JAMIA Open, 4(3), 2021, 1–8
doi: 10.1093/jamiaopen/ooab064
Research and Applications

Research and Applications

A comparison of attentional neural network architectures for modeling with electronic medical records

Anthony Finch [1], Alexander Crowell[1], Yung-Chieh Chang[1], Pooja Parameshwarappa[1], Jose Martinez[1] and Michael Horberg[1,2]

[1]Kaiser Permanente Mid-Atlantic Permanente Medical Group, Rockville, Maryland, USA and [2]Kaiser Permanente Mid-Atlantic Permanente Research Institute, Rockville, Maryland, USA

Corresponding Author: Anthony Finch, Kaiser Permanente Mid-Atlantic Permanente Medical Group, 2101 E. Jefferson St, Rockville, MD 20852, USA (Anthony.J.Finch@kp.org)

Received 19 March 2021; Revised 8 July 2021; Editorial Decision 14 July 2021; Accepted 16 July 2021

ABSTRACT

Objective: Attention networks learn an intelligent weighted averaging mechanism over a series of entities, providing increases to both performance and interpretability. In this article, we propose a novel time-aware transformer-based network and compare it to another leading model with similar characteristics. We also decompose model performance along several critical axes and examine which features contribute most to our model's performance.

Materials and methods: Using data sets representing patient records obtained between 2017 and 2019 by the Kaiser Permanente Mid-Atlantic States medical system, we construct four attentional models with varying levels of complexity on two targets (patient mortality and hospitalization). We examine how incorporating transfer learning and demographic features contribute to model success. We also test the performance of a model proposed in recent medical modeling literature. We compare these models with out-of-sample data using the area under the receiver-operator characteristic (AUROC) curve and average precision as measures of performance. We also analyze the attentional weights assigned by these models to patient diagnoses.

Results: We found that our model significantly outperformed the alternative on a mortality prediction task (91.96% AUROC against 73.82% AUROC). Our model also outperformed on the hospitalization task, although the models were significantly more competitive in that space (82.41% AUROC against 80.33% AUROC). Furthermore, we found that demographic features and transfer learning features which are frequently omitted from new models proposed in the EMR modeling space contributed significantly to the success of our model.

Discussion: We proposed an original construction of deep learning electronic medical record models which achieved very strong performance. We found that our unique model construction outperformed on several tasks in comparison to a leading literature alternative, even when input data was held constant between them. We obtained further improvements by incorporating several methods that are frequently overlooked in new model proposals, suggesting that it will be useful to explore these options further in the future.

Key words: patient modeling, artificial intelligence, neural network, electronic medical record, attention network

LAY SUMMARY

Recent advances in machine learning techniques have enabled researchers to build highly predictive and interpretable models of patient outcomes using only data from electronic medical records (EMRs). One type of model, called Attentional Neural Networks, has demonstrated great promise in both performance and interpretability. This technique has been widely used in several fields, especially Natural Language Processing. Such models have also been successfully applied to patient outcomes using EMR data. In this article, we propose a novel adaptation of attentional neural networks and compare it to a leading literature alternative. Our results indicate that our model outperforms on several key metrics, and that the model can be further improved by incorporating demographic features and by pretraining part of the model on an unsupervised target.

BACKGROUND

Introduction

With recent advances in computational resources, deep learning has become an increasingly popular methodology for producing models of expected outcomes.[1,2] In general, deep learning has been most successful in domains where unstructured or semistructured data have rendered more conventional models impractical.

One application which has benefited greatly from the advent of deep learning has been the modeling of patient outcomes based on their electronic medical records (EMRs).[3–20] This domain has been particularly ripe for exploration by deep learning models because EMR data is typically sufficiently large to construct high-quality deep learning models and sufficiently complex that prior methods left some significant facets of the data underexploited.

Modeling modalities

The starting point for most patient modeling typically consists of the patient's record of diagnoses, pharmaceutical prescriptions, surgical procedures, and lab tests (hereafter summarized as medical entities or medical concepts). Some studies have also examined provider notes and demonstrated significant utility from this data, although it is often much less structured and may contain similar content to the semistructured diagnostic information provided by medical entities.[10] Beginning with techniques such as Med2Vec, it has become standard for modelers to employ entity embedding methods to construct dense representations of these entities. This practice allows models to learn more efficiently by sharing information about similar diagnoses.[21,22]

Frequently, physician decisions can be traced to a relatively small subset of a patient's health record. To approximate this intelligent filtering process, modelers can use a neural attention mechanism which operates similarly.[5–7,9,18,19] In such architectures, the model is trained to assign a weight to each embedding. Then, entities with nonzero weights are combined using a weighted averaging function to construct a dense representation of the patient's health state. This process is interpretable, since the relative sizes of entity weights indicate rough measures of the entities' relative importance in the model's decision. Furthermore, some attention mechanisms can be constructed so as to sparsify the weighting function, resulting in a smaller set of entities which could have contributed to the model's conclusion.[23]

