including medication, procedure, test, BP measurement, BMI, drinking status, and smoking status. Since they provided additional information for modelling, as expected, the Hi-BEHRT model achieved better performance with the inclusion of more modalities. Therefore, our work highlighted the great benefit of including rich medical history for accurate risk prediction. However, we further observed that the contribution of a modality to the model performance is in general highly related to its frequency of recording in the dataset. It is partially due to the fact that higher frequencies of recording can provide more information for prediction. Another possible explanation is that the information provided by modalities with low frequency are overpowered by the modalities with much higher frequency. This phenomenon has been described in the natural language processing literature for the embedding of words as well which tend to be biased towards higher word frequencies [30]. Therefore, one potential future work will be to investigate how to incorporate features or modalities with low frequency in a more meaningful way for risk prediction.

Furthermore, to better understand the advantage of including patients’ complete medical history for modelling, we compared Hi-BEHRT to the benchmark models in terms of handling patients with different learning periods. We conducted a subgroup analysis to evaluate model performance on patients who have EHR length within the capacity of Transformer models (i.e., less than or equal to 256 in our study) and longer than the Transformer models’ capacity (i.e., more than 256) in the learning period. We found that the Hi-BEHRT model showed similar or better performance than the benchmark models on risk prediction tasks for patients within the relatively short EHR length group, but it greatly improved model performance with the inclusion of more records (i.e., > 256). However, due to the limitation of sequence length in the benchmark Transformer models, the difference of model performance between patients with long EHR and short EHR is relatively small. Additionally, we notice for very imbalanced outcomes, for example HF, the majority of the positive cases occur in patients with longer EHR and these patients can have different contextual patterns compared to the patients with short EHR. By making risk prediction with only a fraction of the latest records in the benchmark Transformer models when patients have long EHR sequence, they treated patients with long EHR the same way as making prediction for patient with short EHR sequence. Therefore, the models, which rely on the global attention, can be driven to have better discrimination performance for the positive cases with long EHR records in the training and have relatively poor capability of identifying positive cases with short EHR sequence. On the contrary, with the inclusion of the entire EHR, together with local feature extractor and global feature aggregator to identify temporal and global patterns, the Hi-BEHRT model is more capable of distinguishing different patterns of positive cases in both long and short EHR sequences. Considering the majority (70%) of the population have relatively short EHR length (less than 256) in our risk prediction tasks and probably in most of the cases in reality, this can be an important additional feature of our proposed model.

In addition to model architecture, we also evaluated the usability of a contrastive learning pre-training strategy, BYOL, in this work. We combined the framework, which was originally designed for image pre-training, with the MLM task, and adapted it to pre-train our sequential model. With the pre-training, the Hi-BEHRT model can achieve similar performance using only 1% of training data as the model trained without pre-training using 5% of training data. With additional ablation analysis, we concluded that the pre-training can potentially expanded the power-law region [29] and allowed the model to reach power-law region with smaller data size. However, our results also indicated that the model performance almost saturated when using 50% of training dataset. It means the model achieves the irreducible error region. Future work should investigate more robust model architectures to shift the power-law curve and improve the model accuracy.

One of the major contributions of this work is the provision of a framework for risk prediction with the inclusion of long and comprehensive EHR. With the growing accessibility and usability of EHR systems, risk prediction using long EHR can be inevitable and have important implications for medical practice. To our best knowledge, long sequence modelling and its application in the context of healthcare and EHR remains unexplored. Our work proposed a potential solution to tackle this problem and investigated its benefit comparing to model that makes prediction using only a fraction of the EHR. Moreover, we provided a self-supervised pre-training framework for the proposed model, and pre-training can adapt risk prediction model to handle tasks with less training data available, which is highly desired in most of the scenarios. We also encourage future work to further explore other long sequence modelling strategies (e.g., Longformer [11]) for EHR modelling.

Our study also has limitations. First, we focused on the risk of HF, diabetes, CKD, and stroke. As such, the conclusion may not generalize to other diseases. Additionally, our work relied on internal validation and the model performance under data shifts or in the external cohorts requires further investigation.

