<https://www.frontiersin.org/about/author-guidelines>

- Word count: ~7,500
- #figures + #tables = 10 + 2 = 12

An Explainable Multi-Modal Neural Network Architecture for Predicting Epilepsy Comorbidities Based on Administrative Claims Data

Thomas Linden [née Gerlach]¹^{a,b,c,**}, Johann de Jong^c, Chao Lu^d, Victor Kiri^d, Kathrin Haeffs^c, Holger Fröhlich^{a,b,c,*}

^aDepartment of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, Sankt Augustin, Germany

^bBonn-Aachen International Center for Information Technology (B-IT), University of Bonn, Bonn, Germany

^cUCB Biosciences GmbH, Monheim, Germany

^dUCB Ltd., Raleigh, USA

*) corresponding author: holger.froehlich@scai.fraunhofer.de

**) most of the presented work has been done while the first and last author were affiliated with UCB Biosciences GmbH, Monheim

1 Abstract

Epilepsy is a complex brain disorder characterized by repetitive seizure events. Epilepsy patients often suffer from various and severe physical and psychological co-morbidities (e.g. anxiety, migraine, stroke, etc.). While general comorbidity prevalences and incidences can be estimated from epidemiological data, such an approach does not take into account that actual patient specific risks can depend on various individual factors, including medication. This motivates to develop a machine learning approach for predicting risks of future comorbidities for the individual epilepsy patient.

In this work we use inpatient and outpatient administrative health claims data of around 19,500 US epilepsy patients. We suggest a dedicated multi-modal neural network architecture (Deep personalized **L**ongitudinal convolutional **R**isk model - DeepLORI) to predict the time dependent risk of six common comorbidities of epilepsy patients. We demonstrate superior performance of DeepLORI in a comparison with several existing methods. Moreover, we show that DeepLORI based predictions can be interpreted on the level of individual patients. Using a game theoretic approach, we identify relevant features in DeepLORI models and demonstrate that model predictions are explainable in the light of existing knowledge about the disease. Finally, we validate the model on independent data from around 97,000 patients, showing good generalization and stable prediction performance over time.

keywords: *precision medicine, personalized medicine, P4 medicine, comorbidity prediction, administrative claims data, epilepsy, neural networks, machine learning*

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

¹ Surname changed from Gerlach to Linden in 2018

2 Introduction

Epilepsy is a complex, life threatening brain disorder characterized by repetitive seizure events. Epilepsy patients often suffer from various and severe physical and psychological comorbidities, such as overweight and obesity, anxiety, migraine, bipolar disorder and cardiovascular diseases (Seidenberg, Pulsipher and Hermann, 2009; Ottman *et al.*, 2011; Keezer, Sisodiya and Sander, 2016). Some comorbidities confer a poor disease prognosis, because they complicate pharmacological treatment owing to possible drug-drug interactions and adverse events (Verrotti and Mazzocchi, 2016). The actual development of comorbidities is dependent on patient specific factors and may be modulated by anti-epileptic drug (AED) treatment (Zaccara, 2009). Early identification and treatment of comorbidities has thus been identified as highly relevant to improve the quality of life of epilepsy patients (Verrotti and Mazzocchi, 2016). However, there is a high subject to subject variability. Methods from the field of Artificial Intelligence (AI) and more specifically machine learning (ML) have the potential to predict comorbidity risks on an individual subject basis, hence fulfilling one of the promises of a more individualized patient care in the sense of precision medicine. More specifically, ML based approaches can be used to aid disease prevention by predicting the time dependent risk of an individual epilepsy patient to develop several common comorbidities in the future, such as (1) anxiety, (2) bipolar disorder & schizophrenia, (3) diabetes type 2, (4) migraine, (5) overweight & obesity, and (6) stroke & ischemic attacks.

Machine learning models to predict individualized comorbidity risks of diseases different from epilepsy have recently been published e.g. in (Dworzynski *et al.*, 2020) and (Noh *et al.*, 2020) using clinical routine data from the Danish national registry and hospital electronic health records, respectively. For epilepsy, (Glauser *et al.*, 2020) proposed an ML model for psychiatric comorbidities based on survey data from 122 patients. In our earlier work (Gerlach, Lu and Fröhlich, 2017) we proposed an ML model (Random Survival Forests) using US administrative health claims data from ~10,000 epilepsy patients to predict several major comorbidities (anxiety, bipolar disorder & schizophrenia, diabetes type 2, migraine, overweight & obesity, stroke & ischemic attacks) of epilepsy patients.