The general construction of attention-based networks was most simply implemented in the natural language processing literature by Luong et al[24] for the purpose of machine translation. This construction can be adapted in the medical domain by substituting the occurrence of words for the occurrence of medical codes, leading to a simple and easily interpretable medical prediction.[18] While there are advantages to the simplicity of this model, there are also reasons to suspect that it is inefficient. This model cannot make use of complex relationships between diseases, observe the effects of disease progression over time, or adapt disease representation based on the context within a patient's medical record.

To capture the basic relationships between diseases, several models have incorporated hierarchical structures in which patient visits are aggregated in an intermediary step between entity embedding and patient summary.[9,14,17,19] Such models benefit by encoding logically related information into a structural element of the model. This can help to address questions related both to the relationships between entities and the complexities of capturing time information in the model. Visit-level vectors may be aggregated from individual entities by incorporating within-visit attention, simple average, or through more complex mechanisms.

In addition to aggregating visit data, several models have employed more sophisticated self-attention architectures to allow for complex relationships between entities.[9,14,17,19] Typically, self-attention mechanisms incorporate a pairwise matching approach, whereby the model learns to assign weights to the relationships between entities and then re-embed entities to reflect those relationships. Thus, models can learn the differences between, for example, treated and untreated versions of the same disease. This technique has proved extremely useful and interpretable in the natural language processing literature; since its introduction there, it has enjoyed great success in modeling patient data as well.[9,13,14,19,25]

In order to incorporate time as an element in patient modeling, researchers have taken two divergent approaches. Traditionally, models such as RETAIN have incorporated time using a recurrent neural network (RNN).[20] RNNs employ decay to steadily reduce the impact of older data and allow the proximity of observations to influence the strength of relationships between data. RNNs operate explicitly over sequences and can therefore ignore the distance between visits unless this data is provided separately.

Several authors have also proposed purely attentional methods to incorporate time information into models. Typically, these models have been inspired by the transformer model proposed by Vaswani et al[25] for the task of machine translation. This methodology adapts the Transformer's architecture to encode time data into the entity embedding structure and relies on the attentional structures to interpret this data.[9,19]

In addition to patient diagnoses, several studies have demonstrated that patient demographic features can be useful inputs.[10–13] Unfortunately, many leading deep learning models such as RETAIN, HiTANet, and ConCare have omitted these features as primary inputs when proposing new models.[9,19,20] Furthermore, several studies have examined the impact of improving medical concept embedding initialization by pretraining embeddings with an alternative

model.[14–16] Recently, Rasmy et al demonstrated in Med-BERT that such initialization can substantially improve even sophisticated modern architectures.[14]

Objective

In this article, we propose an efficient new transformer-based architecture for predicting patient outcomes from EMR data. Our model synthesizes improvements described by several authors and simplifies the architecture. Critically, we then examine the impact of each element of the model on overall performance and compare it against a leading alternative. Our model differs from prior works in several significant ways. We simplify the construction of the time-awareness by incorporating a trigonometric decomposition. We also flatten the hierarchical embedding structure used in previous works, relying instead on the time encoding to capture the relevant information. This adjustment is conceptually simple and easy to implement. We employ a sparsified global attention mechanism to maximize interpretability and incorporate both demographic data and transfer learning to optimize model performance.

METHODS

In this study, we propose a new adaptation of the Transformer model. This model implements a Trigonometrically encoded Time-aware Transformer Network (T^3Net). To assess the efficacy of T^3Net in patient prediction, we compare it against a leading recent alternative, HiTANet.[9] We tested these models on two targets (mortality and hospitalization) trained on EMR data from a large regional medical group. Models were trained using medical records from 2017 with targets in 2018 and validated using medical records from 2018 with targets occurring in 2019. Models were evaluated for their performance on average precision (AP) and area under the receiver-operator characteristic (AUROC).

Model architecture

T^3Net takes as its primary input a set of patient medical entities, including diagnoses, procedures, lab tests, and pharmaceutical codes. Each entity is converted to a numeric vector embedding. Code embeddings are then decomposed using a trigonometric decomposition and submitted to a transformer-style self-attention encoder, as Vaswani et al.[25] The re-embedded codes are then concatenated with their original embeddings and submitted to a traditional attention layer, as Luong et al.[24] This attention layer yields a single vector which we consider to be a numeric summary of the patient's known health state. This patient vector is concatenated with a demographic feature vector and submitted to a traditional feedforward neural classifier. For a more complete discussion of T^3Net's architecture, please refer to Supplementary Appendix A: Model Architecture.