**APPENDIX**

**A. Additional Information on Dataset**

In this section, we provide more information on modalities that are not commonly included in the modelling. More specifically, we will introduce procedure and test.

1) **Procedure:** Procedure is CPRD linked data collected from Hospital Episode Statistics (HES) Admitted Patient Care (EHS APC) data. It is recorded at the point of admission to, or attendances at NHS healthcare providers. All procedure information is coded using the U.K. Office of Population, Censuses and Surveys classification (OPCS) 4.6, and procedures that are not covered by OPCS code is not included in the system. Each record in the system is specified with a start and an end date, as well as event date. We used OPCS code and event date to structure the timeline of a patient’s EHR history for modelling.
2) **Test:** Test is recorded in the CPRD test table and coded as Read code. It includes information on history/symptoms, examination/signs, diagnostic procedures, and laboratory procedures. In the experiment, we only used the information in the Read code level, which represents what examinations or procedures are carried out. More detailed quantitative information was excluded.

**B. Clinical Codes for HF, Diabetes, CKD, and Stroke**

**TABLE V**

| ICD Code | Description |
|----------|-------------|
| I09.9    | Rheumatic heart failure |
| I11.0    | Hypertensive heart disease with (congestive) heart failure |
| I13.2    | Hypertensive heart and renal disease with (congestive) heart failure |
| I25.5    | Ischemic cardiomyopathy |
| I27.9    | Chronic cor pulmonale |
| I38      | Congestive heart failure due to valvular disease |
| I42.1    | Obstructive hypertrophic cardiomyopathy |
| I42.2    | Nonobstructive hypertrophic cardiomyopathy |
| I42.6    | Alcoholic cardiomyopathy |
| I42.8    | Other cardiomyopathies |
| I42.9    | Cardiomyopathy NOS |
| I50.0    | Congestive heart failure |
| I50.1    | Left ventricular failure |
| I50.2    | Systolic (congestive) heart failure |
| I50.3    | Diastolic (congestive) heart failure |
| I50.8    | Other heart failure |
| I50.9    | Cardiac, heart or myocardial failure NOS |
| I38      | is mapped from Read code G580400 |

**TABLE VI**

| ICD Code | Description |
|----------|-------------|
| E10      | Type 1 diabetes mellitus |
| E11      | Type 2 diabetes mellitus |
| E12      | Malnutrition-related diabetes mellitus |
| E13      | Other specified diabetes mellitus |
| E14      | Unspecified diabetes mellitus |
| O24.2    | Pre-existing malnutrition-related diabetes mellitus |

**C. Model Evaluation Stratified By Baseline Age**

We evaluated model performance stratified by the baseline age. The comparison was conducted on three subgroups of patients: 1) patients with baseline age between 35 and 50 years old (young adult); 2) patients with baseline age between 50 and 70 years old (middle-aged adult), and 3) patients with baseline age 70–90 years old (older adult). Table IX shows that the hierarchical BEHRT model has better performance across all subgroups, and it substantially outperforms for BEHRT model on HF and diabetes risk prediction tasks, especially for patients with younger age.

**D. Size and Overlap of Sliding Window**

For Hi-BEHRT model, we used sliding window to segment the raw EHR into segments. As shown in Table X when window size is relatively small (i.e., 50), the size of the stride does not have significant impact in terms of predictive performance, and the bigger stride size can potentially decrease the number of segments and reduce model complexity. However, for the larger window size (i.e., 100), the stride size becomes more important, and some level of overlap between segments is necessary. Without any overlap for window size 100, the AUPRC decreases 4% comparing to the model with stride size 50. Additionally, the analysis shows that not larger window size always the better choice. For instance, AUPRC of window size 100 without overlap decreases 2% comparing to AURPC of window size 50 without overlap. Without overlap, larger window can lead to shorter length in the segment level, and a balance between window size and length of segment might be more preferred in the hierarchical structure.
TABLE IX
BASELINE AGE STRATIFIED SUBGROUP ANALYSIS