Administrative health claims data have generally been shown useful for developing ML models in the epilepsy field. For example (An *et al.*, 2018) used claims data of more than 1.3 Million epilepsy patients to predict antiepileptic drug resistance. Examples from other disease areas include prediction of Alzheimer's Disease (Park *et al.*, 2020), osteoporotic hip fractures (Engels *et al.*, 2020), and heart failure (Desai *et al.*, 2020). The opportunities of healthcare claims data for ML based modeling have further been discussed in (Fröhlich *et al.*, 2018; Miotto *et al.*, 2018; Xiao, Choi and Sun, 2018; Thesmar *et al.*, 2019; Kwak and Hui, 2020).

In our earlier work we demonstrated the possibility to augment claims data with biomedical background knowledge, hence enabling the interpretation of machine learning models down to the level of disease associated biological processes (Gerlach, Lu and Fröhlich, 2017). The particular novelty of the present work is a dedicated multi-modal neural network architecture for administrative claims data, which we call **Deep** personalized **L**ongitudinal convolutional **R**isk model (DeepLORI). We show that DeepLORI more accurately predicts the time dependent risk for six common comorbidities on the level of individual patients than several competing methods, including our own previously proposed model. Using a game theoretic approach based on Shapley Additive Explanations (Lundberg and Lee, 2017) we show that DeepLORI models are explainable, also on the level of predictions for individual patients.

3 Data

3.1 Claims Based Electronic Health Records

US commercial inpatient and outpatient data were obtained from the IBM® MarketScan® Truven Health databases. The Commercial Claims and Encounters database within MarketScan® is a nationally representative collection of de-identified patient-specific inpatient, outpatient, and pharmaceutical claims from more than 200 insurance carriers and large, self-insuring companies. All dates and timestamps were transformed from a daily to a monthly scale (1 month := 30 days) for a more robust representation. The data generally comprises demographic (age, gender) and regional information (major metropolitan area), days in hospital, health insurance plan, as well as time dependent diagnosis codes and prescriptions (plus prescription duration and quantity).

We used two cohorts, (1) the *original data* covering years 2011-2015 for model-training and –evaluation within a nested cross-validation scheme, and (2) the *external validation data* covering years 2008-2018 to validate the models trained on the “original data”. In agreement to our earlier publication (Gerlach, Lu and Fröhlich, 2017) and common practice at UCB, epilepsy patients in the original data were identified matching at least one of the following criteria:

1. an occurrence of ≥ 2 ICD-9-CM codes of 345.xx (i.e. *epilepsy*, except 345.3 – *grand mal status*) among separate medical encounters (separate dates in any care venue)
2. an occurrence of ≥ 1 ICD-9-CM code of 345.xx (except for 345.3) AND ≥ 1 ICD-9-CM code of 780.39 (*convulsions*) among separate medical encounters
3. an occurrence of 1 ICD-9-CM code of 345.xx (except for 345.3) AND code(s) for AED prescription at least a day after the 345.xx code
4. an occurrence of ≥ 2 ICD codes of 780.39 among separate medical encounters AND code(s) for AED treatment. The code(s) for the AED treatment should occur at least a day after the second 780.39 irrespective of the presence or absence of an AED code after the first 780.39 code
5. Individuals with ICD-9-CM code 345.3 will be required to have an occurrence of ≥ 2 ICD-9-CM codes of 345.3 separated by at least 30 days, or an occurrence of the 345.3 code and ≥ 1 ICD-9-CM code 780.39 separated by at least 30 days, or ≥ 1 ICD-9-CM code 345.3 and ≥ 1 ICD-9-CM code 345.xx encounters on separate days

The index date for each patient was defined as the time point of the first epilepsy diagnosis, and for definitions requiring at least 2 ICD-9-CM codes the **first** diagnosis code was the index date. The data was further filtered by requiring for each patient a) at least 365 days of medical history before, and 365 days follow up after the index date; b) age between 18 and 65; c) any AED treatment during observation period. Altogether this yielded 7,430,840 records from 19,510 patients. For part of these patients diagnoses after index date were coded in ICD10, which we mapped to ICD-9-CM via the Thomas ReutersTM public web resource² and manual curation.