Trigonometric time decomposition

Our model incorporates an original trigonometric time decomposition. Prior to self-attention re-embedding, each code is decomposed into two elements by multiplying the code by $\sin^2 \frac{t}{2\pi P}$ and by $\cos^2 \frac{t}{2\pi P}$, where P indicates the desired period (in our case, 365 days) and t indicates the time since code assignment. This decomposition can also be applied multiple times with a variety of periods to construct a more nuanced time encoding, although we do not incorporate multiple periods in this study. By decomposing codes in this way, we allow the model to perfectly reconstruct the original embedding while losslessly (with an appropriately selected period) encoding time data. In principle, this compares favorably with additive methods, where it may not be possible for the model to perfectly reconstruct either element from the available data.

Data

We employed data from patients in the Kaiser Permanente Mid-Atlantic States (KPMAS) medical system. KPMAS is an integrated medical system serving approximately 780 000 members in Maryland, Virginia, and the District of Columbia. We trained models on EMR records for 294 698 patients with active coverage on January 1st, 2018; then, we validated models on a set of 311 156 patients with active coverage on January 1st, 2019. Patients were included in each year of data if they were age 45 or older by the model date. We excluded younger patients from our sample based on preliminary analysis of target prevalence, which showed that mortality and hospitalization were both very rare among younger patients. In addition, patients without any recorded medical history (typically new patients) were excluded.

Each patient record included both demographic and EMR data. We modeled two targets for patient outcomes: 12-month mortality and 6-month hospitalization. We chose to incorporate a 6-month hospitalization target instead of 12 months because preliminary experimentation demonstrated that the hospitalization target became less predictable after 6 months. For the hospitalization model, we included all inpatient hospital admissions submitted as claims to the health system. For the 1-year mortality target, we identified death records based on operational records available in our EMR system.

Table 1 displays the rates of various demographic groups and outcome measures across the two years studied. We observed no significant departures between these two sets.

Our research was approved by the Kaiser Permanente Mid-Atlantic State Institutional Review Board (IRB).

Comparison model

We compare our model to the state-of-the-art HiTANet model.[9] This model incorporates similar architectural elements, particularly an alternative version of time-aware self-attention and a secondary attention model applied over code embeddings. In contrast, HiTANet incorporates a hierarchical visit-based embedding structure prior to applying self-attention. HiTANet also omits demographic data and transfer learning to initiate code embeddings.

Table 1. Patient demographic summary

Demographic/outcome group	2018 frequency (%)	2019 frequency (%)
Male	45.7	45.6
Female	54.3	54.4
Asian/Pacific Islander	12.8	13.5
Black/African American	38.7	38.2
Hispanic/LatinX	10.2	10.4
White	31.7	31.2
Unknown/other	6.6	6.6
1-year mortality	0.9	0.9
6-month hospitalization	3.8	3.7

This table describes a patient demographic breakdown for our training and validation data sets. Overall, demographic statistics were very similar across the two data sets.

Ablation studies

One of the primary drawbacks to employing HiTANet as a comparison is that T^{3}Net incorporates certain elements not present in HiTANet. In particular, HiTANet's published implementation does not offer a method for initializing embeddings through transfer learning or inputting demographic data. For this reason, we also examine the results of several ablation studies. In these studies, we omit either transfer learning, demographic inputs, or both. Note that models which did not incorporate transfer learning to initialize code embeddings were instantiated with random embeddings generated from an independent (by vector and dimension) uniform distribution between −0.05 and 0.05.

Model training

For each version of T^{3}Net, we employed a single attentional head (both self-attention and simple attention, where applicable) to maximize interpretability. We examined results for 16-dimensional embeddings. For the final classification task, we employed a deep network with an initial layer size of 256 and 4 layers deep. Models were trained with a dropout rate of 0.3 for layers in the primary classification network and the simple attentional network. Each layer in both networks was also regularized with a penalty of 0.001.

Entity embeddings were initialized using pretrained vectors obtained from a Word2Vec model applied to the same training data; however, entity embeddings were also allowed to co-train with the model. Entities were deduplicated prior to submission to either the embedding algorithm or to the core model. For models which incorporated time-sensitivity, we retained the most recent allocation of any individual code as the canonical diagnosis time. This yielded a final count of 8.3 million deduplicated diagnosis, procedure, pharmaceutical, and lab test codes in our training set and 8.9 million deduplicated codes in our validation set.