| Sample size | No. (%) of positive cases | Baseline age | AUR | AUP | AUR | AUP | AUR | AUP |
|-------------|---------------------------|--------------|-----|-----|-----|-----|-----|-----|
| HF          | 154,032                   | 1,008 (0.7)  | 35-50| 0.84| 0.40| 0.90| 0.56|     |
|             | 180,416                   | 6,878 (3.8)  | 50-70| 0.88| 0.64| 0.93| 0.72|     |
|             | 111,044                   | 17,670 (15.9)| 70-90| 0.86| 0.75| 0.90| 0.80|     |
| Diabetes    | 149,308                   | 4,764 (3.1)  | 35-50| 0.87| 0.69| 0.92| 0.69|     |
|             | 167,753                   | 12,443 (7.4) | 50-70| 0.87| 0.69| 0.91| 0.76|     |
|             | 103,866                   | 7,932 (7.6)  | 70-90| 0.89| 0.69| 0.90| 0.75|     |
| CKD         | 145,889                   | 4,343 (3.0)  | 35-50| 0.88| 0.62| 0.89| 0.64|     |
|             | 176,422                   | 13,037 (7.4) | 50-70| 0.90| 0.74| 0.92| 0.76|     |
| Stroke      | 111,727                   | 24,875 (22.3)| 70-90| 0.89| 0.83| 0.91| 0.84|     |
|             | 136,090                   | 11,325 (8.3) | 35-50| 0.88| 0.70| 0.88| 0.71|     |
|             | 157,789                   | 21,392 (13.6)| 50-70| 0.88| 0.76| 0.90| 0.79|     |
|             | 93,159                    | 22,793 (24.5)| 70-90| 0.87| 0.82| 0.89| 0.84|     |

TABLE X
PERFORMANCE OF HF RISK PREDICTION WITH DIFFERENT WINDOW AND STRIDE SIZE

| Window size | Stride size | AUROC | AUPRC |
|-------------|-------------|-------|-------|
| 50          | 50          | 0.96  | 0.77  |
| 50          | 50          | 0.95  | 0.76  |
| 100         | 50          | 0.96  | 0.78  |
| 100         | 100         | 0.95  | 0.74  |
| 150         | 150         | 0.95  | 0.74  |

TABLE XI
HI-BEHRT HYPER-PARAMETER TUNING

| Hidden size | Intermediate size | AUROC | AUPRC |
|-------------|-------------------|-------|-------|
| 150         | 108               | 0.96  | 0.77  |
| 90          | 108               | 0.95  | 0.74  |
| 240         | 108               | 0.96  | 0.77  |
| 150         | 256               | 0.96  | 0.77  |

E. Hyper-Parameter Tuning

We set up hierarchical BEHRT with similar hyper-parameters as the BEHRT model and used it as a reference model to tune the hidden size and intermediate size of the Transformer. More specifically, we applied grid search for hidden size among [90, 150, 240] and intermediate size among [108, 256]. All experiments were conducted on the 5-year HF risk prediction task. Table XI shows that hidden size 150 and intermediate size 108 can achieve similar performance as the model with larger size.

F. Evaluation for Multiple Levels of Hierarchy

In this section, we investigated how the number of levels of hierarchy in Hi-BEHRT can influence the model performance in risk prediction. Specifically, we compared the performance of Hi-BEHRT with two and three levels of hierarchy. This is because each additional level can substantially reduce the sequence length. For instance, a sequence with maximum length 1225 would reduce to sequence length 118 with window size 50 and stride size 10 after the first level of hierarchy and would further reduce to 7 after the second level of hierarchy. Therefore, our dataset limited the number of levels we can investigate, and it would not make sense to investigate Hi-BEHRT with more than three levels of hierarchy. We encourage future work to replicate our work to more comprehensively investigate Hi-BEHRT with more levels of hierarchy. In our experiment, we only modified the feature extractor and kept the total number of layers in feature extractor the same for both comparators. More specifically, the two-level Hi-BEHRT had one level of hierarchy with four layers of Transformer for the extractor while the three-level Hi-BEHRT included two levels of hierarchy with a two-layer Transformer for each hierarchy. Both comparators used window size 50 and stride size 10 and the rest parameters were the same as reported in the manuscript. The results show that both models achieved AUROC 0.96 and AUPRC 0.76 for HF risk prediction, and there is no material difference between two-level and three-level Hi-BEHRT in our dataset.