

Deep Learning for Electronic Health Records: A Comparative Review of Multiple Deep Neural Architectures

Jose Roberto Ayala Solares^{a,b}, Francesca Elisa Diletta Raimondi^a, Yajie Zhu^{a,*}, Fatemeh Rahimian^a, Dexter Canoy^{a,b,c}, Jenny Tran^a, Ana Catarina Pinho Gomes^a, Amir Payberah^a, Mariagrazia Zottoli^a, Milad Nazarzadeh^{a,d}, Nathalie Conrad^a, Kazem Rahimi^{a,b}, Gholamreza Salimi-Khorshidi^a

^a*The George Institute for Global Health (UK), University of Oxford, United Kingdom*

^b*NIHR Oxford Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, United Kingdom*

^c*Faculty of Medicine, University of New South Wales, Sydney, Australia*

^d*Collaboration Center of Meta-Analysis Research, Torbat Heydariyeh University of Medical Sciences, Torbat Heydariyeh, Iran*

Abstract

Despite the recent developments in deep learning models, their applications in clinical decision-support systems have been very limited. Recent digitalisation of health records, however, has provided a great platform for the assessment of the usability of such techniques in healthcare. As a result, the field is starting to see a growing number of research papers that employ deep learning on electronic health records (EHR) for personalised prediction of risks and health trajectories. While this can be a promising trend, vast paper-to-paper variability (from data sources and models they use to the clinical questions they attempt to answer) have hampered the field's ability to simply compare and contrast such models for a given application of interest.

*Corresponding Author

Thus, in this paper, we aim to provide a comparative review of the key deep learning architectures that have been applied to EHR data. Furthermore, we also aim to: (1) introduce and use one of the world’s largest and most complex linked primary care EHR datasets (i.e., Clinical Practice Research Datalink, or CPRD) as a new asset for training such data-hungry models; (2) provide a guideline for working with EHR data for deep learning; (3) share some of the best practices for assessing the “goodness” of deep-learning models in clinical risk prediction; (4) and propose future research ideas for making deep learning models more suitable for the EHR data. Our results highlight the difficulties of working with highly imbalanced datasets, and show that sequential deep learning architectures such as RNN may be more suitable to deal with the temporal nature of EHR.

Keywords: Deep learning, Representation learning, Neural networks, Electronic health records, CPRD

1. Introduction

Electronic health records (EHR) systems store data associated with each individual’s health journey (including demographic information, diagnoses, medications, laboratory tests and results, medical images, clinical notes, and more) [1]. While the primary use of EHR was to improve the efficiency and ease of access of health systems [2], it has found a lot of applications in clinical informatics and epidemiology [3, 4, 5]. In particular, EHR have been used for medical concept extraction [6, 7], disease and patient clustering [8, 9], patient trajectory modelling [10], disease prediction [11, 12], and data-driven clinical decision support [13, 14], to name a few.

11 The early analyses of EHR relied on simpler and more traditional sta-
 12 tistical techniques [15]. More recently, however, statistical machine learning
 13 techniques such as logistic regression [16], support vector machines (SVM)
 14 [17], Cox proportional hazard model [18], and random forest [19] have also
 15 been employed for mining reliable predictive patterns in EHR data. While
 16 the simplicity and interpretability of such statistical models are desirable for
 17 medical applications, their weakness in dealing with high-dimensional input,
 18 their reliance on many assumptions, both statistical and structural, and their
 19 need for hand-crafted features/markers (guided by a domain expertise) make
 20 their use for comprehensive analyses of EHR data impractical [20, 21, 22]. To
 21 alleviate these issues, one needs to analyse each individual’s entire medical
 22 history (i.e., a series mixed-type and multimodal data packed in irregular
 23 intervals) [22], using modelling techniques that can discover and take into
 24 account complex nonlinear interactions among variables [23, 24].

25 The past decade has witnessed the profound impact of DL in a broad
 26 range of applications, including but not limited to its striking performance
 27 in natural language processing [25], speech recognition [26], computer vision
 28 [27], and reinforcement learning [28]. As a result, in recent years, the field of
 29 health informatics has seen a growth in the use of DL for a broad range of
 30 tasks [29]. For instance, DL has been extensively applied in the analysis of
 31 medical images [30], with Google’s work on detection of diabetic retinopathy
 32 [31], and detection of cancer metastases on gigapixel pathology images [32],
 33 being the most widely known examples. In genomics, DL has helped iden-
 34 tify common genetic variations more accurately than conventional methods
 35 [33, 34], opening the doors to new directions in genomics research [35, 36]

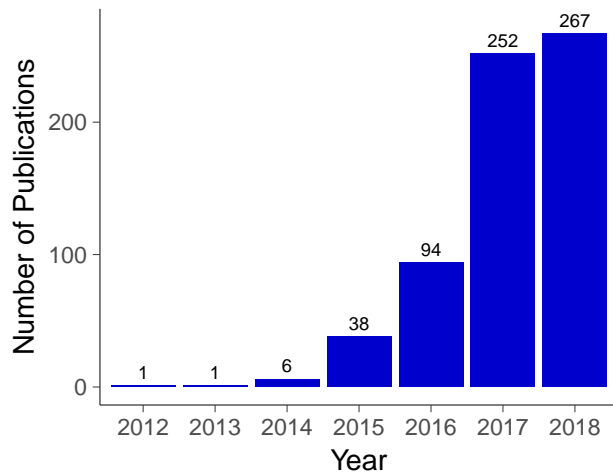


Figure 1: Published studies found on Semantic Scholar (<https://www.semanticscholar.org/>) through October 2018 using keywords “deep learning” AND (“electronic health records” OR “electronic medical records”) in the title or abstract.

and drug discovery for personalised medicine [37]. In all these applications, DL algorithms analysed big datasets to produce a compact general-purpose set of features that can be used for personalised risk prediction, treatment recommendations, and clinical trial recruitment [38, 39].

The number of research projects using DL for the analysis of EHR has been growing rapidly in the past few years (see Fig. 1). A quick search will show that the key DL architectures (e.g., feed-forward neural networks (FFNN), convolutional neural networks (CNN), and recurrent neural networks (RNN)), have been employed for the analysis and modelling of EHR. In one of the earliest such works, Tran et al. [40] introduced eNRBM (electronic medical records-driven nonnegative restricted Boltzmann machines), for learning a universal representation from complete EHR data (i.e., an

automatic feature extraction). eNRBM introduced constraints for nonnegative coefficients (for explainability of the learned parameters) and structural smoothness (to account for disease hierarchy in classification coding schemes) to standard Bernoulli RBM; the resulting model showed a superior performance for suicide risk prediction, when compared with “manual feature engineering”. Furthermore, eNRBM was shown to be capable of providing clinically meaningful clustering of patients and diseases. Similarly, Miotto et al. [38] chose deep stacked denoising autoencoder (SDA) for training a universal feature extractor, and showed that it outperforms the expert-driven feature engineering in a variety of clinical risk-prediction tasks, including diabetes mellitus with complications, cancer of rectum and anus, cancer of liver and intrahepatic bile duct, congestive heart failure (non-hypertensive), among others. In both approaches, a supervised learning model (such as logistic regression, SVM, or random forests) was needed to map the learned representation to the outcome of interest. For this reason, we refer to such approaches as “modular”.

Neither eNRBM nor SDA considered the temporal information in the EHR, explicitly; the attempt by Tran et al. [40] to split patients’ journeys into non-overlapping time intervals, and concatenating them (see the details in Section 2) was the closest that modular models have become to considering time. In order to address this shortcoming, Nguyen et al. [20] introduced a CNN architecture called Deepr (Deep record), where a patient’s journey is modelled as a long sentence of medical codes, with each code embedded into a new space to facilitate the statistical and algebraic operations (i.e., similar to word embedding in natural language processing [41]), and denoting the

time between events as “special words”. DeepR is an example of an “end-to-end” model, trained to map the EHR history directly to the outcome of interest. Such a model was validated on hospital data to predict unplanned readmission after discharge, obtaining a slightly better performance with an AUC of 0.80 (3-month prediction) and 0.819 (6-month prediction), compared to logistic regression with bag-of-words representation, where an AUC of 0.797 (3-month prediction) and 0.811 (6-month prediction) was obtained.

In another end-to-end modelling endeavour, Choi et al. introduced Med2Vec [42], an FFNN model for learning representations for both visits (a group of medical codes) and medical codes, providing word embeddings comparable to other techniques like Skip-gram [43], GloVe [44], and stacked autoencoders [38]. Furthermore, in multiple studies [45, 46, 47], Choi et al. extended their work by using an RNN architecture to detect influential past visits and significant clinical variables while remaining clinically interpretable. In particular, the RETAIN (REverse Time AttentIoN) model [45] used two RNNs and a two-level neural attention model to process sequential data using an end-to-end approach. The model obtained an AUC of 0.8705 when tested on heart failure diagnosis from Sutter Health medical records, suggesting that the model had a high predictive accuracy, while providing interpretability in the results. Since its development, the RETAIN model has been enhanced considerably. For instance, Kwon et al. [48] produced an interactive visual interface named RetainVis, that offers insights into how individual medical codes contribute to making risk predictions, Ma et al. [49] developed an attention-based bidirectional RNN that measures the relationships of past and future visits (from Medicaid claims in the US) to produce diagnosis pre-

98 diction, while providing clinically meaningful interpretations, and Choi et al.
99 [50] built a graph-based attention model, capable of using hierarchical infor-
100 mation inherent to medical ontologies, resulting in a 10% higher accuracy for
101 predicting rare diseases and 3% improved AUC for predicting heart failure
102 using an order of magnitude less training data.

103 RNN’s power in modelling language and other sequences [51, 52], when
104 paired with their early success in learning new representations of EHR (as in-
105 troduced above) made it a popular choice for DL researchers in this domain.
106 Pham et al. [21] introduced an RNN architecture (called DeepCare) to pre-
107 dict future medical outcomes based on the patients’ health trajectories. To
108 do so, they used a modified long short-term memory (LSTM) unit [53, 54]
109 with a forgetting mechanism and the ability to handle irregular inter-visit
110 intervals. The model was applied to prediction of unplanned readmission
111 within 12 months, where an F-score of 0.791 was obtained, an improvement
112 over traditional machine learning techniques like SVM (F-score of 0.667) and
113 random forests (F-score of 0.714). Extending this idea, Rajkomar et al. [22]
114 introduced an ensemble model [55], which combined the strengths of three
115 different models: a weighted RNN, a feedforward model with time-aware at-
116 tention, and a boosted embedded time series model. The authors applied
117 their model to EHR data from the University of California, San Francisco,
118 and the University of Chicago Medicine. They demonstrated the effective-
119 ness of this DL model in a variety of clinical tasks, including mortality pre-
120 diction (AUC of 0.93-0.95), unexpected readmission (AUC of 0.76-0.77), and
121 increased length of stay (AUC of 0.85-0.86). In addition, Ma et al. [56] de-
122 veloped a hybrid network structure composed of both an RNN and a CNN

123 in order to extract comprehensive multifaceted patient information patterns
124 with attentive and time-aware modulars. Such a network was applied in the
125 prediction of heart failure (SNOW dataset) and diabetes mellitus (EMRbots
126 dataset), obtaining AUCs of 0.729 (heart failure) and 0.943 (diabetes mel-
127 litus), outperforming other models like logistic regression (0.650/0.790) and
128 RETAIN (0.668/0.767).

129 Despite the growth in the number of DL papers for EHR, the majority of
130 these papers have used data from different countries and healthcare systems,
131 employed different data cleaning and preparation processes (e.g., inclusion /
132 exclusion criteria, input fields, and sample size), developed and used differ-
133 ent DL architectures, while answering different clinical questions; even when
134 assessing the models for risk prediction, the reviewed papers were not con-
135 sistent in the quantification of model performance (see Table 1). Such wide
136 paper-to-paper variability will hamper the ability of the field to compare and
137 contrast the relevant papers for a clinical question at hand.