Note that in medical practice, confirmation of the final diagnosis “epilepsy” can be complicated and often requires a number of visits. Moreover, reporting of a dedicated diagnosis within our data does not necessarily correspond to the actual time point of the medical condition within the patient. To capture this uncertainty we defined a three month time interval starting from the index date as „epilepsy diagnosis period“. That means the actual medical history of each patient after application of the above mentioned filter criteria was 365 + 91 days, i.e. 456 days.

Diagnosis codes after 1st Oct 2015 were provided as ICD-10-CM codes. Accordingly, the following modified inclusion criteria were applied to select epilepsy patients in the external validation data (covering 2008-2018):

² <http://www.tdrdata.com/ipd/ipd-ICD10ToICD9List>

1. The presence of at least 1 ICD-9-CM of 345.xx or ICD-10-CM of G40.xx (*epilepsy*)
2. The presence of at least 2 ICD-9-CM of 780.39 or ICD-10-CM of R56.9 (*convulsions*) within one year.

After applying the same filter criteria as for the original data this resulted in 112,755 patients. Within those patients we prefiltered diagnosis codes and substances observed in ≤ 10 patients or with a frequency $\leq 1\%$.

One of the main issues with claims data is that one and the same diagnosis may be coded with different ICD codes. Moreover, observations related to one specific ICD9/10 code could be rare. To address these issues, we mapped all ICD-9-CM codes to PheWAS terms, which describe a higher level aggregate of several ICD codes (Carroll, Bastarache and Denny, 2014). In addition, a mapping to MeSH (Rogers, 1963) was performed to allow for integration with other data sources (see Methods).

3.2 Definition of Focused Comorbidities and Compilation of Training Data

Based on the medical literature as well as observed frequency in our data we focused on six common comorbidities of epilepsy patients: (1) Anxiety, (2) Bipolar and Schizophrenia, (3) Diabetes type 2, (4) Migraine, (5) Overweight and Obesity, (6) Stroke and Ischemic Attack. These comorbidities were defined according to a set of PheWAS codes provided in the Supplements (Table S1).

The number of patients with these comorbidities being diagnosed at least 6 months after the epilepsy diagnosis period differs widely across comorbidities, see **Table 1**. We would like to highlight that our data is in principle *right censored*, i.e. the diagnosis of a specific comorbidity might happen after the end of the period covered by our training data. Moreover, a significant proportion of those individuals where a diagnosis of a specific comorbidity is observed (i.e. incident cases) have already been diagnosed with at least one of the other 5 comorbidities during their medical history, see **Table 1**. Note that for training a machine learning model to predict a specific comorbidity we should not have an observation of the same comorbidity in the medical history of any of the training samples. For this reason the number of patients in the training data is different per comorbidity, and we developed separate machine learning models for each comorbidity.

Each diagnosis and prescription in our data has an associated time stamp. Due to the fact that the appearance of a record in our data does not necessarily correspond to the observation of the actual medical condition, each time stamp was mapped to a monthly (= 30 days time interval) resolution.

comorbidity	<i>n</i> total	<i>n</i> incident	% incident	<i>mean</i> $\pm$ <i>sd</i> , no. of other comorbidities in medical history per incident patient
Anxiety	15,274	4,037	26%	1.3 $\pm$ 0.6
Bipolar / Schizophrenia	16,908	1,279	8%	1.5 $\pm$ 0.8
Diabetes type 2	16,929	1,377	8%	1.5 $\pm$ 0.8
Migraine	16,036	2,063	13%	1.3 $\pm$ 0.7
Overweight / Obesity	17,273	3,307	19%	1.4 $\pm$ 0.7
Stroke / Ischemic Attack	17,208	1,157	7%	1.5 $\pm$ 0.8

Table 1: Proportion of incident patients by comorbidity.

4 Methods

4.1 Proposed Model: DeepLORI

4.1.1 DeepLORI Architecture

As highlighted before, our aim was to develop separate machine learning models for each of 6 typical comorbidities of epilepsy patients. Each of these models aims for predicting the time dependent risk of an individual to be diagnosed with one specific comorbidity.