Models were implemented in Tensorflow version 2.3.0, primarily using the Keras functional interface. They were optimized using the Adam optimization algorithm for 20 epochs each. All models were trained on a locally hosted IBM CloudPak4Data instance with 2 cores and 64 GB of RAM.

To train the HiTANet model, we used the code published with the original article.[9] Models were trained in PyTorch version 1.3.1. Once again, we used a locally hosted IBM CloudPak4Data instance with 2 cores and 64 GB of RAM to train models. To alter our data set for use with HiTANet, we had to aggregate patient codes into visit lists. For each patient, we combined all codes which occurred on the same date and labeled those as a single visit.

Due to computational constraints, we were only able to train the HiTANet model once for each target. To remain consistent with settings for T^{3}Net, we trained the model with an embedding dimensionality of 16. We also employed a classification hidden layer size of 256, which was the same as T^{3}Net and which was also the default suggested by the authors. We used default values for all other hyperparameters. We trained each HiTANet model for 20 epochs and recorded performance at the end of each epoch.

Model comparison

For each model, we examined model performance using the AUROC and the average precision (AP) when using the model to predict our validation data. For each version of T^{3}Net, we performed model training 10 times independently and computed performance on our evaluation metrics after each epoch. We then examined the median performance of the given set of models over all epochs and selected the epoch with the best median AP as our representative for that class of models. We employed a median-based strategy to minimize the impacts of individual runs of each model, which could sometimes produce highly variant results due to poor random starting conditions. Although we took the best epoch of each model as our canonical result, results were robust to our selection of epoch.

RESULTS

Here, we examine the performance of T^{3}Net in comparison to HiTANet on our mortality and hospitalization targets. In Supplementary Appendix B: Attention Analysis, we analyze the interpretability of T^{3}Net by observing the network's prediction and attentional responses to a synthetic patient profile. In Supplementary Appendix C: Architectural Ablation Studies, we present a detailed breakdown of results from ablating specific architectural elements of T^{3}Net.

Convergence and overfitting

We found that the models which excluded Word2Vec pretraining tended to converge very quickly (optimal epochs were 2–3), but that these models quickly began overfitting (Figures 1–4). This was a tendency which was not evidenced by the full T^{3}Net model or by the model which ablated only demographics. These observations held true for both the mortality target and the hospitalization target.

In both experiments, HiTANet's optimal epoch was quite late[19]; however, HiTANet demonstrated an asymptotic convergence and was no longer improving significantly by the final stopping point. On the mortality target, HiTANet did not appear to substantially improve after the fourth epoch, although it continued improving until about the 10th epoch on the hospitalization target.

One-year mortality prediction

T^{3}Net significantly outperformed the HiTANet benchmark on the 1-year mortality target (Figures 1 and 2). In its optimal epoch, median performance from T^{3}Net achieved an AUROC of 91.96% and an AP of 20.35% with the full model. With the fully ablated model (which included neither transfer learning nor demographic data), T^{3}Net achieved an AUROC of 90.17% and an AP of 19.81%. This compares favorably with results achieved by HiTANet, which achieved a maximum AUROC of 73.82% and an AP of 3.90%. As demonstrated by Table 2, the median performance of each version of T^{3}Net outperformed HiTANet on both metrics. Interestingly, ablating only demographic features severely limited model performance and yielded the worst performance over all T^{3}Net models with an optimal median AUROC of 89.01% and AP of 15.89%.

Six-month hospitalization prediction

Median performance by the full T^{3}Net model outperformed HiTANet on both metrics using the hospitalization target, even after ablating both demographics and transfer learning (Figures 3 and 4). The full T^{3}Net model achieved an AUROC of 82.41% and an AP of 23.80%, with the fully ablated model achieving an AUROC of 81.72% and an AP of 23.16%. HiTANet's performance was much closer to our own models' on this target, achieving an optimal AUROC of 80.16% and AP of 21.85%. HiTANet outperformed the model which ablated only demographics data, which achieved an optimal AUROC of 80.33% and AP of 20.80%.

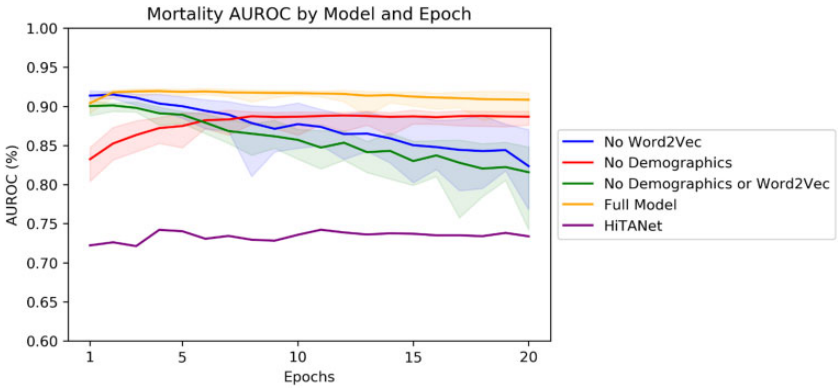

Figure 1. Mortality AUROC by epoch. This figure demonstrates the minimum, median, and maximum area under the receiver-operator characteristic (AUROC) performance of each model on the mortality target after each epoch. Model performance peaked early for most model categories; however, our model which omitted Demographics took several more epochs to converge. Furthermore, the models without Word2Vec pretraining displayed a tendency to overfit after only a few epochs.