138 This paper primarily aims to provide a comparative review of the key
139 DL architectures that have been used for the analysis of EHR data; de-
140 scribing their corresponding strengths and weaknesses for various real-world
141 challenges when applied to a single complex EHR. We focused our work
142 on getting an efficient patient representation to tackle two important but di-
143 verse clinical prediction challenges: predicting emergency admission, or heart
144 failure. In addition to being the first comparative review of key DL archi-
145 tectures for EHR data, this paper also aims to: (1) introduce and use one of
146 the world’s largest and most complex databases of linked primary care EHR
147 (i.e., Clinical Practice Research Datalink, or CPRD) that captures 30 years

148 of medical history of patients, as a new asset for training such data-hungry
149 models; (2) review some of the approaches for working with such data for
150 DL; (3) share some of the best practices for assessing the performance of
151 DL models in clinical risk prediction tasks; (4) and propose ideas of future
152 research for making DL models even more suitable for EHR data.

153 **2. Materials and Methods**

154 *2.1. Overview*

155 Our comparative review assessed four main DL architectures that have
156 been employed so far for the analysis of EHR data: AE (autoencoders), SDA,
157 CNN, and RNN (for complete reference, please see Table 1 with the corre-
158 sponding abbreviations available in Appendix A). In particular, we used the
159 following models: eNRBM [40], Deep Patient [38], Deepr [20], and RETAIN
160 [45]; this is the best representative subset for all the key DL architectures in
161 the field [3, 4, 5].

Table 1: Main approaches of Deep Learning for EHR (the abbreviations' meaning can be found in Appendix A)

Model	Architecture	Learning Process	# Patients considered	Predictors Considered	Outcome	Performance Metrics
eNRBM [40]	AE	Modular	7,578	Diagnoses (ICD-10), procedures (ACHI), Elixhauser comorbidities, diagnosis related groups, emergency attendances and admissions, demographic variables (ages in 10 year intervals and gender)	suicide risk prediction	F-score Recall Precision
Deep Patient [38]	SDA	Modular	704,587	Demographic variables (i.e., age, gender, and race), diagnoses (ICD-9), medications, procedures, lab tests, free-text clinical notes	future disease prediction	AUROC Accuracy F-score Precision
Deepr [20]	CNN	End-to-end	300,000	Diagnoses (ACS), procedures (ACHI)	unplanned readmission prediction	AUROC
DeepCare [21]	RNN	End-to-end	7,191	Diagnoses (ICD-10), procedures (ACHI), medications (ATC)	disease progression, unplanned readmission prediction	Precision F-score Recall
Med2Vec [42]	FFNN	End-to-end	ND	Diagnoses (ICD-9), procedures (CPT), medications (NDC)	medical codes in previous/future visits	AUROC Recall
Doctor AI [46]	RNN	End-to-end	263,706	Diagnoses (ICD-9), procedures (CPT), medications (GPI)	medical codes in future visits, duration until next visit	Recall R^2
RETAIN [45]	RNN	End-to-end	32,787	Diagnoses (ICD-9), procedures (CPT), medications (GPI)	heart failure prediction	AUROC
RetainVis [48]	RNN	End-to-end	63,030 (heart failure) 117,612 (cataract)	Diagnoses (KCD-9), procedures, medications	heart failure prediction, cataract prediction	AUROC AUPRC
Ensemble model [22]	RNN FFNN	End-to-end	216,221	Demographics, provider orders, diagnoses, procedures, medications, laboratory values, vital signs, flowsheet data, free-text medical notes	inpatient mortality, 30-day unplanned readmission, long length of stay, diagnoses	AUROC

162 2.2. *Clinical Practice Research Datalink (CPRD)*

163 Over 98% of the UK population are registered with a general practice
164 (GP), the first point of contact for healthcare in the UK National Health
165 Service [57]. The CPRD is a service that collects de-identified longitudinal
166 primary care data from a network of GPs in the UK, which are linked to
167 secondary care and other health and area-based administrative databases
168 [58]. Among these linked databases include the Hospital Episode Statistics
169 (for data on hospitalisations, outpatient visits, accident and emergency at-
170 tendances, and diagnostic imaging), the Office of National Statistics (death
171 registration), Public Health England (cancer registration), and the Index
172 of Multiple Deprivation. Around 1 in 10 GP units contribute data to the
173 CPRD. To date, the CPRD covers 35 million patient lives among whom 10
174 million were currently registered patients from 674 GP units, making it one
175 of the largest primary care EHR databases in the world.

176 Patients included in the CPRD database were largely nationally repre-
177 sentative in terms of age, sex and ethnicity, covering around 7% of the UK
178 population. Given the data on demographics, diagnoses, therapies, and tests
179 together with its linkage to other health-related databases, the CPRD is a
180 valuable source of healthcare data [57]. Because CPRD contains detailed
181 personal information, the dataset is not readily available to the public, and
182 its usage depends on approval from the CPRD Research Ethics Committee
183 [58].

184 For this work, we only considered practices providing healthcare data that
185 met research quality standards within the period from 01/January/1985 to
186 31/December/2014, and whose records were linked to the Hospital Episode

Table 2: Statistics of CPRD dataset from 01/January/1985 to 31/December/2014. (SD: Standard Deviation, IQR: Interquartile Range, D:Diagnosis, M:Medication)

Number of patients	4, 272, 833
Number of visits	283, 996, 690
Number of visits per patient, Mean (SD)	66.47 (91.1)
Number of visits per patient, Median (IQR)	30 (79)
Number of medical codes	326 (D:222, M:104)
Number of codes in a visit, Mean (SD)	2.34 (1.92)
Number of codes in a visit, Median (IQR)	2 (2)

187 Statistics database. Furthermore, we focused on patients aged 16 years or
188 older who have been registered with their GP for at least 1 year. This resulted
189 in a dataset of 4, 272, 833 patients, and is profiled in Table 2.

190 2.3. Modeling Pipeline

191 While CPRD contains many data fields, we limited our analyses to two
192 main scenarios: demographics + diagnoses, and demographics + diagnoses +
193 medications (denoted as DD and DDM, respectively), as these are common
194 predictors that were used by all models in Table 1, and were very likely to
195 exist in all EHR systems. The demographics variables that we took into
196 account were sex (binary), age (continuous), ethnicity (categorical with 12
197 classes), and region (categorical with 10 classes). In CPRD, diagnoses were
198 coded using Read codes [59] (for primary care) and ICD-10 codes [60] (for
199 Hospital Episode Statistics), while medications were coded using prodcodes
200 (CPRD unique code for the treatment selected by the GP). In their raw
201 format, these coding schemes were unsuitable to work with given their high

cardinality, i.e., there were around 110,000 Read codes, 14,000 ICD-10 codes and 55,000 prodcodes. For this reason, we mapped diagnoses from Read and ICD-10 to the CCS (Clinical Classifications Software) [61] coding scheme. For medications, we mapped the prodcodes to level 2 of BNF (British National Formulary) [62] codes. These mappings are known as “clinical groupers” and are commonly used in clinical tasks to reduce the number of medical codes commonly found in EHR.

In our work, we decided to use CCS because it was the most common approach in the papers that we compared. Furthermore, this approach helps address the sparsity issue that is common in medical records, and makes statistical analyses and reporting easier [61]; defining the diseases at a higher granularity will lead to lower frequencies of their occurrences and hence hampers the models’ ability of learning meaningful patterns about them. Furthermore, using clinical groupers can help avoid the memory issues, as raised by Choi et al. [45], [50], and from a practical point of view, using the full vocabulary of ICD codes (i.e., more than 10,000 diseases) is not a tractable option for the eNRBM and Deep Patient models, given that such architectures do not use embedding layers and therefore, the memory in the GPUs does not have enough space to allocate such information. In addition to this, diagnoses and medications with a cumulative sum of less than 0.5% of the total number of medical codes (recorded from 01/January/1985 to 31/December/2014), were grouped into the **Drare** and **Mrare** categories, respectively. This resulted in a total of 222 codes for diagnoses and 104 for medications.

For both the DD and DDM scenarios, we focused on two separate out-

227 comes: predicting emergency admission, or heart failure. We chose these two
 228 outcomes to capture two important but diverse clinical prediction challenges.
 229 Overall, emergency hospital admission is of great importance in healthcare
 230 service delivery. Predicting such an outcome helps in planning to ensure
 231 services are available to meet unscheduled hospitalisations [63]. The reader
 232 can find the emergency admission codes used in this study in Appendix B.
 233 Furthermore, heart failure is an important cause of mortality and morbid-
 234 ity, which could be associated with poor prognosis. Recent trends suggest
 235 that the burden of heart failure is increasing, not least because of the trend
 236 towards an ageing population [64]. In this study we focused on CCS code
 237 108, i.e. Congestive Heart Failure; Nonhypertensive. The reader can find the
 238 related ICD codes in Appendix C. In both cases, we considered a prediction
 239 window of 6 months, and the outcome was treated as a binary classification
 240 task. Patients were split into separate groups, randomly keeping 53% for
 241 training of DL models, 7% for training the top layer classifier in the modular
 242 DL architectures (i.e., a random forest in our study), 20% for validation (hy-
 243 perparameter tuning), and 20% for testing. All groups included data from
 244 01/January/1985, but the time window was different for each of them. The
 245 training group covered until 31/December/2012, the validation group until
 246 31/December/2013 and the test group until 31/December/2014. The pur-
 247 pose of this time-splitting is to have a form of external validation (i.e., out
 248 of time, out of sample) that is typical in clinical studies, where we want to
 249 predict future events based on everything we know up to a certain baseline.
 250 In all cases, the prediction interval was used to create the target variable
 251 and corresponded to the last six months previous to the end of the corre-

Table 3: Statistics of the outcome variables for the different data partitions for the Demographics + Diagnoses scenario

Data partition	# of Emergency Admission cases	# of Heart Failure cases
Training (53%)	20,430 (1.08%)	2,754 (0.15%)
Training (modular models) (7%)	2,773 (1.08%)	378 (0.15%)
Validation (20%)	8,262 (1.14%)	1,114 (0.15%)
Testing (20%)	7,115 (0.98%)	982 (0.14%)

sponding time window (i.e. whether a patient had an emergency admission, or whether he had a heart failure during these six months). Overall, the ratio of emergency admission cases is 0.011, while the ratio of heart failure cases is 0.0015. The statistics of the outcome variables are shown in Tables 3 and 4. Data before these six-month windows are used for feature generation (see a visual representation of this approach in Fig. 2). Patients who died before the corresponding prediction interval are excluded from the analysis. For heart failure cases, we did not focus on new incidences since we wanted to identify any heart failure presentation (whether first event or recurrence) given the patients history up to a baseline.