We came up with a dedicated neural network model for our purposes, which we call **Deep** personalized **L**ongitudinal convolution **R**isk model (DeepLORI). We start by explaining the principle architecture of DeepLORI. In agreement to our former work, one of the key ideas is that claims data has an inherent hierarchical structure (Gerlach, Lu and Fröhlich, 2017): The data initially contains three major types of features: 1) prescribed substance codes, 2) diagnoses codes (mapped to PheWAS terms, see above) and 3) general demographic information, such as age, gender and major metropolitan area information. Monthly reported prescriptions and diagnoses can typically be represented via a one-hot vector encoding. However, individual substance and diagnose codes are typically rather sparsely observed over time, which can potentially lead to challenges for a machine learning algorithm to find regularities.

Based on this consideration our idea was to use additional background knowledge available in databases to impose further hierarchical sub-structure: For example, each prescribed substance may have one or several known targets, and it can have a number of side effects reported in clinical studies. Diagnoses have associated symptoms, and in some cases biomarkers may exist. Based on this background information each domain of features (e.g. diagnosis) can be further associated to several sub-domains (e.g. biomarkers, impaired biological pathways, etc.). Sub-domain features can subsequently be represented via a one-hot vector encoding. **Figure 1** and **Table 2** provide an overview about the domains and according sub-domains we defined in our data. More details about our previously published approach to augment claims data with biomedical knowledge can be found in the Supplements.

In this work we propose a multi-modal neural network architecture to reflect the specific structure of the augmented claims data, see **Figure 2**: In this architecture, each of the feature domains and sub-domains are initially treated as separate data modalities. Note that each feature derived from diagnosis and substance codes has an additional time stamp (30 days interval), i.e. each subdomain is a three-dimensional data cube. Each of these tensors is projected down to a lower dimensional representation via a bottleneck feed-forward architecture with 1 – 4 hidden layers, where the exact number of hidden layers and units per layer are treated as tunable hyperparameters in our framework, details in Supplements. In conclusion, at the first layers for each sub-domain, specific latent features are extracted in a non-linear manner from the original data and subsequently concatenated.

After latent feature extraction, DeepLORI models the time dependency in the data within each feature domain. For that purpose we apply a pooling function (max or mean) over the entire time series, individually for each component of our one-hot-vector representations. The exact choice of the pooling function is a hyperparameter. In addition to pooling we allow for the application of different time convolutional kernels (with multiple filters per kernel), similar to a sliding window. Whether or not convolutional filters are applied, and which sizes these filters have, is again determined within hyperparameter optimization.

After modeling time dependency, there is another feedforward bottleneck structure (same tunable design as for the initial latent feature extraction) in our network. Finally, DeepLORI concatenates latent features extracted from each feature domain, and feeds them through the last feedforward bottleneck structure into one output unit, representing a patient specific comorbidity risk score. That means we have one DeepLORI model per comorbidity.

A more detailed view on the DeepLORI architecture, including an overview of all tunable hyperparameters can be found in the Supplementary material.

Figure 1: The hierarchical data model structures the data in 3 input domains (General, Drug, Diagnosis) and according subdomains. Features retrieved from the Claims Data are highlighted red, all others are augmented features (including those highlighted blue).

domain	subdomain	feature origin		#time dependent_features (per 30-days time interval)	#static_features
		claims data	background knowledge		
diagnosis	Biomarker		x	699	
diagnosis	GO Term		x	3,759	
diagnosis	MeSH disease		x	137	
diagnosis	Pathway		x	453	
diagnosis	PheWAS	x		1,196	
diagnosis	Symptoms		x	235	
general	Age, Sex, Insurance, Hospital	x			17
general	medical risk		x		53
general	Region	x			403
drug	GO Term		x	1,240	
drug	Pathway		x	435	
drug	PheWAS (known indications)		x	986	
drug	Sideeffects x Frequency		x	1,489	
drug	Substancename	x		453	
drug	Target Tissue		x	32	
drug	Therapeutic Class		x	104	
drug	Therapeutic Group		x	19	
				$\Sigma = 11,237$	$\Sigma = 473$

Table 2: Number of features by domain, subdomain and feature origin (claims data or biomedical background knowledge).

Figure 2: DeepLORI model architecture: The input is organized as a multi-modal data cube, which is sub-divided into several feature domains. Medication and diagnosis related features are time dependent, whereas features in the „General“ domain are not. The three different colours at stage (3) symbolize the 3 different pooling/convolution kernel sizes.