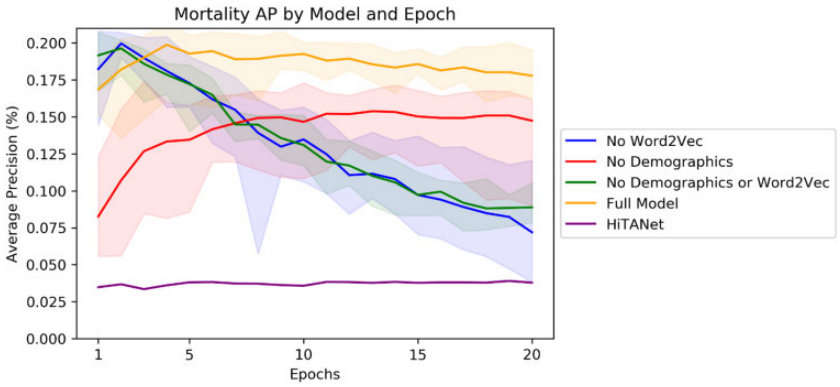

Figure 2. Mortality AP by epoch. This figure demonstrates the minimum, median, and maximum average precision (AP) performance of each model on the mortality target after each epoch. Model performance peaked early for most model categories; however, our model which omitted Demographics took several more epochs to converge. Furthermore, the models without Word2Vec pretraining displayed a tendency to overfit after only a few epochs.

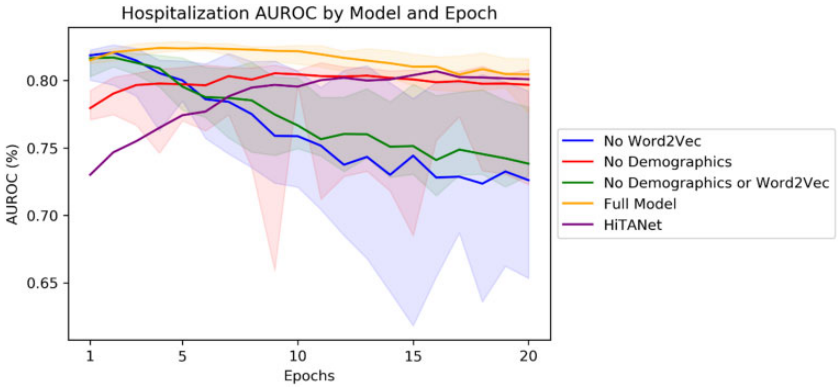

Figure 3. Hospitalization AUROC by epoch. This figure demonstrates the minimum, median, and maximum area under the receiver-operator characteristic (AUROC) performance of each model on the hospitalization target after each epoch. Model performance peaked early for most model categories; however, the HiTANet model took several epochs to converge. Furthermore, the models without Word2Vec pretraining displayed a tendency to overfit after only a few epochs.

Runtime

As demonstrated by Table 2, T^{3}Net achieved significantly shorter runtimes than HiTANet in our experiments. On training tasks, the full T^{3}Net model averaged approximately 640 seconds per epoch, compared to approximately 18 000 seconds per epoch for HiTANet. Over the entire training cycle, we found that HiTANet took 55 633 s (15.45 h) on average to train, store model weights, predict the validation set, and perform other internal maintenance for each epoch using a slightly altered version of the authors' provided training scripts. In comparison, we observed an average runtime of 13,101 seconds per complete run of T^{3}Net (including 20 epochs of training, forecasting at each epoch, and model saving).