While using the same source of data and the same inclusion and exclusion criteria, each model in Table 1 requires a different format for its input data, as described below:

- *eNRBM*: Each patient’s medical history (i.e., the data before dashed line in Fig 2) was split into five non-overlapping intervals: $(0 - 3)$,

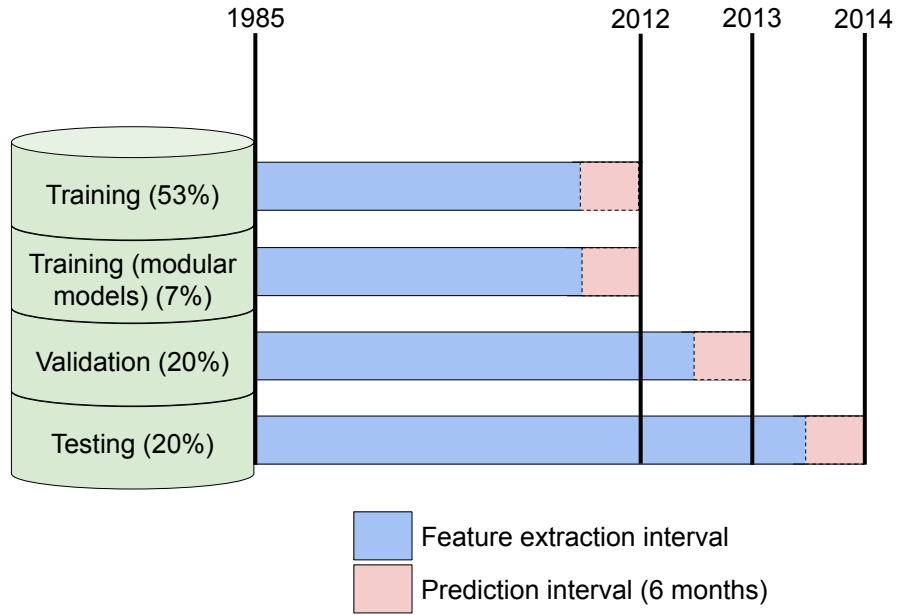


Figure 2: Study Design. A total of 4,272,833 patients were split into training, validation and test sets. A small portion of the training (7%) was kept apart for the modular deep learning architectures. Red region corresponded to the prediction interval, which consisted of the last six months of the respective time window. This region was used to create the two separate outcomes (i.e. whether or not a patient had an emergency admission, or a heart failure during these six months). All data previous to these six months (marked with a dashed line) was used for feature extraction. Patients who died before the corresponding prediction interval are excluded from the analysis.

Table 4: Statistics of the outcome variables for the different data partitions for the Demographics + Diagnoses + Medications scenario

Data partition	# of Emergency Admission cases	# of Heart Failure cases
Training (53%)	23,200 (1.18%)	2,764 (0.14%)
Training (modular models) (7%)	3,064 (1.16%)	385 (0.15%)
Validation (20%)	9,308 (1.24%)	1,032 (0.14%)
Testing (20%)	7,998 (1.07%)	947 (0.13%)

(3 – 6), (6 – 12), (12 – 24), and (24+) months before the prediction window. The information in each interval is formatted as a long sparse vector of length equal to 222 for DD scenario and 326 for DDM scenario, where each entry corresponds to the number of times a single disease (or medication) was diagnosed (or prescribed) in the given interval. All categorical features are converted to dummy variables, and age is scaled between 0 and 1 using information from the training set. This resulted in a sparse vector that, together with demographics information, has 1,135 input variables for DD scenario, and 1,655 for DDM scenario.

- *Deep Patient*: Each patient’s medical history was aggregated into a sparse vector, where each entry corresponds to the number of times a single disease (or medication) was diagnosed (or prescribed). All categorical features are converted to dummy variables, and age is scaled between 0 and 1 using information from the training set. Together with demographic information, this resulted in 247 input variables for DD

scenario, and 351 for DDM scenario.

- *Deeprr*: It deals with the sequence of events (i.e., diagnoses and medications) directly, while discretising the time between two events as: $(0 - 1]$, $(1 - 3]$, $(3 - 6]$, $(6 - 12]$, and $12+$ months, e.g. $(0 - 1]$ corresponds to a duration that goes from 0 (exclusive) to 1 (inclusive) months between medical events. Each time interval is assigned a unique identifier, which is treated as a word; for instance, `0-1m` is a word for the $(0 - 1]$ interval gap. Each patient’s medical history is expressed as a long sentence of medical codes, where the duration between two consecutive visits will appear in the sequence as one of the five preserved interval words. Similar to the approach of Nguyen et al. [20], the sentences were trimmed to keep the last 100 words in order to avoid the effects of some patients, who have very long sentences, that could severely skew the data distribution.
- *RETAIN*: Each patient’s information was split in three components. The first one was made of a list of sublists, where each sublist contained all the medical codes that were recorded on a single date. These sublists were sequentially ordered from oldest to newest. The second component consisted of a list with the demographics information. Finally, the third component, was made of a list of values indicating the difference in days between consecutive visits. These values were ordered from oldest to newest, and the first of these differences was set to zero, given that this information was not available for the first visit.

This input formatting was carried out after the data partitioning process.

306 A short description of the inclusion criteria and outcomes of interest of the
307 selected works are shown in the Appendix D. For more details, readers are
308 directed to the corresponding publications.

309 All four models selected were coded based on the description found in
310 the corresponding publications, except for the RETAIN model, which has
311 the code available online [65]. Note that the original model in Nguyen et
312 al. [20] for the Deepr architecture only considered dynamic information as
313 the input of the convolutional layer (after embedding). However, for the
314 sake of a fair comparison with the other selected models, we had to include
315 static information (i.e. demographics) into the network. This resulted in a
316 slight modification of the original architecture, via the parallel processing of
317 two kinds of inputs (multi-input network). This adjustment amounted to
318 merging two networks: the first (processing dynamic information) had the
319 same structure as Deepr, whereas the second (processing static information)
320 consisted of one fully connected layer. The two intermediate outputs were
321 then merged together through a further fully connected layer, before the final
322 classification task (see Fig. 3).

323 The implementations were done with PyTorch 0.4.1 [66], Keras 2.2.2 [67]
324 and scikit-learn 0.20.0 [68] using two NVIDIA Titan Xp graphics cards. Fur-
325 thermore, all DL models selected have several hyperparameters as shown in
326 Appendix E. In order to tune these hyperparameters, we used a tree-based
327 Bayesian Optimisation approach [69, 70], as implemented in Scikit-Optimize
328 [71]. We chose a batch size of 1,024 patients for all models, except for RE-
329 TAIN, where we used a batch size of 256 patients due to memory constraints
330 in the GPUs. To benchmark our results, we used bag-of-words representa-

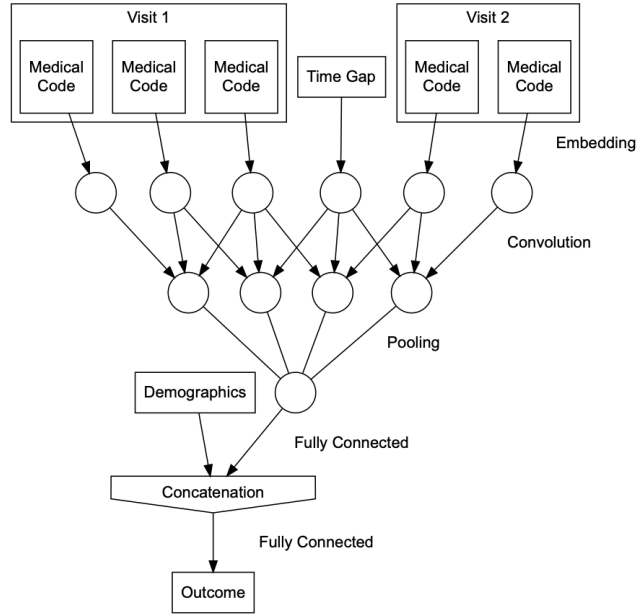


Figure 3: Modified DeepR architecture in order to consider demographics and have a fair comparison with the other selected models. EHR are processed through a series of steps that include sequencing, embedding, convolution, pooling and classification. Static information (i.e., demographics) is merged together and passed through a fully connected layer before the final classification task.

tion with logistic regression (BOW + LR), and a simple RBM with the same input format as in Tran et al. [40]. For the random forest in the modular DL models, we used 200 trees while the remaining hyperparameters were optimised using the same tree-based Bayesian Optimisation approach¹.

2.4. Performance Metrics

Model performance was evaluated using the area under the ROC curve (AUROC), similar to what the majority of the articles in Table 2 reported. The AUROC is a discrimination metric equivalent to the probability that a classification model will rank a randomly chosen positive instance higher than a randomly chosen negative instance [72, 73]. Thus, the AUROC ranges from 0.5 (no discrimination ability) to 1 (perfect discrimination). However, this metric can be misleading when the classes are very imbalanced [74, 75, 76]. To avoid this, the area under the precision-recall curve (AUPRC) is useful as it considers the tradeoff between precision (a measure of result relevancy) and recall (a measure of how many truly relevant results are returned) for different thresholds [77, 78]. The AUPRC range goes from 0 to 1, with higher values preferred overall [79]. In addition, the F1-Score is also taken into account given that it is a weighted average of the precision and recall, ranging from 0 (worst score) to 1 (best score) [80].

3. Results

The performances of all the tested models are shown in Tables 5 and 6 for the DD scenario, and Tables 7 and 8 for the DDM scenario. The models

¹For the interested readers, our codes will be shared upon request.

353 were trained 5 times independently to provide summary statistics and 95%
 354 confidence intervals. In general, the RETAIN model has the best performance
 355 when compared to other DL models; in both scenarios, and across a range
 356 of metrics. This agrees with the results presented in Choi et al. [45] in
 357 terms of AUROC, and emphasises the advantage of using RNN to exploit the
 358 sequential nature of EHR. In addition, the inclusion of medications provides
 359 an improvement in the AUROC and AUPRC for the majority of the models
 360 in these two prediction tasks. The best hyperparameters found for all the
 361 DL and baseline models are shown in Appendix E.

362 When compared to simpler techniques such as BOW + LR, all DL mod-
 363 els show better performance. For instance, when considering Emergency
 364 Admission in the DD scenario, the RETAIN model outperforms the BOW
 365 + LR model by 25% in AUROC, 264% in AUPRC, and 174% in F1-Score.
 366 Similarly, for Heart Failure cases, the RETAIN model outperforms the same
 367 by 18% in AUROC, 490% in AUPRC, and 350% in F1-Score. On the other
 368 hand, when considering Emergency Admission in the DDM scenario, the RE-
 369 TAIN model outperforms the BOW + LR model by 31% in AUROC, 336%
 370 in AUPRC, and 183% in F1-Score. Similarly, for Heart Failure cases, the
 371 RETAIN model outperforms the same by 39% in AUROC, 800% in AUPRC,
 372 and 515% in F1-Score. Overall, this is likely due to the ability of the DL
 373 models to reduce a complex input space (made out of hundreds of diagnosis
 374 and medication codes) to a small yet predictive representation of the EHR
 375 sequence for predicting either emergency admission or heart failure. Simi-
 376 larly, when compared with the simple RBM model, the eNRBM has a similar
 377 performance in both classification tasks, as expected, given the similar layout

Table 5: Comparison for the Demographics + Diagnoses scenario (Emergency Admission)

Model	AUROC	AUPRC	F1-Score
eNRBM	0.803 (0.803-0.804)	0.051 (0.051-0.051)	0.052 (0.052-0.052)
Deep Patient	0.801 (0.801-0.802)	0.047 (0.047-0.047)	0.053 (0.053-0.053)
Deepr	0.815 (0.813-0.817)	0.053 (0.052-0.054)	0.112 (0.107-0.117)
RETAIN	0.822 (0.819-0.826)	0.062 (0.061-0.064)	0.118 (0.109-0.127)
BOW + LR	0.654 (0.627-0.681)	0.017 (0.015-0.018)	0.043 (0.031-0.054)
RBM	0.807 (0.807-0.808)	0.050 (0.050-0.050)	0.053 (0.053-0.053)

* Data represented as: Mean (95% Confidence Interval)

of both models (see Table E.11). Further, both Deep Patient and Deepr have a slightly worse performance compared to the simple RBM when predicting emergency admission, but this is improved when predicting heart failure. Finally, the RETAIN model has the best performance in both tasks for the DD and DDM scenarios.