4.1.2 Loss Function

Let the training data denote as $D = \{(x_{it}, y_i, \delta_i) \mid x_{it} \in \mathbb{R}^{d \times T}, y_i \in \mathbb{R}, \delta_i \in \{0, 1\}, i = 1, 2, \dots, n, t = 1, 2, \dots, T\}$, where T is the maximum number of time stamps in the data (in our case 15 months) and y_i is the observed time of the first diagnosis of a comorbidity (in case the data is not censored) after index date or the maximum observed event free time. Moreover $\delta_i \in \{0, 1\}$ is a binary variable indicating whether or not y_i is right censored ($\delta_i = 0$) or not ($\delta_i = 1$). Following (Katzman *et al.*, 2018) we here use the negative partial log-likelihood of a Cox proportional hazard's model (Cox, 1972) as a loss function for training DeepLORI:

$$l(W, y) = \sum_{i=1}^n \delta_i \left[N_W(x_i) - \log \left(\sum_{j: y_j \geq y_i} \exp^{N_W(x_j)} \right) \right]$$

Here $N_W(\cdot)$ denotes the risk score learned by DeepLORI, parameterized by weights W , and x_i the medical history of patient i .

To avoid overfitting we regularize DeepLORI in multiple ways:

- We use drop-out units in the input and hidden layers.
- We perform batch normalization (Ioffe and Szegedy, 2015) before each activation function.
- We impose groupwise elastic net penalties for weights (Zou and Hastie, 2005).

The elastic net is an extension of the classical lasso algorithm (Tibshirani, 1996), which has originally been introduced in the context of generalized linear models. It combines an ℓ_1 penalty of coefficients (like in lasso regression) with an ℓ_2 penalty like in ridge regression (Hoerl and Kennard, 1970). The elastic net enforces a sparse regression model by jointly pushing coefficients towards zero via the ℓ_1 penalty, i.e. there is a feature selection. At the same time

the ℓ_2 penalty promotes a joint selection of correlated features (Zou and Hastie, 2005). The idea of the elastic net can also be extended to neural networks. More specifically, by adding groupwise elastic net penalties we modify our training objective as:

$$l(\mathbf{W}, y) + \left(\sum_{s \in S} \lambda_s^{\ell_1} \|W_s\|_1 + \lambda_s^{\ell_2} \|W_s\|_2 \right) + \lambda_D^{\ell_1} \|W_D\|_1 + \lambda_D^{\ell_2} \|W_D\|_2$$

where S the set of feature subdomains. Furthermore, $\lambda_s^{\ell_1}, \lambda_s^{\ell_2}$ denote tunable hyperparameters, W_s refers to the set of weights connecting the input to the first hidden layer within feature domain s , and W_D are the weights of the connections feeding into the output layer.

4.1.3 Hyperparameters Optimization

A comprehensive overview of hyperparameters of DeepLORI can be found in the Supplements (Table S2). We performed Bayesian Hyperparameter Optimization (Bergstra, Yamins and Cox, 2013) to tune DeepLORI on the training data. Each candidate hyperparameter set was evaluated via a 5-fold cross-validation. Hyperparameter optimization was run for 100 trials per fold, a maximum number of 100 epochs per trial, or, if the cross-validated prediction performance did not increase within 10 sequential epochs. Prediction performance was measured via Harrell's C-index (Harrell *et al.*, 1982), which is a generalization of the area under receiver operating characteristic curve (AUC), frequently used for classification models.

4.1.4 Shapley Additive Explanations (SHAP)

One of the main criticisms of neural network based approaches is the difficulty to interpret them. Recently, (Lundberg and Lee, 2017) proposed a model agnostic game theoretic framework to address this issue. Briefly, the idea behind Shapley Additive exPlanations (SHAP) is that the relevance $\Phi_i(x)$ of feature i on model output $f(x)$ can be regarded as the average weighted difference between outputs from all possible models trained on (1) all subsets S of features including feature i , against (2) all subsets of features excluding feature i :

$$\Phi_i(x) = \sum_{S \subseteq F/\{i\}} \frac{|S|! (|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)]$$

with F as the set of all features. The authors propose several local approximation techniques, which can circumvent the exact combinatorial calculation of $\Phi_i(x)$, one which is specifically tailored towards neural networks (Deep SHAP). Deep SHAP effectively combines SHAP values calculated for smaller components of a neural network into SHAP values for the entire network. We refer to (Lundberg and Lee, 2017) for details. In this work we used SHAP to understand the impact of individual feature domains, sub-domains, and AEDs on the comorbidity risk score that we learned via DeepLORI. SHAP results in a patient specific score that may be positive (feature i increases $f(x)$) or negative (feature i decreases $f(x)$) compared to the average patient. In agreement to Lundberg et al. we considered the mean absolute values of the SHAP values per feature to score the overall impact of a variable. Moreover, we investigated the overall mean absolute SHAP values per feature domain and sub-domain, respectively. This is possible, because SHAP values are additive. That means the sum of two SHAP values can indeed be interpreted as the overall impact of the corresponding features.