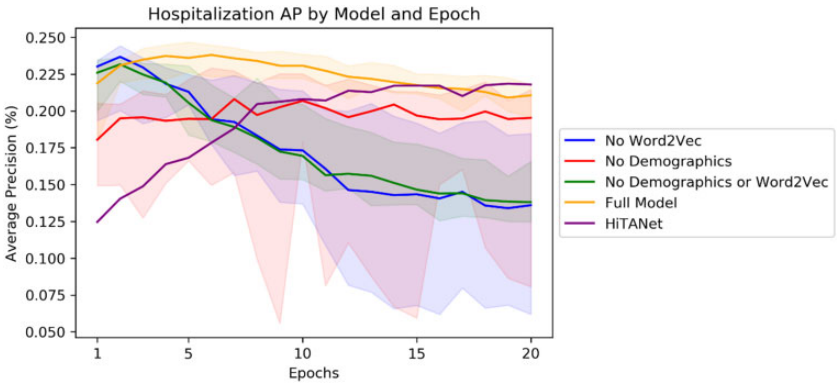

Figure 4. Hospitalization AP by epoch. This figure demonstrates the minimum, median, and maximum average precision (AP) performance of each model on the hospitalization target after each epoch. Model performance peaked early for most model categories; however, the HiTANet model took several epochs to converge. Furthermore, the models without Word2Vec pretraining displayed a tendency to overfit after only a few epochs.

Table 2. Results summary by model

Model name	Epoch (M)	AUROC (M)	AP (M)	Epoch (H)	AUROC (H)	AP (H)
T^3Net (FULL MODEL)	4	**91.96%** (**91.61%, 92.26%**)	**20.35%** (**16.15%, 21.50%**)	6	**82.41%** (**82.24%, 82.74%**)	**23.80%** (**22.31%, 24.46%**)
T^3Net (Ablate demographics)	11	89.01% (87.67%, 89.26%)	15.89% (11.38%, 16.64%)	7	80.33% (77.47%, 81.00%)	20.80% (15.81%, 22.71%)
T^3Net (Ablate Word2vEC)	2	91.59% (90.99%, 91.89%)	20.12% (19.02%, 20.71%)	2	82.09% (79.68%, 82.64%)	23.67% (20.03%, 24.42%)
T^3Net (Ablate Word2vEC and demographics)	2	90.17% (89.37%, 90.44%)	19.81% (17.82%, 20.19%)	2	81.72% (81.02%, 82.09%)	23.16% (22.01%, 23.61%)
HiTANET	19	73.82%	3.90%	19	80.16%	21.85%

This table displays the value of the median score achieved by each model in its best-performing epoch by average precision (AP). The table also indicates performance on the area under the receiver-operator characteristic curve (AUROC). Columns marked with an (M) display values for model performance on the mortality target; columns marked with an (H) display values for model performance on the hospitalization target. Each column also indicates the best and worst scores for models on the given metric at the indicated epoch. Note that the best performance in both cases (as indicated by bold font) is achieved by the full T^3Net model, although the most appropriate comparison network is the fully ablated model.

DISCUSSION

Contributions

In this study, we have proposed a novel implementation of the Transformer architecture which incorporates Trigonometrically embedded Time data, named T^3Net. Our model incorporates several innovations over current state-of-the-art patient EMR models, including our unique implementation of time encoding. Furthermore, our model incorporates transfer learning and demographic data; while other studies have demonstrated the value of these features, they have been underutilized in the deep learning EMR literature, especially in Transformer models. When applied to a real-world data set, our model outperformed a leading alternative in the space.

In addition, we performed a comprehensive series of ablation studies. Interestingly, we found that a self-attentional Transformer model significantly underperformed (Supplementary Appendix C: Architectural Ablation Studies) when time encoding was omitted, lending considerable weight to the observations by Luo et al which indicated that time data was critical to the performance of EMR models.

In additional ablation studies, we found that including demographic features significantly improved model performance when keeping all other features constant and that using transfer learning to initialize medical entity embeddings had a significant impact on the model's ability to overfit. Models with pretrained embeddings tended to both outperform and to be more robust to overfitting (Figures 1–4).

Comparison to HiTANet

We compared our model construction to the recently published alternative HiTANet. HiTANet employs several important architectural features similar toT^3Net, including the use of time-aware self-attention; however, there are important distinctions between these models that led to substantial differences in performance. To account for these differences, we examined the performance of several ablated versions of T^3Net.

We found substantial differences in performance between the ablated versions of T^3Net and HiTANet. Our model significantly outperformed HiTANet on the mortality target and marginally outperformed HiTANet on the hospitalization target. By construction, these differences must be attributable to differences in the mathematical construction of our network and HiTANet.

Model interpretation

We found that it was possible to construct a reliable total attention weight which incorporates the impacts of both global attention and self-attention (Supplementary Appendix B: Attention Analysis). Our

construction of total attention was strongly correlated with the absolute difference in risk score obtained by adding or removing the given diagnosis, demonstrating that this measure is a reliable tool for interpreting T^3Net's decisions. Finally, we observed that T^3Net's decisions reflected clinical intuition. For example, it learned that time information was significantly more important when incorporating the effects of acute codes than when examining long-term chronic diagnoses.