4. Discussions and Future Work

DL is becoming the ubiquitous approach for the analysis of EHR data, because of the models' ability to process huge amounts of data without the need to perform explicit feature engineering by domain experts unlike more traditional statistical models. In this work, we carried out a comparative review of the common DL architectures for EHR, to compare their properties while standardising the common sources of paper-to-paper variability (e.g., dataset, sample size, medical codes, clinical questions, and performance metrics). To the best of our knowledge, this is the first paper that has carried

Table 6: Comparison for the Demographics + Diagnoses scenario (Heart Failure)

Model	AUROC	AUPRC	F1-Score
eNRBM	0.912 (0.911-0.913)	0.018 (0.017-0.018)	0.017 (0.017-0.017)
Deep Patient	0.948 (0.948-0.949)	0.040 (0.039-0.041)	0.024 (0.024-0.024)
Deepr	0.938 (0.907-0.969)	0.049 (0.036-0.063)	0.110 (0.089-0.131)
RETAIN	0.951 (0.949-0.952)	0.065 (0.061-0.069)	0.135 (0.126-0.144)
BOW + LR	0.801 (0.775-0.827)	0.011 (0.009-0.012)	0.030 (0.025-0.034)
RBM	0.897 (0.897-0.898)	0.016 (0.016-0.017)	0.015 (0.015-0.015)

* Data represented as: Mean (95% Confidence Interval)

Table 7: Comparison for the Demographics + Diagnoses + Medications scenario (Emergency Admission)

Model	AUROC	AUPRC	F1-Score
eNRBM	0.831 (0.831-0.832)	0.071 (0.071-0.071)	0.063 (0.062-0.063)
Deep Patient	0.813 (0.813-0.813)	0.060 (0.060-0.061)	0.059 (0.059-0.059)
Deepr	0.829 (0.828-0.831)	0.069 (0.067-0.071)	0.131 (0.118-0.144)
RETAIN	0.847 (0.845-0.849)	0.083 (0.082-0.083)	0.153 (0.151-0.154)
BOW + LR	0.646 (0.576-0.717)	0.019 (0.015-0.023)	0.054 (0.046-0.063)
RBM	0.840 (0.840-0.840)	0.072 (0.072-0.073)	0.066 (0.066-0.066)

* Data represented as: Mean (95% Confidence Interval)

Table 8: Comparison for the Demographics + Diagnoses + Medications scenario (Heart Failure)

Model	AUROC	AUPRC	F1-Score
eNRBM	0.920 (0.920-0.921)	0.020 (0.019-0.021)	0.014 (0.014-0.014)
Deep Patient	0.947 (0.947-0.948)	0.040 (0.039-0.041)	0.023 (0.022-0.023)
DeepR	0.949 (0.947-0.952)	0.039 (0.032-0.046)	0.085 (0.049-0.120)
RETAIN	0.950 (0.946-0.954)	0.054 (0.053-0.056)	0.117 (0.098-0.136)
BOW + LR	0.682 (0.613-0.752)	0.006 (0.002-0.009)	0.019 (0.011-0.027)
RBM	0.917 (0.917-0.917)	0.023 (0.022-0.023)	0.014 (0.014-0.014)

* Data represented as: Mean (95% Confidence Interval)

392 out such research.

393 Most of the DL research in this field, however, analysed EHR data that
394 came from private hospitals [40, 38, 20, 45]. These datasets vary in terms of
395 sample size and the information they contain (demographics, clinical data,
396 vital signs, laboratory tests and medications, among others), and hence DL’s
397 performance on them might not be reproduced across other EHR datasets
398 (and hence different populations). Among EHR data sources, CPRD is one
399 of the largest accessible EHR datasets for a patient population that is largely
400 representative of the UK population in terms of age, sex, and ethnicity. The
401 CPRD not only includes data on demographics, diagnoses, therapies, and
402 tests, but it is also linked to data on hospitalisations, outpatient visits, ac-
403 cident and emergency attendances, diagnostic imaging, death registration,
404 cancer registration, and socioeconomic status [57]. It is a valuable source
405 of healthcare data that enables models to be trained in a real healthcare

406 ecosystem.

407 Although CPRD has been around for over a decade with hundreds of
408 publications and widespread use in epidemiology [81, 82, 83, 64, 84], its
409 potential for machine learning and DL is currently untapped and its use only
410 in early stages (i.e., Rahimian et al. [19]). With this study being the first
411 to employ DL on CPRD, we aimed to show the richness of this database
412 for various types of DL frameworks, given its large sample size, variety of
413 information, and representativeness of the UK population, which can be used
414 on a large range of analyses on risk predictions and disease trajectories.

415 In this work, we used four models that covered the main DL architectures
416 found in the literature. The eNRBM and Deep Patient models were trained in
417 a modular way. First, they aimed to get a meaningful unsupervised represen-
418 tation that identifies hierarchical regularities and dependencies in the input
419 data; such a representation is then used as input for a supervised learning
420 algorithm (i.e., a random forest in our study). Given both models' inability
421 to deal with sequential data explicitly, they require to represent a patient's
422 journey as a sparse vector, which ignores the elapsed time between visits. In
423 practice, however, time between and since events is clinically relevant. For
424 instance, Rahimian et al. [19] have shown a 10.8% higher AUC (0.848 with a
425 gradient boosting classifier compared to 0.740 with a Cox proportional haz-
426 ards model) for prediction of risk of emergency admission within 24 months
427 using engineering features that represented time since diagnoses. Thus, most
428 recent developments in the use of DL for modelling EHR data has been using
429 sequential models such as RNN.

430 DeepR and RETAIN models, on the other hand, are trained using an

431 end-to-end approach, directly connecting the input (e.g., medical codes and
 432 other information) to the outcome of interest. This can be an advantage if
 433 the aim is to identify meaningful patterns for a specific clinical question, but
 434 if this is not the case, a different model needs to be trained for each given
 435 outcome. Nevertheless, since these models incorporate the elapsed time be-
 436 tween visits, they are more suitable for modelling patient trajectories. While
 437 DeepR employed an interesting approach to characterise time between events
 438 by defining special words that represent time intervals, RETAIN considered
 439 time between visits as an additional feature. Overall, the RETAIN model
 440 has the best performance in both tasks for the DD and DDM scenarios. This
 441 highlights the strength of RNN and attention mechanisms for analysing se-
 442 quential data, thus various modifications to RNN are currently the state of
 443 the art research in DL for EHR.

444 Regardless of which classification model one uses, one of the most com-
 445 mon ways to assess such model performance is the AUROC. In our study, we
 446 achieved a high AUROC that is comparable to what other papers reported
 447 [20, 45, 48]. Furthermore, we employed an additional metric, AUPRC, which
 448 most papers in the literature did not report, and with respect to which our
 449 models performed poorly. Such observations have been previously reported
 450 by others, e.g., Davis and Goadrich [85], when datasets were highly imbal-
 451 anced. It is possible that focusing on more specific or relevant clinical sub-
 452 groups may improve the prediction as models are likely to incorporate more
 453 relevant parameters for the condition of interest than for all the patient pop-
 454 ulation. This highlights the difficulty that even these models encounter when
 455 dealing with imbalanced clinical tasks. Such a scenario is typical in many

456 healthcare applications where the minority class is buried by the majority
457 class. Several approaches have been proposed for dealing with imbalanced
458 datasets, including the work by Chawla, et.al. [86] and Lemaitre, et.al. [87],
459 but further research is required, particularly for DL models.

460 One of the key strengths of DL models is their ability to map one’s full
461 medical record to a low-dimensional representation. In our analyses, the pa-
462 tient representations (projected onto 2D using a UMAP (Uniform Manifold
463 Approximation and Projection) [88] for visual clarity) are shown in Figs 4
464 and 5. Note that the representations from modular models tend to form
465 clusters, while the representations from the end-to-end models seem to form
466 a continuum. Nevertheless, it can be observed that cases of emergency ad-
467 mission (in blue) and heart failure (in red) tend to get grouped in separable
468 regions. Further profiling of these clusters can help identify and assess com-
469 mon characteristics of patients to improve the performance of the models
470 and to guide future clinical research.

471 As part of the objectives of this work to provide a guideline to work with
472 DL for EHR, we identified that the main issues are related to the choice of
473 inclusion criteria and preprocessing of the data. The inclusion criteria usually
474 depends on the clinical question and outcome of interest; it is a key part of
475 a good study design. Furthermore, it is of great importance to properly
476 split the data into the usual training, validation and test sets as this would
477 help to avoid overfitting and properly evaluate the DL models. In terms of
478 data preprocessing, we highlighted the importance of using clinical groupers
479 to reduce the cardinality of both diagnoses and medications. This could
480 be thought as a form of dimensionality reduction that prevents very long

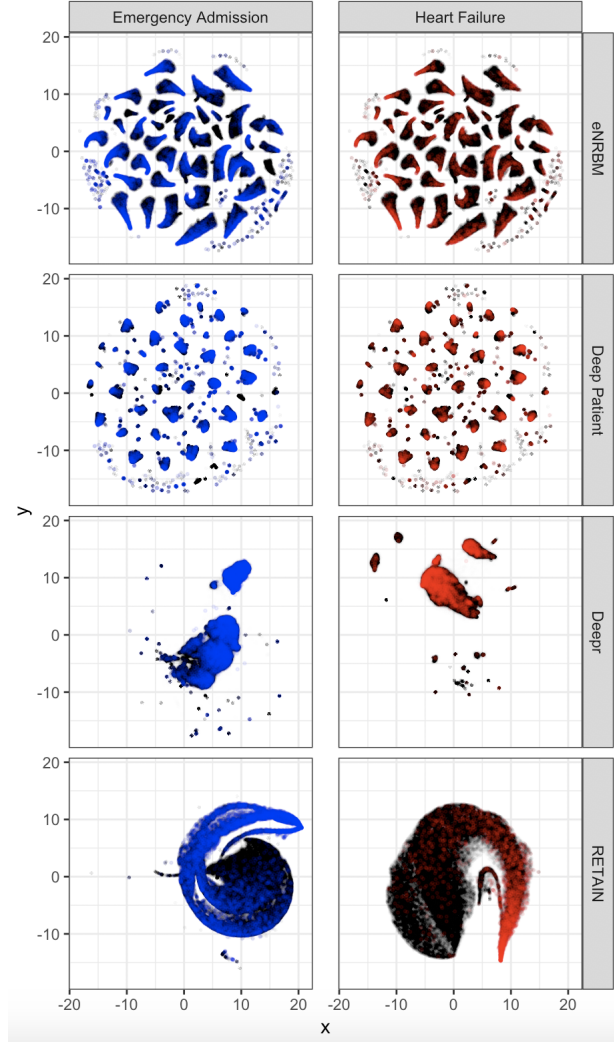


Figure 4: Patient representations obtained by each DL model in the DD scenario, after projection to 2D using UMAP. Each point corresponds to a patient in the test set; blue and red dots correspond to cases of emergency admission or heart failure, respectively.