In practice we found Deep SHAP too computationally costly when using our entire original dataset. We thus repeatedly subsampled 5% of our data with replacement (30 times) and re-

calculated SHAP values. We checked the robustness of the approach via the variance of SHAP values.

4.2 Competing methods

We compared DeepLORI against several competing approaches:

1. Random Survival Forests (Ishwaran *et al.*, 2008): In this earlier published approach (Gerlach, Lu and Fröhlich, 2017) we first combined claims data with biomedical knowledge (akin to this paper) and then used a window of fixed length (3 months) to summarize features via a max-pooling. Features encoding prescriptions and diagnoses within such a time window were concatenated, resulting into an overall number of around 165,000 features per patient. Subsequently we used maximum relevance minimum redundancy (mRMR) (Ding and Peng, 2005) to further reduce the number of features to 500 prior to Random Survival Forest (RSF) model training. For RSF model training we relied on R-package “ranger” (Wright and Ziegler, 2017). The number of decision trees was set to 5000, and the log-rank statistic was used as a split rule for nodes.
2. Stacked denoising autoencoders (SDA) followed by training a RSF (Miotto *et al.*, 2016): In this approach initially an SDA was trained to extract features from the medical history of each patient (diagnosis and prescription codes as well as demographic information) in an unsupervised manner. The same SDA architecture as described in Miotto *et al.* was employed. After feature extraction for each of the 6 comorbidities a RSF was trained.
3. Kaplan-Meier (KM) estimator as “null model”. This approach does not use features of any individual. It only estimates the overall risk curve for a given comorbidity from the data and applies the same estimate to each patient. The purpose of this “null model” was to understand the added value of complex machine learning models.

4.3 Evaluation Approach

DeepLORI was compared with competing methods within a 5-fold cross-validation scheme using the exactly same data splits of the original dataset. Hyperparameters were only tuned on the respective training data, resulting in a nested cross-validation scheme for DeepLORI. We used Uno’s C-index (Uno *et al.*, 2011) as a performance measure:

$$\hat{C}_\tau = \frac{\sum_{i=1}^n \sum_{j=1}^n \Delta_i \{ \hat{G}(X_i)^{-2} I(X_i < X_j, X_i < \tau) I(\widehat{risk}_i > \widehat{risk}_j) \}}{\sum_{i=1}^n \sum_{j=1}^n \Delta_i \{ \hat{G}(X_i)^{-2} I(X_i < X_j, X_i < \tau) \}}$$

Uno’s C-index is a consistent estimator of the concordance index for a population that is independent of censoring. It satisfies this requirement for censored populations using two “tricks”, first by applying a “inverse probability weighting” schema using the censoring distribution estimated with $\hat{G}(\cdot)$ (e.g. the Kaplan-Meier estimator), second by evading instable tail parts for times $\geq \tau$ of the estimated survival function with a prespecified time point τ as constraint. We refer to (Uno *et al.*, 2011) for further details.

In addition, we evaluated DeepLORI on our external validation data. This validation was done separately in two different ways:

1. Follow-up of existing patients: We selected patients that had already been in our original dataset, but for which a right censoring was observed.

2. New patients: We evaluated DeepLORI on patients that had no records in the original dataset.

In both cases we recorded Uno's C-index over the entire time series and as a function of time (named $AUC(t)$ in the following) to measure the prediction performance.

5 Results

5.1 DeepLORI Outperforms Competing Methods

Despite high censoring rates all 6 comorbidities could be predicted rather accurately by DeepLORI, and the 5-fold cross-validated performance of Uno's C-index ranged from 71% for overweight and obesity up to 77% for stroke & ischemic attacks (**Figure 3, Table S3**). At the same time the Kaplan-Meier estimator (i.e. the "null"-model without any features) was consistently at chance level (50% Uno's C-index), indicating that all of our tested machine learning models (DeepPatient, DeepLORI and MRMR + RSF) extracted relevant predictive signal from the data. At the same time DeepLORI showed significantly higher C-indices than all competing methods.

