2017) Attention is all you need. In: Advances in neural information processing systems, vol 30

Vig J (2019) A multiscale visualization of attention in the transformer model. In: Proceedings of the 57th annual meeting of the Association for Computational Linguistics: system demonstrations, pp. 37–42

Wac K (2016) mQoL: experimental methodology for longitudinal, continuous quality of life assessment via unobtrusive, context-rich mobile computing in situ. In: The International Society for Quality-of-Life Studies Conference (ISQOLS 2016)

Wang X et al (2019) Assessing depression risk in Chinese microblogs: a corpus and machine learning methods. In: 2019 IEEE International conference on healthcare informatics (ICHI), pp 1–5

Wang C, Nulty P, Lillis D (2020) A comparative study on word embeddings in deep learning for text classification. In: Proceedings of the 4th International conference on natural language processing and information retrieval, pp 37–46

Wen Q, Zhou T, Zhang C, Chen W, Ma Z, Yan J, Sun L (2022) Transformers in time series: a survey, arXiv preprint arXiv:2202.07125

Yan T, Meng H, Liu S, Parada-Cabaleiro E, Ren Z, Schuller BW (2022) Convoluational transformer with adaptive position embedding for Covid-19 detection from cough sounds. In: 2022 IEEE International conference on acoustics, speech and signal processing (ICASSP), pp 9092–9096

Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C et al (2022) A large language model for electronic health records. NPJ Digit Med 5(1):194

Yao Y, Yu W, Gao Y, Dong J, Xiao Q, Huang B, Shi Z (2022) W-Transformer: accurate Cobb angles estimation by using a transformer-based hybrid structure. Med Phys 49(5):3246–3262

Ye M, Luo J, Xiao C, Ma F (2020) Lsan: modeling long-term dependencies and short-term correlations with hierarchical attention for risk prediction. In: 29th ACM International conference on information & knowledge management, pp 1753–1762

Zeng X, Linwood SL, Liu C (2022) Pretrained transformer framework on pediatric claims data for population specific tasks. Sci Rep 12(1):1–13

Zhao J et al (2019) Learning from longitudinal data in electronic health record and genetic data to improve cardiovascular event prediction. Sci Rep 9(1):1–10