Limitations

In this article, we present the results from applying several complex models to a single dataset. Further study will be necessary to verify that our results are generalizable to alternative datasets. Unfortunately, there is a notable lack of large, publicly available Electronic Health Record data that focus on long-term, chronic conditions. The most popular research dataset, MIMIC-III, focuses on critical care patients; we do not believe that our methodology is well-optimized for this type of application since it focuses on code occurrences over long periods of time (particularly chronic codes).[26]

The data set we have employed may not have been perfectly accurate. In a health system as large as ours, it is inevitable that some patients will not be assigned diagnoses correctly or that demographic data will be recorded incorrectly. Furthermore, our definition of mortality is based on an operational definition which our health system uses in practice; however, it is possible that some mortality events are not recorded through this operational system. Similarly, it is possible that some patients were hospitalized without submitting claims. In any of these cases, we note that these omissions would be likely to degrade model performance.

To limit the scope of computational resources required, we have employed only models of moderate size. Our experiments did not explore the impacts of incorporating additional attentional heads or compare our results with other modern architectures. We note, however, that these are practical concerns present in most healthcare organizations that may seek to deploy patient outcome modeling. Furthermore, HiTANet has compared well with other leading recent alternatives.[9] All models presented here were able to run on a virtual machine with only 2 cores and 64 GB of RAM.

Our experimentation with HiTANet was limited by model training times. We found that HiTANet epochs took approximately 10× longer to train than our own largest models, with similarly long times required to produce validation predictions. We suspect that this difference was due in large part to our computational infrastructure. HiTANet was implemented and optimized by the original authors for use in a GPU-based environment. Unfortunately, we did not have access to a robust GPU training environment and were therefore unable to make use of these optimizations.

Future directions

This article has indicated several interesting new directions for the future of patient modeling. Our work suggests that incorporating demographic features and transfer learning into other model architectures could improve the performance of those models. Further research will be required to determine how embedding strategy can influence the performance of various modeling architectures, particularly when embedding weights are initialized by training on alternative data sets.[14,27]

We also proposed a new method to encode time data into models. Our results indicate that encoding based on a trigonometric decomposition can drastically improve model performance, although further investigation is required to more thoroughly determine the relative efficiency of the various ways that this data can be encoded.

Finally, we found that incorporating advanced re-embedding structures such as self-attention can complicate model interpretation. The influence of such structures will naturally vary by model architecture; however, our results strongly indicate that it is useful to confirm a researcher's intuition and intentions by correlating attention results with a perturbation study.

CONCLUSION

In this article, we have examined the effectiveness of several attentional models on patient prediction tasks using real EMRs data. Our results indicate that attentional networks can produce strong models of patient outcomes with relatively small computational requirements. Furthermore, our unique adaptation of Vaswani et al's Transformer model (T^3Net) proved to be superior to a leading literature alternative on the given tasks. Finally, our findings have indicated several important considerations to be considered by future models in the space, including the importance of pretraining medical concept embeddings, demographic features, incorporating time data into model architecture, and using the full contribution of a code to the model (total weight) instead of its simple attention weight (global weight) when interpreting a model with self-attentional elements.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

CONTRIBUTORS

Anthony Finch proposed the work, performed the core data analysis, implemented the core neural network models, and supervised other technical members of the team. Alexander Crowell built the original model of patient mortality and accompanying dataset, helped to design several of the core model architectures, and assisted in developing the model code. Yung-Chieh Chang built the original model of patient hospital admissions and the accompanying target data and assisted in developing the model code. Pooja Parameshwarappa contributed significantly to the literature review and assisted in developing the model code. Jose Martinez managed and supported several of the personnel that participated in this project. Michael Horberg directly supervised this project and contributed significantly to the research plan.

Conflict of interest statement. None declared.

DATA AVAILABILITY

All data underlying this article will be shared on reasonable request to the corresponding author.

ACKNOWLEDGMENTS

We would like to acknowledge Richard McCarthy, MD, The Permanente Medical Group Associate Executive Director for the Mid-Atlantic States, for his support and critical input.