Figure 3: 5-fold (nested) cross-validation test set performance benchmark of **Deep LORI** (green) vs. competing methods; **MRMR**: Minimum Redundance Maximum Relevance feature selection; **SDAE**: Stacked Denoising Autoencoder; **RSF**: Random Survival Forest; **KM**: Kaplan-Meier estimator.

5.2 DeepLORI Shows Stable Prediction Performance on External Validation Data

Evaluation of DeepLORI on the external validation data showed roughly comparable C-indices to those observed on the original data when focusing on the follow-up of the ~15,000 patients, which had medical history in the original data (**Figure 4**). This highlights that DeepLORI, despite of high censoring rates in the original data, was not overfitted. C-indices for new / so far unseen patients ($n = \sim 97,000$) in the external validation data were around 6% lower (**Table S3**).

When investigating the $AUC(t)$ we found a rather stable prediction performance for all comorbidities over time (**Figure 5**). Remarkably this held true for a time interval of up to 6 years after initial diagnosis of epilepsy, and it was true for the follow-up of existing patients as well as for new patients, again highlighting the fact that DeepLORI generalizes rather well.

Figure 4: Prediction performance of Deep LORI on external validation data. Green: time-split validation, i.e. follow up of patients included into training data. Blue: Prediction performance on new patient data. Red: 5-fold cross-validated prediction performance on original data for comparison purposes.

Figure 5: Time-dependent prediction performance (AUC(t)) of DeepLORI on external validation data.

Figure 6: Relative impact of original vs augmented features in DeepLORI models for 6 different comorbidities

5.3 DeepLORI Models are Explainable

We next investigated the relevance of feature domains via SHAP in DeepLORI models trained on the entire original dataset. This analysis highlighted for most comorbidities the relevance of features derived via augmentation of the original data with additional information (**Figure 6**): For example, in the model predicting anxiety augmented features made up around 55% of the total feature importance, for migraine we found 43%. Notably, among augmented features the “medical risk” subdomain, covering various unspecific comorbidity indices derived from ICD diagnosis codes (Charlson *et al.*, 1987; Romano, Roos and Jollis, 1993; Lee *et al.*, 1999; Schneeweiss *et al.*, 2003; Boersma *et al.*, 2005; Quan *et al.*, 2005; Sessler *et al.*, 2010; Sigakis, Bittner and Wanderer, 2013) using the R-package “medicalrisk” (McCormick and Joseph, 2020) in the medical history of epilepsy patients, had a significant impact (**Figure 7**), suggesting that the risk of developing any of our six comorbidities increases with a generally worse medical condition upfront. Importantly, none of the comorbidity indices are specific to any of the six comorbidities focused by DeepLORI.

In the following we discuss one of the six DeepLORI models in more detail, namely the one for migraine (**Figure 8**). Figures related to DeepLORI models for the other five comorbidities can be found in the Supplementary Material (**Figures S1 – S6**). According to SHAP analysis the most relevant features in the DeepLORI model for migraine relate to the prior existence of headaches and the use of drugs for the nervous system, which are typically used to treat headaches. In addition, many drugs used for treating headaches are known to affect the liver (Mathew and Klein, 2019; Valade, 2019). It is known that females are more affected by migraine than males, and that migraine is age dependent (Victor *et al.*, 2010). The anti-epileptic drug (AED) topiramate, known to be well tolerated by this group of patients (Spritzer, Bravo and Draskowski, 2016; Silberstein, 2017), ranks among the top 15 most relevant features: **Figure 9** shows the marginal dependency of DeepLORI model predictions on AED prescription frequency, suggesting that patients treated with topiramate are slightly more likely to be diagnosed with migraine later on compared to those without such treatment in the past. In fact, topiramate is often used as a preventive treatment for migraine (Spritzer, Bravo and Draskowski, 2016), suggesting that patients treated with this AED are often considered at risk of developing migraine by their treating physician. Indeed, many of these patients eventually receive this diagnosis.