481 and sparse input vectors that would hamper the training process of the DL
 482 models. In this work, we used the CCS coding scheme, but other clinical [89]
 483 and drug [90] classification systems could be used depending on the study

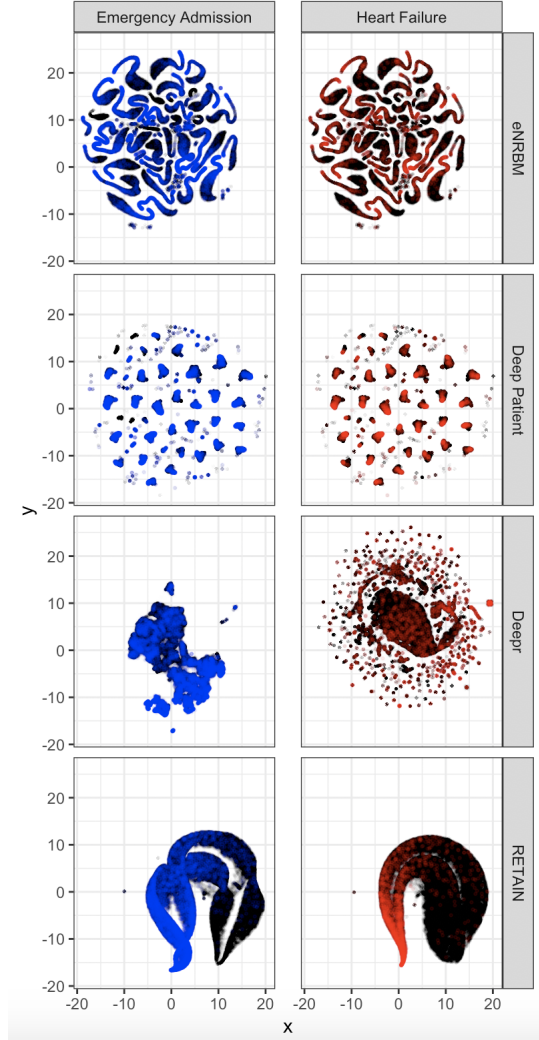


Figure 5: Patient representations obtained by each DL model in the DDM scenario, after projection to 2D using UMAP. Each point corresponds to a patient in the test set; blue and red dots correspond to cases of emergency admission or heart failure, respectively.

484 requirements.

485 One important consideration when working with DL models is the tun-
 486 ing of the corresponding hyperparameters (as shown in Appendix E). In this

work, we employed a tree-based Bayesian Optimisation approach [69, 70], as implemented in Scikit-Optimize [71], to find the most appropriate values. This approach takes fewer evaluations to achieve a similar result compared with exhaustive grid search and random search given that it learns from its evaluation history and suggests better hyperparameters to test as the search progresses. Still, Bayesian optimisation can be time-consuming as the models become larger and more complex, so a lot of research has been done to make this process more robust and scalable [91, 92, 93, 94]. Nowadays, new best practices are proposed constantly in this fast-evolving field, so it is hard to provide a general set of rules for efficient hyperparameter tuning. The work by Howard, et.al. [95] has been highly influential as the authors suggest that the learning rate is the most important hyperparameter when training a DL model, and advocate for a more iterative approach using techniques like differential learning rates, cyclical learning rates [96], learning rate schedulers, and stochastic gradient descent with restarts [97], which seem to work irrespective of the DL architecture.

Overall, there is no DL model that dominates the current state-of-the-art for the analysis of EHR, and that outperforms all expert-driven models yet (as shown in the works of Jacobs, et.al. [98] and Rahimian, et.al. [19]). Furthermore, DL models remain difficult to interpret and unable to provide uncertainty estimation, which is undesirable in clinical settings. Although, the works by Gal, et.al. [99, 100], Tran, et.al. [101] and Chakraborty, et.al. [102] have contributed to overcome these issues, further research is required.

Finally, there seems to be a trend towards the use of RNN architectures given the sequential nature of medical records. Several works have been de-

veloped recently where RNN are applied to a series of case studies like early
detection of sepsis [103], hospital readmission for lupus patients [104], and
detection of adverse medical events in EHR [105]. Furthermore, the field
of natural language processing has greatly advanced as evidenced by the
works of Transformers [106, 107], contextualised word representations [108],
and transfer learning for language models [109]. Given the similarities be-
tween natural language processing tasks and EHR, the application of these
techniques provides interesting research directions that could result in the
development of medical embeddings for analysis of EHR [110, 111], under-
standing of disease clusters and trajectories [112], generation of synthetic
EHR [113], and the use of pretrained models to avoid the need for starting
from scratch every time a new clinical question comes. The application of
DL to EHR seems to have an exciting future.

5. Conclusions

The use of DL to analyse EHR data has increased over the past years; a
growth that is continuing to be facilitated by the availability of more data
(EHR and beyond), developments in DL (specifically, models for sequential
data), and innovative ways of combining these two trends. In this work,
we implemented key deep learning architectures to learn an efficient patient
representation for predicting emergency admission, and heart failure. Our
objective here was to help the field have a comparative view of these ap-
proaches, and to assess their strengths and weaknesses when it comes to
EHR.

Along this work, we introduced CPRD, which is one of the world’s largest

primary care databases, and showed how data from primary care can provide predictions that can be of value in policy and practice of care. Given the complexity of primary care EHR (heterogeneous events recorded in irregular intervals with varying degree of richness and quality across different individuals), and its importance in provision of care in many parts of the world, we believe that the set of best practices we shared for them (e.g., inclusion criteria, preprocessing, medical codes/grouping, performance metrics, and hyperparameter tuning) will be of great value in helping DL research in EHR.

Our work showed the strength of recurrent neural networks in dealing with the temporal nature of the EHR. This was consistent with the developments in modelling EHR-like data from other domains, such as natural language and time series data. Our future research aims to explore techniques and methodologies from such domains, and apply them to other types of data from different healthcare systems.

6. Acknowledgements

This research was funded by the Oxford Martin School (OMS) and supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC). It also got support from the PEAK Urban programme, funded by UK Research and Innovation’s Global Challenge Research Fund, Grant Ref: ES/P011055/1. The views expressed are those of the authors and not necessarily those of the OMS, the UK National Health Service (NHS), the NIHR or the Department of Health and Social Care.

This work used data provided by patients and collected by the NHS as

part of their care and support and would not have been possible without access to this data. The NIHR recognises and values the role of patient data, securely accessed and stored, both in underpinning and leading to improvements in research and care.

We also thank Wayne Dorrington for his help in making the graphical abstract that summarises the contents of this work, and Emma Copland for revising the first draft of this manuscript.

7. Author contributions statement

G.S.K. conceived the idea for this work; J.R.A.S worked on the data cleaning and formatting, as well as on the eNRBM, Deep Patient and RETAIN models; F.E.D.R worked on the Deepr model; Y.Z. reviewed the data pipelines and model pipelines for eNRBM, Deep Patient, Deepr, and RETAIN; J.R.A.S reported the results for the eNRBM, Deep Patient and RETAIN models; F.E.D.R reported the results for the Deepr model; J.R.A.S wrote the rest of the paper; G.S.K. and Y.Z. contributed to the discussion part of the paper. All authors contributed to multiple parts of the manuscript review, as well as the style and overall contents.

Appendix A. Abbreviations

The following abbreviations are used throughout the document:

- EHR: Electronic Health Records
- SVM: Support Vector Machines
- DL: Deep Learning

- 582 • FFNN: Feed-Forward Neural Networks
- 583 • CNN: Convolutional Neural Networks
- 584 • RNN: Recurrent Neural Networks
- 585 • eNRBM: Electronic Medical Records-driven nonnegative restricted Boltz-
- 586 mann machines
- 587 • SDA: Stacked Denoising Autoencoder
- 588 • RETAIN: Reverse Time Attention
- 589 • CPRD: Clinical Practice Research Datalink
- 590 • GP: General Practice
- 591 • ICD: International Statistical Classification of Diseases and Related
- 592 Health Problems
- 593 • ACHI: Australian Classification of Health Interventions
- 594 • ACS: Australian Coding Standard
- 595 • ATC: Anatomical Therapeutic Chemical Classification
- 596 • CPT: Current Procedural Terminology
- 597 • NDC: National Drug Codes
- 598 • GPI: Generic Product Identifier
- 599 • KCD: Korean Statistical Classification of Diseases and Related Health
- 600 Problems

- 601 • SD: Standard Deviation
- 602 • IQR: Interquartile Range
- 603 • BOW: Bag of Words
- 604 • LR: Logistic Regression
- 605 • AUROC: Area Under the Receiver Operating Characteristic Curve
- 606 • AUPRC: Area Under the Precision-Recall Curve
- 607 • ND: Not Defined
- 608 • DD: Demographics + Diagnoses
- 609 • DDM: Demographics + Diagnoses + Medications

610 **Appendix B. Emergency Admission Codes**

611 Table B.9 shows the Medcodes considered to identify emergency admis-
612 sions.

Table B.9: Medcodes considered to identify emergency admissions

Medcode	Description
140	Cauterisation of internal nose
314	Diagnostic endoscopic examination and biopsy lesion larynx
1081	Herpes simplex disciform keratitis
1158	Other operation on cornea NOS

3488	Rebound mood swings
6269	Herpesviral infection of perianal skin and rectum
6885	Congenital aortic valve stenosis
7058	Admit diabetic emergency
7059	[V]Palliative care
7242	FESS/Post operative division of adhesions
7503	[D]Heart murmur, undiagnosed
8082	Other specified other repair of vagina
8265	Closed fracture lumbar vertebra, wedge
9409	[X]Cut by glass
11413	Other finger injuries NOS
11963	Repair of chest wall
12038	[SO]Leg region
12243	[SO]Tricuspid valve
13706	Stroke / transient ischaemic attack referral
22296	Trucut transperineal biopsy of prostate
22374	Fracture of carpal bone
23106	Serum alkaline phosphatase NO
25057	Seen by ear, nose and throat surgeon
29190	Sensorineural hear loss, unilateral unrestricted hear/contralat side
29988	Gittes endoscopic bladder neck suspension
30027	Fracture of malar and maxillary bones
32898	Peritonitis - tuberculous
35328	Pressure ulcer assessment

37543	Ecstasy poisoning
38379	Mechanical complication of tendon graft
43828	Removal of onychophosis
46824	Congenital malformation of aortic and mitral valves unspecified
48729	Acute duodenal ulcer with haemorrhage and perforation
66764	Closed skull vlt no intracranial injury
67786	Encephalitis due to meningococcus
95163	Labyrinthine round window fistula
99761	Acute lower respiratory tract infection
102555	Tiabendazole poisoning

613 **Appendix C. Heart Failure Codes**

614 Table C.10 shows the ICD codes that are mapped to CCS code 108 (Con-
615 gestive Heart Failure; Nonhypertensive).

616 **Appendix D. Data: inclusion criteria and outcome**

617 *Appendix D.1. eNRBM*

618 Tran et al. [40] used a mental health cohort extracted from a large re-
619 gional hospital in Australia and collected between January 2009 and March
620 2012. Any patient who had at least one encounter with the hospital services,
621 as well as one risk assessment for suicide, was included. Each assessment
622 was considered as a data point from which a prediction would be made. The
623 sample size was 7,578 patients (49.3% male, 48.7% under 35), for a total
624 of 17,566 assessments. The future outcomes within 3 months following an

Table C.10: ICD codes mapped to CCS code 108 (Congestive Heart Failure; Nonhypertensive)

ICD	CCS	Description
I0981	108	Rheumatic heart failure
I501	108	Left ventricular failure, unspecified
I5020	108	Unspecified systolic (congestive) heart failure
I5021	108	Acute systolic (congestive) heart failure
I5022	108	Chronic systolic (congestive) heart failure
I5023	108	Acute on chronic systolic (congestive) heart failure
I5030	108	Unspecified diastolic (congestive) heart failure
I5031	108	Acute diastolic (congestive) heart failure
I5032	108	Chronic diastolic (congestive) heart failure
I5033	108	Acute on chronic diastolic (congestive) heart failure
I5040	108	Unspecified combined systolic and diastolic (congestive) heart failure
I5041	108	Acute combined systolic and diastolic (congestive) heart failure
I5042	108	Chronic combined systolic and diastolic heart failure
I5043	108	Acute on chronic combined systolic and diastolic heart failure
I50810	108	Right heart failure, unspecified
I50811	108	Acute right heart failure
I50812	108	Chronic right heart failure
I50813	108	Acute on chronic right heart failure
I50814	108	Right heart failure due to left heart failure
I5082	108	Biventricular heart failure
I5083	108	High output heart failure
I5084	108	End stage heart failure
I5089	108	Other heart failure
I509	108	Heart failure, unspecified

assessment were stratified in 3 ordinal levels of risk: no-risk, moderate-risk (non-fatal consequence), and high-risk (fatal consequence). The risk classes related to a diagnosis were decided in relation to the ICD-10 codes. If there were more than one outcome class, the highest risk class was chosen. There were 86.9% no-risk outcomes, 8.2% moderate-risk and 4.9% high-risk. Since the completed suicides were rare, the class distributions were quite imbalanced.