REFERENCES

1. Shrestha A, Mahmood A. Review of deep learning algorithms and architectures. *IEEE Access* 2019; 7: 53040–65.
2. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015; 521 (7553): 436–44.
3. Osmani V, Li L, Danieletto M, *et al.* Processing of electronic health records using deep learning: a review. ArXiv180401758 Cs. Published Online First: 5 April 2018; http://arxiv.org/abs/1804.01758. Accessed January 27, 2021.
4. Miotto R, Li L, Kidd BA, *et al.* Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Sci Rep* 2016; 6: 26094.
5. Sha Y, Wang MD. Interpretable predictions of clinical outcomes with an attention-based recurrent neural network. *ACM BCB* 2017; 2017: 233–40.
6. Kaji DA, Zech JR, Kim JS, *et al.* An attention based deep learning model of clinical events in the intensive care unit. *PLoS One* 2019; 14 (2): e0211057.
7. Kim YJ, Lee Y-G, Kim JW, *et al.* Highrisk prediction from electronic medical records via deep attention networks. ArXiv171200010 Cs Stat. Published Online First: 30 November 2017; http://arxiv.org/abs/1712.00010. Accessed January 27, 2021.
8. Zhang J, Kowsari K, Harrison JH, *et al.* Patient2Vec: a personalized interpretable deep representation of the longitudinal electronic health record. *IEEE Access* 2018; 6: 65333–46.
9. Luo J, Ye M, Xiao C, *et al.* HiTANet: hierarchical time-aware attention networks for risk prediction on electronic health records. In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. New York, NY, USA: Association for Computing Machinery 2020: 647–56.
10. Zhang D, Yin C, Zeng J, *et al.* Combining structured and unstructured data for predictive models: a deep learning approach. *BMC Med Inform Decis Mak* 2020; 20 (1): 280.
11. Liu J, Zhang Z, Razavian N. Deep EHR: chronic disease prediction using medical notes. arXiv:180804928. https://arxiv.org/abs/1808.04928
12. Beaulieu-Jones BK, Yuan W, Brat GA, *et al.* Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians? *NPJ Digit Med* 2021; 4 (1): 62.
13. Badgeley MA, Zech JR, Oakden-Rayner L, *et al.* Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ Digit Med* 2019; 2: 31.
14. Rasmy L, Xiang Y, Xie Z, *et al.* Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digit Med* 2021; 4: 1–13.
15. Shang J, Ma T, Xiao C, *et al.* Pre-training of graph augmented transformers for medication recommendation. ArXiv190600346 Cs. Published Online First: 26 November 2019; http://arxiv.org/abs/1906.00346 Accessed June 22, 2021.
16. Xiang Y, Xu J, Si Y, *et al.* Time-sensitive clinical concept embeddings learned from large electronic health records. *BMC Med Inform Decis Mak* 2019; 19 (Suppl 2): 58.
17. Li Y, Rao S, Solares JRA, *et al.* BEHRT: transformer for electronic health records. *Sci Rep* 2020; 10 (1): 7155.
18. Chen P, Dong W, Wang J, *et al.* Interpretable clinical prediction via attention-based neural network. *BMC Med Inform Decis Mak* 2020; 20 (Suppl 3): 131.
19. Ma L, Zhang C, Wang Y, *et al.* ConCare: personalized clinical feature embedding via capturing the healthcare context. ArXiv191112216 Cs Stat. Published Online First: 27 November 2019; http://arxiv.org/abs/1911.12216 Accessed June 4, 2021.
20. Choi E, Bahadori MT, Kulas JA, *et al.* RETAIN: an interpretable predictive model for healthcare using reverse time attention mechanism. ArXiv160805745 Cs. Published Online First: 26 February 2017; http://arxiv.org/abs/1608.05745. Accessed June 22, 2021.
21. Choi E, Bahadori MT, Searles E, *et al.* Multi-layer representation learning for medical concepts. ArXiv160205568 Cs. Published Online First: 17 February 2016; http://arxiv.org/abs/1602.05568. Accessed January 27, 2021.
22. Mikolov T, Chen K, Corrado G, *et al.* Efficient estimation of word representations in vector space. ArXiv13013781 Cs. Published Online First: 6 September 2013; http://arxiv.org/abs/1301.3781. Accessed January 27, 2021.
23. Martins AFT, Astudillo RF. From softmax to sparsemax: a sparse model of attention and multi-label classification. ArXiv160202068 Cs Stat. Published Online First: 8 February 2016; http://arxiv.org/abs/1602.02068. Accessed January 27, 2021.
24. Luong M-T, Pham H, Manning CD. Effective approaches to attention-based neural machine translation. ArXiv150804025 Cs. Published Online First: 20 September 2015; http://arxiv.org/abs/1508.04025. Accessed January 27, 2021.
25. Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need. ArXiv170603762 Cs. Published Online First: 5 December 2017; http://arxiv.org/abs/1706.03762. Accessed January 27, 2021.
26. MIMIC. https://mimic.physionet.org/about/mimic/. Accessed January 27, 2021.
*27. Finch A, Crowell A, Bhatia M, *et al.* Exploiting hierarchy in medical concept embedding. *JAMIA Open* 2021; 4 (1): doi:10.1093/jamiaopen/ooab022.