Appendix D.2. Deep Patient

In this work, the authors used the Mount Sinai data warehouse. All patients with at least one diagnosed disease (expressed as numerical ICD-9) between 1980 and 2014, inclusive, were considered in Miotto et al. [38]. This led to a dataset of about 1.2 million patients, with every patient having an average of 88.9 records. All patients with at least one recorded ICD-9 code were split in three independent datasets for evaluation purposes (i.e., every patient appeared in only one dataset). First, the authors retained 81,214 patients having at least one new ICD-9 diagnosis assigned in 2014 and at least ten records before that. These patients composed validation (i.e., 5,000 patients) and test (i.e., 76,214 patients) sets for the supervised evaluation (i.e., future disease prediction). In particular, all the diagnoses in 2014 were used to evaluate the predictions computed using the patient data recorded before December 31, 2013. The requirement of having at least ten records per patient was set to ensure that each test case had some minimum of clinical history that could lead to reasonable predictions. A random subset of 200,000 different patients was sampled with at least five records before the split-point (December 31, 2013) to use as training set for the disease

650 prediction experiment. Finally, the authors created the training set for the
651 unsupervised feature learning algorithms using the remaining patients having
652 at least five records by December 2013. This led to a dataset composed of
653 704,587 patients and 60,238 clinical descriptors.

654 *Appendix D.3. Deepr*

655 Data in Nguyen et al. [20] were collected from a large private hospital
656 chain in Australia from July 2011 to December 2015, using Australian Cod-
657 ing Standards (ICD-10-AM for diagnoses and the Australian Classification
658 of Health Interventions for procedures). Data consisted of 590,546 records
659 (around 300,000 patients), each corresponding to an admission (characterised
660 by an admission time and a discharge time). The outcome consisted in un-
661 planned readmission (corresponding to nearly 2% of the sample). The risk
662 group was made of patients with at least one unplanned readmission within
663 6 months (4,993 patients), or within 3 months (3,788 patients) from a dis-
664 charge, regardless of the admitting diagnosis. For each risk case, a control
665 case was randomly picked from the remaining patients. For each risk/control
666 group, 16.7% of patients were used for model tuning, 16.7% for testing, and
667 the rest for training. A discharge (except for the last one in the risk groups)
668 was randomly selected as the prediction point, from which the future risk
669 would be predicted.

670 *Appendix D.4. RETAIN*

671 In Choi et al. [45], the authors focused on heart failure prediction. The
672 dataset consisted of EHR from Sutter Health, including patients from 50 to
673 80 years of age. Diagnosis, medication and procedure codes were extracted

674 from the encounter records, medication orders, procedure orders and prob-
675 lem lists, and then aggregated into existing medical groupers so as to reduce
676 the dimensionality while preserving the clinical information of the input vari-
677 ables. From the source dataset, 3,884 cases were selected and approximately
678 10 controls were considered for each case correspondingly (28,903 controls).
679 Medical codes were extracted in the 18-month window before the baseline.
680 The patient cohort was divided into the training, validation and test sets in
681 a 0.75 : 0.1 : 0.15 ratio. The validation set was used to determine the values
682 of the hyperparameters.

683 **Appendix E. Hyperparameters**

684 Tables E.11, E.14, E.16 show the different hyperparameters for each of
685 the selected DL models together with the values identified using a tree-based
686 Bayesian Optimisation approach [69, 70], as implemented in Scikit-Optimize
687 [71]. Tables E.12 and E.13 show the corresponding hyperparameters for the
688 random forest used as the top layer classifier in the modular DL architec-
689 tures. Tables E.15 and E.17 show the corresponding hyperparameters and
690 the identified values for the baseline models.

691 **References**

- 692 [1] G. S. Birkhead, M. Klompas, N. R. Shah, Uses of Electronic Health
693 Records for Public Health Surveillance to Advance Public Health, An-
694 nual Review of Public Health 36 (1) (2015) 345–359. doi:10.1146/
695 annurev-publhealth-031914-122747.

- 696 [2] L. L. Weed, Medical records that guide and teach, *The New England*
697 *Journal of Medicine* 278 (1968) 652–657.
- 698 [3] T. Botsis, G. Hartvigsen, F. Chen, C. Weng, Secondary use of EHR:
699 data quality issues and informatics opportunities, *Summit on Transla-*
700 *tional Bioinformatics* 2010 (2010) 1.
- 701 [4] P. B. Jensen, L. J. Jensen, S. Brunak, Mining electronic health records:
702 towards better research applications and clinical care, *Nature Reviews*
703 *Genetics* 13 (6) (2012) 395.
- 704 [5] B. Shickel, P. J. Tighe, A. Bihorac, P. Rashidi, Deep EHR: A Survey
705 of Recent Advances in Deep Learning Techniques for Electronic Health
706 Record (EHR) Analysis, *IEEE Journal of Biomedical and Health Infor-*
707 *matics* 22 (5) (2018) 1589–1604. doi:10.1109/JBHI.2017.2767063.
- 708 [6] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, J. F. Hurdle,
709 Extracting information from textual documents in the electronic health
710 record: a review of recent research, *Yearbook of medical informatics*
711 17 (01) (2008) 128–144.
- 712 [7] M. Jiang, Y. Chen, M. Liu, S. T. Rosenbloom, S. Mani, J. C. Denny,
713 H. Xu, A study of machine-learning-based approaches to extract clin-
714 ical entities and their assertions from discharge summaries, *Journal of*
715 *the American Medical Informatics Association* 18 (5) (2011) 601–606.
- 716 [8] F. Doshi-Velez, Y. Ge, I. Kohane, Comorbidity Clusters in Autism
717 Spectrum Disorders: An Electronic Health Record Time-Series Analy-
718 sis, *Pediatrics* 133 (1) (2014) e54–e63. doi:10.1542/peds.2013-0819.

- 719 [9] L. Li, W.-Y. Cheng, B. S. Glicksberg, O. Gottesman, R. Tamler,
720 R. Chen, E. P. Bottinger, J. T. Dudley, Identification of type 2 diabetes
721 subgroups through topological analysis of patient similarity, *Science*
722 *Translational Medicine* 7 (311) (2015) 311ra174.
- 723 [10] S. Ebadollahi, J. Sun, D. Gotz, J. Hu, D. Sow, C. Neti, Predicting
724 patient’s trajectory of physiological data using temporal trends in sim-
725 ilar patients: a system for near-term prognostics, in: *AMIA Annual*
726 *Symposium Proceedings*, Vol. 2010, American Medical Informatics As-
727 sociation, 2010, p. 192.
- 728 [11] D. Zhao, C. Weng, Combining PubMed knowledge and EHR data to
729 develop a weighted bayesian network for pancreatic cancer prediction,
730 *Journal of Biomedical Informatics* 44 (5) (2011) 859–868.
- 731 [12] P. C. Austin, J. V. Tu, J. E. Ho, D. Levy, D. S. Lee, Using meth-
732 ods from the data-mining and machine-learning literature for disease
733 classification and prediction: a case study examining classification of
734 heart failure subtypes, *Journal of Clinical Epidemiology* 66 (4) (2013)
735 398–407.
- 736 [13] G. J. Kuperman, A. Bobb, T. H. Payne, A. J. Avery, T. K. Gandhi,
737 G. Burns, D. C. Classen, D. W. Bates, Medication-related Clinical De-
738 cision Support in Computerized Provider Order Entry Systems: A Re-
739 view, *Journal of the American Medical Informatics Association* 14 (1)
740 (2007) 29–40.
- 741 [14] R. Miotto, C. Weng, Case-based reasoning using electronic health

- 742 records efficiently identifies eligible patients for clinical trials, Jour-
743 nal of the American Medical Informatics Association 22 (e1) (2015)
744 e141–e150.
- 745 [15] F. E. Harrell Jr, K. L. Lee, R. M. Califf, D. B. Pryor, R. A. Rosati,
746 Regression modelling strategies for improved prognostic prediction,
747 Statistics in medicine 3 (2) (1984) 143–152.
- 748 [16] I. Kurt, M. Ture, A. T. Kurum, Comparing performances of logistic
749 regression, classification and regression tree, and neural networks for
750 predicting coronary artery disease, Expert Systems with Applications
751 34 (1) (2008) 366–374.
- 752 [17] R. J. Carroll, A. E. Eyler, J. C. Denny, Naïve electronic health record
753 phenotype identification for rheumatoid arthritis, in: AMIA Annual
754 Symposium Proceedings, Vol. 2011, American Medical Informatics As-
755 sociation, 2011, p. 189.
- 756 [18] J. Hippisley-Cox, C. Coupland, Predicting risk of emergency admis-
757 sion to hospital using primary care data: derivation and validation of
758 QAdmissions score, BMJ Open 3 (8).
- 759 [19] F. Rahimian, G. Salimi-Khorshidi, J. Tran, A. Payberah, J. R. Ay-
760 ala Solares, F. Raimondi, M. Nazarzadeh, D. Canoy, K. Rahimi, Pre-
761 dicting the risk of emergency hospital admissions in the general pop-
762 ulation: development and validation of machine learning models in a
763 cohort study using large-scale linked electronic health records, PLOS
764 Medicine 15 (11).

- 765 [20] P. Nguyen, T. Tran, N. Wickramasinghe, S. Venkatesh, **Deepr**: A con-
 766 volutional net for medical records, *IEEE Journal of Biomedical and*
 767 *Health Informatics* 21 (1) (2017) 22–30. doi:10.1109/JBHI.2016.
 768 2633963.
- 769 [21] T. Pham, T. Tran, D. Phung, S. Venkatesh, Predicting healthcare
 770 trajectories from medical records: A deep learning approach, *Journal of Biomedical Informatics* 69 (2017) 218 – 229. doi:https:
 771 //doi.org/10.1016/j.jbi.2017.04.001.
 772
- 773 [22] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt,
 774 P. J. Liu, X. Liu, J. Marcus, M. Sun, P. Sundberg, H. Yee, K. Zhang,
 775 Y. Zhang, G. Flores, G. E. Duggan, J. Irvine, Q. Le, K. Litsch,
 776 A. Mossin, J. Tansuwan, D. Wang, J. Wexler, J. Wilson, D. Ludwig,
 777 S. L. Volchenboum, K. Chou, M. Pearson, S. Madabushi, N. H. Shah,
 778 A. J. Butte, M. D. Howell, C. Cui, G. S. Corrado, J. Dean, Scalable
 779 and accurate deep learning with electronic health records, *npj Digital*
 780 *Medicine* 1 (1) (2018) 18. doi:10.1038/s41746-018-0029-1.
- 781 [23] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, *Nature* 521 (2015) 436
 782 EP –. doi:http://dx.doi.org/10.1038/nature14539.
- 783 [24] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, MIT Press,
 784 2016, <http://www.deeplearningbook.org>.
- 785 [25] J. Hirschberg, C. D. Manning, *Advances in Natural Language Process-*
 786 *ing*, *Science* 349 (2015) 261–266.

- 787 [26] A. Graves, A. rahman Mohamed, G. E. Hinton, Speech recognition with
788 deep recurrent neural networks, 2013 IEEE International Conference
789 on Acoustics, Speech and Signal Processing (2013) 6645–6649.
- 790 [27] C. Szegedy, S. Ioffe, V. Vanhoucke, Inception-v4, Inception-ResNet and
791 the Impact of Residual Connections on Learning, in: AAAI, 2017.
- 792 [28] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou,
793 D. Wierstra, M. A. Riedmiller, Playing Atari with Deep Reinforcement
794 Learning, Computing Research Repository abs/1312.5602.
- 795 [29] D. Ravì, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez,
796 B. Lo, G. Yang, Deep Learning for Health Informatics, IEEE Jour-
797 nal of Biomedical and Health Informatics 21 (1) (2017) 4–21. doi:
798 10.1109/JBHI.2016.2636665.
- 799 [30] D. Shen, G. Wu, H.-I. Suk, Deep Learning in Medical Image Analysis,
800 Annual Review of Biomedical Engineering 19 (2017) 221–248.
- 801 [31] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu,
802 A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams,
803 J. Cuadros, et al., Development and Validation of a Deep Learning
804 Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus
805 Photographs, JAMA 316 (22) (2016) 2402–2410.
- 806 [32] Y. Liu, K. K. Gadepalli, M. Norouzi, G. Dahl, T. Kohlberger, S. Venu-
807 gopalan, A. S. Boyko, A. Timofeev, P. Q. Nelson, G. Corrado, J. Hipp,
808 L. Peng, M. Stumpe, Detecting Cancer Metastases on Gigapixel

- 809 Pathology Images, Tech. rep., arXiv (2017).
 810 URL <https://arxiv.org/abs/1703.02442>
- 811 [33] S. Webb, Deep learning for biology, Nature.
- 812 [34] R. Poplin, P.-C. Chang, D. Alexander, S. Schwartz, T. Colthurst,
 813 A. Ku, D. Newburger, J. Dijamco, N. Nguyen, P. T. Afshar, et al.,
 814 A universal snp and small-indel variant caller using deep neural net-
 815 works, Nature Biotechnology.
- 816 [35] C. Angermueller, T. Pärnamaa, L. Parts, O. Stegle, Deep learning
 817 for computational biology, Molecular Systems Biology 12 (7). doi:
 818 10.15252/msb.20156651.
- 819 [36] T. Yue, H. Wang, Deep Learning for Genomics: A Concise Overview,
 820 arXiv.
- 821 [37] E. Gawehn, J. A. Hiss, G. Schneider, Deep Learning in Drug Discovery,
 822 Molecular Informatics 35 (1) (2016) 3–14.
- 823 [38] R. Miotto, L. Li, B. A. Kidd, J. T. Dudley, Deep Patient: An Un-
 824 supervised Representation to Predict the Future of Patients from the
 825 Electronic Health Records, Scientific Reports 6.
- 826 [39] J. Jameson, D. L. Longo, Precision medicine - personalized, problem-
 827 atic, and promising, The New England Journal of Medicine 372 (2015)
 828 2229–34.
- 829 [40] T. Tran, T. D. Nguyen, D. Phung, S. Venkatesh, Learning vector rep-
 830 resentation of medical objects via EMR-driven nonnegative restricted

- 831 Boltzmann machines (eNRBM), Journal of Biomedical Informatics 54
832 (2015) 96 – 105. doi:[https://doi.org/10.1016/j.jbi.2015.01.](https://doi.org/10.1016/j.jbi.2015.01.012)
833 012.
- 834 [41] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed
835 representations of words and phrases and their compositionality, in:
836 Advances in Neural Information Processing Systems, 2013, pp. 3111–
837 3119.
- 838 [42] E. Choi, M. T. Bahadori, E. Searles, C. Coffey, M. Thompson, J. Bost,
839 J. Tejedor-Sojo, J. Sun, Multi-layer representation learning for medical
840 concepts, in: Proceedings of the 22nd ACM SIGKDD International
841 Conference on Knowledge Discovery and Data Mining, ACM, 2016,
842 pp. 1495–1504.
- 843 [43] T. Mikolov, K. Chen, G. S. Corrado, J. Dean, Efficient Estimation of
844 Word Representations in Vector Space, Computing Research Repository.
845
- 846 [44] J. Pennington, R. Socher, C. Manning, Glove: Global Vectors for Word
847 Representation, in: Proceedings of the 2014 Conference on Empirical
848 Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–
849 1543.
- 850 [45] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, W. Stewart,
851 RETAIN: An Interpretable Predictive Model for Healthcare using Re-
852 verse Time Attention Mechanism, in: Advances in Neural Information
853 Processing Systems, 2016, pp. 3504–3512.

- 854 [46] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, J. Sun, Doc-
 855 tor AI: Predicting Clinical Events via Recurrent Neural Networks, in:
 856 Proceedings of the 1st Machine Learning for Healthcare Conference,
 857 Vol. 56 of Proceedings of Machine Learning Research, PMLR, 2016,
 858 pp. 301–318.
 859 URL <http://proceedings.mlr.press/v56/Choi16.html>
- 860 [47] E. Choi, A. Schuetz, W. F. Stewart, J. Sun, Using recurrent neural
 861 network models for early detection of heart failure onset, Journal of
 862 the American Medical Informatics Association 24 (2) (2017) 361–370.
 863 doi:10.1093/jamia/ocw112.
- 864 [48] B. C. Kwon, M.-J. Choi, J. T. Kim, E. Choi, Y. B. Kim, S. Kwon,
 865 J. Sun, J. Choo, RetainVis: Visual Analytics with Interpretable and
 866 Interactive Recurrent Neural Networks on Electronic Medical Records,
 867 IEEE Transactions on Visualization and Computer Graphics.
- 868 [49] F. Ma, R. Chitta, J. Zhou, Q. You, T. Sun, J. Gao, Dipole: Diagnosis
 869 Prediction in Healthcare via Attention-based Bidirectional Recurrent
 870 Neural Networks, in: Proceedings of the 23rd ACM SIGKDD Inter-
 871 national Conference on Knowledge Discovery and Data Mining, ACM,
 872 2017, pp. 1903–1911.
- 873 [50] E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, J. Sun, GRAM:
 874 Graph-based Attention Model for Healthcare Representation Learning,
 875 in: Proceedings of the 23rd ACM SIGKDD International Conference on
 876 Knowledge Discovery and Data Mining, KDD '17, 2017, pp. 787–795.

- 877 [51] J. T. Connor, R. D. Martin, L. E. Atlas, Recurrent neural networks and
878 robust time series prediction, *IEEE Transactions on Neural Networks*
879 5 2 (1994) 240–54.
- 880 [52] M. Schuster, K. K. Paliwal, Bidirectional Recurrent Neural Networks,
881 *IEEE Trans. Signal Processing* 45 (1997) 2673–2681.
- 882 [53] S. Hochreiter, J. Schmidhuber, Long Short-Term Memory, *Neural Com-*
883 *putation* 9 (1997) 1735–1780.
- 884 [54] Z. C. Lipton, D. C. Kale, C. Elkan, R. C. Wetzel, Learning to Di-
885 agnose with LSTM Recurrent Neural Networks, *Computing Research*
886 *Repository*.
- 887 [55] L. Rokach, Ensemble-based classifiers, *Artificial Intelligence Review* 33
888 (2009) 1–39.
- 889 [56] T. Ma, C. Xiao, F. Wang, Health-ATM: A deep architecture for mul-
890 tifaceted patient health record representation and risk prediction, in:
891 *Proceedings of the 2018 SIAM International Conference on Data Min-*
892 *ing*, SIAM, 2018, pp. 261–269.
- 893 [57] E. Herrett, A. M. Gallagher, K. Bhaskaran, H. Forbes, R. Mathur,
894 T. van Staa, L. Smeeth, Data Resource Profile: Clinical Practice Re-
895 search Datalink (CPRD), *International Journal of Epidemiology* 44 (3)
896 (2015) 827–836.
- 897 [58] Clinical Practice Research Datalink, www.cprd.com, Accessed:
898 11/September/2018.

- 899 [59] Read Codes, <https://digital.nhs.uk/services/terminology-and-classifications/read-codes>, Accessed: 19/Oc-
900 tober/2018.
- 902 [60] ICD-10 online versions, <https://www.who.int/classifications/icd/icdonlineversions/en/>, Accessed: 19/October/2018.
- 904 [61] Clinical Classifications Software (CCS) for ICD-9-CM, <https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp>, Accessed: 12/Oc-
905 tober/2018.
- 907 [62] BNF Publications, <https://www.bnf.org/>, Accessed: 12/Octo-
908 ber/2018.
- 909 [63] A. Steventon, S. Deeny, R. Friebe, T. Gardner, R. Thorlby, Briefing:
910 Emergency hospital admissions in England Which may be avoidable
911 and how?, Tech. rep., The Health Foundation (May 2018).
- 912 [64] N. Conrad, A. Judge, J. Tran, H. Mohseni, D. Hedgecott, A. P.
913 Crespillo, M. Allison, H. Hemingway, J. G. Cleland, J. J. Mc-
914 Murray, et al., Temporal trends and patterns in heart failure inci-
915 dence: a population-based study of 4 million individuals, *The Lancet*
916 391 (10120) (2018) 572–580.
- 917 [65] Reimplementation of RETAIN Recurrent Neural Network in Keras,
918 <https://github.com/Optum/retain-keras>, Accessed: 11/Octo-
919 ber/2018.
- 920 [66] A. Paszke, et al., PyTorch, <https://pytorch.org/> (2017).

- 921 [67] F. Chollet, et al., Keras, <https://keras.io> (2015).
- 922 [68] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller,
923 O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Lay-
924 ton, J. VanderPlas, A. Joly, B. Holt, G. Varoquaux, API design for ma-
925 chine learning software: experiences from the scikit-learn project, in:
926 ECML PKDD Workshop: Languages for Data Mining and Machine
927 Learning, 2013, pp. 108–122.
- 928 [69] M. Pelikan, D. E. Goldberg, E. Cantú-Paz, BOA: The Bayesian Op-
929 timization Algorithm, in: Proceedings of the 1st Annual Conference
930 on Genetic and Evolutionary Computation - Volume 1, GECCO’99,
931 Morgan Kaufmann Publishers Inc., 1999, pp. 525–532.
- 932 [70] J. Snoek, H. Larochelle, R. P. Adams, Practical Bayesian Optimiza-
933 tion of Machine Learning Algorithms, in: F. Pereira, C. J. C. Burges,
934 L. Bottou, K. Q. Weinberger (Eds.), Advances in Neural Information
935 Processing Systems, 2012, pp. 2951–2959.
- 936 [71] `skopt` module, <https://scikit-optimize.github.io/>, Accessed:
937 12/September/2018.
- 938 [72] T. Fawcett, ROC graphs: Notes and practical considerations for re-
939 searchers, Machine Learning 31 (1) (2004) 1–38.
- 940 [73] T. Fawcett, An introduction to ROC analysis, Pattern Recognition
941 Letters 27 (8) (2006) 861–874.
- 942 [74] H. He, E. A. Garcia, Learning from Imbalanced Data, IEEE Transac-
943 tions on Knowledge and Data Engineering 21 (2009) 1263–1284.

- 944 [75] T. Saito, M. Rehmsmeier, The Precision-Recall Plot Is More Informa-
945 tive than the ROC Plot When Evaluating Binary Classifiers on Imbal-
946 anced Datasets, PloS one 10 (3) (2015) e0118432.
- 947 [76] Precision-Recall (scikit-learn), [http://scikit-learn.org/stable/
948 auto_examples/model_selection/plot_precision_recall.html](http://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html),
949 Accessed: 05/November/2018.
- 950 [77] Average Precision score (scikit-learn), [http://scikit-learn.
951 org/stable/modules/generated/sklearn.metrics.average_
952 precision_score.html](http://scikit-learn.org/stable/modules/generated/sklearn.metrics.average_precision_score.html), Accessed: 05/November/2018.
- 953 [78] Y. Yuan, W. Su, M. Zhu, Threshold-free measures for assessing the
954 performance of medical screening tests, Frontiers in Public Health 3
955 (2015) 57.
- 956 [79] K. Boyd, K. H. Eng, C. D. Page, Area Under the Precision-Recall
957 Curve: Point Estimates and Confidence Intervals, in: Joint Euro-
958 pean Conference on Machine Learning and Knowledge Discovery in
959 Databases, Springer, 2013, pp. 451–466.
- 960 [80] F1-Score (scikit-learn), [https://scikit-learn.org/stable/
961 modules/generated/sklearn.metrics.f1_score.html](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html), Accessed:
962 01/June/2019.
- 963 [81] C. A. Emdin, S. G. Anderson, T. Callender, N. Conrad, G. Salimi-
964 Khorshidi, H. Mohseni, M. Woodward, K. Rahimi, Usual blood pres-
965 sure, peripheral arterial disease, and vascular risk: cohort study of 4.2
966 million adults, British Medical Journal 351 (2015) h4865.

- [82] C. A. Emdin, P. M. Rothwell, G. Salimi-Khorshidi, A. Kiran, N. Conrad, T. Callender, Z. Mehta, S. T. Pendlebury, S. G. Anderson, H. Mohseni, et al., Blood pressure and risk of vascular dementia: evidence from a primary care registry and a cohort study of transient ischemic attack and stroke, *Stroke* 47 (6) (2016) 1429–1435.
- [83] K. Rahimi, H. Mohseni, C. M. Otto, N. Conrad, J. Tran, M. Nazarzadeh, M. Woodward, T. Dwyer, S. MacMahon, Elevated blood pressure and risk of mitral regurgitation: A longitudinal cohort study of 5.5 million United Kingdom adults, *PLoS medicine* 14 (10).
- [84] CPRD Bibliography, <https://www.cprd.com/bibliography>, Accessed: 17/December/2018.
- [85] J. Davis, M. Goadrich, The relationship between precision-recall and roc curves, in: *Proceedings of the 23rd International Conference on Machine learning*, ACM, 2006, pp. 233–240.
- [86] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, *Journal of Artificial Intelligence Research* 16 (2002) 321–357.
- [87] G. Lemaître, F. Nogueira, C. K. Aridas, Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning, *Journal of Machine Learning Research* 18 (17) (2017) 1–5.
URL <http://jmlr.org/papers/v18/16-365>
- [88] L. McInnes, J. Healy, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, arXiv preprint arXiv:1802.03426.

- [89] Understanding the High Prevalence of Low Prevalence Chronic Disease Combinations: Databases and Methods for Research, <https://goog1/srsZs2>, 4/December/2018.
- [90] WHO, Introduction to Drug Utilization Research, World Health Organization, 2003.
- [91] J. Snoek, O. Rippel, K. Swersky, R. Kiros, N. Satish, N. Sundaram, M. M. A. Patwary, Prabhat, R. P. Adams, Scalable Bayesian Optimization Using Deep Neural Networks, in: ICML, 2015.
- [92] J. T. Springenberg, A. Klein, S. Falkner, F. Hutter, Bayesian Optimization with Robust Bayesian Neural Networks, in: NIPS, 2016.
- [93] R. Miikkulainen, J. Z. Liang, E. Meyerson, A. Rawal, D. Fink, O. Francon, B. Raju, H. Shahrzad, A. Navruzian, N. Duffy, B. Hodjat, Evolving Deep Neural Networks, Computing Research Repository abs/1703.00548.
- [94] J. K. Dutta, J. Liu, U. Kurup, M. Shah, Effective Building Block Design for Deep Convolutional Neural Networks using Search, Computing Research Repository abs/1801.08577.
- [95] J. Howard, et al., fastai, <https://github.com/fastai/fastai> (2018).
- [96] L. N. Smith, Cyclical Learning Rates for Training Neural Networks, 2017 IEEE Winter Conference on Applications of Computer Vision (WACV) (2017) 464–472.

- 1012 [97] I. Loshchilov, F. Hutter, SGDR: Stochastic Gradient Descent with
1013 Warm Restarts, 2016.
- 1014 [98] L. Jacobs, L. Efremov, J. P. Ferreira, L. Thijs, W.-Y. Yang, Z.-Y.
1015 Zhang, R. Latini, S. Masson, N. Agabiti, P. Sever, et al., Risk for
1016 Incident Heart Failure: A Subject-Level Meta-Analysis From the Heart
1017 "OMics" in AGEing (HOMAGE) Study, *Journal of the American Heart*
1018 *Association* 6 (5) (2017) e005231.
- 1019 [99] Y. Gal, Z. Ghahramani, Dropout as a Bayesian Approximation: Rep-
1020 resenting Model Uncertainty in Deep Learning, in: *Proceedings of The*
1021 *33rd International Conference on Machine Learning*, Vol. 48 of *Pro-*
1022 *ceedings of Machine Learning Research*, New York, New York, USA,
1023 2016, pp. 1050–1059.
- 1024 [100] Y. Gal, *Uncertainty in Deep Learning*, University of Cambridge.
- 1025 [101] D. Tran, D. Mike, M. van der Wilk, D. Hafner, Bayesian Lay-
1026 ers: A Module for Neural Network Uncertainty, *arXiv preprint*
1027 *arXiv:1812.03973*.
- 1028 [102] S. Chakraborty, R. Tomsett, R. Raghavendra, D. Harborne, M. Alzan-
1029 tot, F. Cerutti, M. Srivastava, A. Preece, S. Julier, R. M. Rao, T. D.
1030 Kelley, D. Braines, M. Sensoy, C. J. Willis, P. Gurram, Interpretabil-
1031 ity of deep learning models: A survey of results, in: *2017 IEEE*
1032 *SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted*
1033 *Computed, Scalable Computing Communications, Cloud Big Data*

1034 Computing, Internet of People and Smart City Innovation (Smart-
1035 World/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), 2017, pp. 1–6.

1036 [103] H. J. Kam, H. Y. Kim, Learning representations for the early detec-
1037 tion of sepsis with deep neural networks, *Computers in Biology and*
1038 *Medicine* 89 (2017) 248–255.

1039 [104] B. K. Reddy, D. Delen, Predicting hospital readmission for lupus pa-
1040 tients: An RNN-LSTM-based deep-learning methodology, *Computers*
1041 *in Biology and Medicine* 101 (2018) 199–209.

1042 [105] J. Chu, W. Dong, K. He, H. Duan, Z. Huang, Using neural atten-
1043 tion networks to detect adverse medical events from electronic health
1044 records, *Journal of Biomedical Informatics*.

1045 [106] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N.
1046 Gomez, L. Kaiser, I. Polosukhin, Attention Is All You Need, in: *NIPS*,
1047 2017.

1048 [107] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of
1049 Deep Bidirectional Transformers for Language Understanding, *Com-*
1050 *puting Research Repository* abs/1810.04805.

1051 [108] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee,
1052 L. Zettlemoyer, Deep contextualized word representations, in: *Proc. of*
1053 *NAACL*, 2018.

1054 [109] S. Ruder, J. Howard, Universal Language Model Fine-tuning for Text
1055 Classification, in: *ACL*, 2018.

- 1056 [110] M. Zeng, M. Li, Z. Fei, Y. Yu, Y. Pan, J. Wang, Automatic ICD-9
1057 coding via deep transfer learning, *Neurocomputing*.
- 1058 [111] Y. Wang, S. Liu, N. Afzal, M. Rastegar-Mojarad, L. Wang, F. Shen,
1059 P. Kingsbury, H. Liu, A comparison of word embeddings for the
1060 biomedical natural language processing, *Journal of Biomedical Infor-*
1061 *matics* 87 (2018) 12–20.
- 1062 [112] B. Lim, M. van der Schaar, Disease-Atlas: Navigating Disease
1063 Trajectories with Deep Learning, *Computing Research Repository*
1064 *abs/1803.10254*.
- 1065 [113] S. Lee, Natural Language Generation for Electronic Health Records,
1066 *Computing Research Repository abs/1806.01353*.

Table E.11: Hyperparameters' values for modular deep learning architectures.

Model	Hyperparameter	Scenario	
		DD	DDM
eNRBM	# hidden units	234	259
	Learning rate	0.332	0.154
	Nonnegativity cost	0.002	0.001
	Smoothness cost	0.014	0.018
	Epochs	38	20
Deep Patient	# hidden units	181	159
	# hidden layers	1	1
	Learning rate	0.018	0.014
	Noise level	0.004	0.13
	Epochs	14	27
RBM	# hidden units	214	154
	Learning rate	0.451	0.332
	Epochs	25	19

Table E.12: Hyperparameters' values for modular deep learning architectures with emergency admission as the outcome.

Model	Hyperparameter	Scenario	
		DD	DDM
eNRBM	Fraction of features to consider	0.209	0.162
	Minimum number of samples at a leaf node	1206	1363
Deep Patient	Fraction of features to consider	0.185	0.477
	Minimum number of samples at a leaf node	1136	1050
RBM	Fraction of features to consider	0.367	0.440
	Minimum number of samples at a leaf node	1225	1244

Table E.13: Hyperparameters' values for modular deep learning architectures with heart failure as the outcome.

Model	Hyperparameter	Scenario	
		DD	DDM
eNRBM	Fraction of features to consider	0.416	0.397
	Minimum number of samples at a leaf node	1153	4976
Deep Patient	Fraction of features to consider	0.493	0.396
	Minimum number of samples at a leaf node	1657	1285
RBM	Fraction of features to consider	0.467	0.286
	Minimum number of samples at a leaf node	2170	3605

Table E.14: Hyperparameters' values for end-to-end deep learning architectures with emergency admission as the outcome.

		Scenario	
Model	Hyperparameter	DD	DDM
Deepr	Filters	45	31
	Kernel size	5	3
	Fully connected units - dynamic info	25	34
	Fully connected units - static info	41	46
	Fully connected units - after merging	8	6
	Dropout - dynamic info	0.256	0.423
	Dropout - static info	0.112	0.117
	Dropout - after merging	0.114	0.439
	Positive class weight	53.022	78.120
	Learning rate	0.002	0.001
RETAIN	Embedding size	135	150
	Recurrent layer size	113	79
	Dropout index	0.333	0.482
	Dropout context	0.302	0.104
	L2 regularization	0.048	0.012
	Epochs	19	8

Table E.15: Hyperparameters' values for baseline models with emergency admission as the outcome.

		Scenario	
Model	Hyperparameter	DD	DDM
BOW + LR	ℓ_1 Regularization coefficient	1.77e-08	1.12e-06
	Positive class weight	60	34

Table E.16: Hyperparameters' values for end-to-end deep learning architectures with heart failure risk as the outcome.

Model	Hyperparameter	Scenario	
		DD	DDM
Deepr	Filters	45	43
	Kernel size	4	4
	Fully connected units - dynamic info	9	50
	Fully connected units - static info	37	24
	Fully connected units - after merging	26	25
	Dropout - dynamic info	0.244	0.406
	Dropout - static info	0.185	0.371
	Dropout - after merging	0.349	0.299
	Positive class weight	67.151	1.969
	Learning rate	0.004	0.005
RETAIN	Embedding size	193	143
	Recurrent layer size	104	116
	Dropout index	0.500	0.457
	Dropout context	0.043	0.047
	L2 regularization	0.008	0.009
	Epochs	11	14

Table E.17: Hyperparameters' values for baseline models with heart failure as the outcome.

Model	Hyperparameter	Scenario	
		DD	DDM
BOW + LR	ℓ_1 Regularization coefficient	1.99e-06	8.97e-03
	Positive class weight	97	73