Another interesting finding from the SHAP analysis of our model is the influence of disorders of the lipid metabolism on migraine risk. Associations between lipid levels and migraine have been reported in (Rist, Tzourio and Kurth, 2011) and (Onderwater *et al.*, 2019). Moreover, the metabolic syndrome and migraine have been associated to each other (Sachdev and Marmura, 2012). It is important to highlight at this point that SHAP analysis does not provide a causal explanation, though.

To further exemplify the possibility of interpreting our DeepLORI models on the level of individual patients we depict in **Figure 10**. SHAP values for two randomly selected patients with high and low risk for developing stroke or ischemic attacks, respectively. As expected, the low risk patient is young and has no diagnosis of hypertension or disorder of the metabolic system. In contrast, the high risk patient is an older person with hypertension who lives in Texas. In fact, significant regional differences in the risk for strokes have been reported throughout the US (Howard *et al.*, 2007), and Eastern Texas belongs to the so-called „Stroke Belt“ (Karp David N. *et al.*, 2016).

Figure 7: SHAP values by feature subdomain. Subdomains with a cumulative relative importance $\leq 5\%$ are not shown.

Figure 8: SHAP values of top 15 most relevant features in the DeepLORI model for migraine, rank by mean absolute values. x-axis: SHAP values, higher values = higher importance. y-axis: features ranked by importance, annotated with their domain and subdomain membership (domain > subdomain > feature),

Figure 9: The plot shows the marginal dependency of the frequency of topiramate prescription on the predicted risk on later migraine diagnosis. One dot represents one patient; x-axis: prescription frequency (0% = never observed, 100% = observed in every month of a patient's medical history); y-axis: change in SHAP-value (=change in individual hazard rate). Marginal dependency plots for each comorbidity for the top-5 features can be found in the Supplemental Material (Figures S7-S12).

SHAP individual explanations for Stroke, Ischemic Attack

Figure 10: Examples of SHAP explanations for a patient predicted at low risk for stroke & ischemic attacks (left) and a high risk patient (right). Red (green) bars indicate higher (lower) risk compared to the average patient: The hazard ratio³ of the low risk patient (left) is around 56% of that of an average patient ($\frac{\exp(-17.29)}{\exp(-16.71)} = 0.56$). The hazard ratio of the high risk patient (right) is around 2.94 times that of an average patient ($\frac{\exp(-15.63)}{\exp(-16.71)} = 2.94$).

6 Discussion and Conclusions

Precision medicine has the vision to bring the right treatment to the right patients. Precision medicine is strongly dependent on machine learning. At present precision medicine is only an

³ Hazard ratios computed similar to Cox proportional hazard's model (Cox, 1972).

emerging reality. Several reasons can be identified (Fröhlich *et al.*, 2018; Miotto *et al.*, 2018; Xiao, Choi and Sun, 2018; Kwak and Hui, 2020): 1) lack of the right data in sufficient quantity; 2) insufficient validation; and 3) difficulties in interpreting complex ML models, which is by itself a prerequisite for generating the necessary confidence in using such models. Realization of precision medicine will only be possible, if all these aspects are addressed jointly. In this context it is essential that ML models can be used in a cost effective and practical manner. Hence, clinical routine data are of extreme relevance and are gaining more and more attention (Weiss *et al.*, 2012; Peissig *et al.*, 2014; Miotto *et al.*, 2016; Choi *et al.*, 2016; Rajkomar *et al.*, 2018; Harutyunyan *et al.*, 2019). Administrative claims data constitute an important source of such clinical routine data. They principally exist in large quantities and allow for obtaining insights into the longitudinal medical history of individual patients under real world conditions. However, these data have not been collected for research purposes. First of all, coding of diagnoses into ICD codes is not unique and mostly done for maximizing economic reasons rather than for providing a precise medical description. Different ICD codes can be used for similar diagnoses, and the relationship between different medical conditions is consequently not always uniquely resolvable from their distances in the ICD ontology. Second, it should be noted that ICD only reflects the medical symptom level, which should not be confused with the biological relationship between disorders. Third, the time of diagnosis encoding might not correspond to the actual appearance of the medical condition. Fourth, it is unclear whether patients take the prescribed medication. Finally, the nature of irregular time series data, different for each patient, imposes specific challenges for data analysis.

In this work we tried to address these challenges by a) mapping ICD codes to PheWAS codes that are at higher granularity; b) augmenting the original data with further information from biological databases; and c) proposing a specific multi-modal neural network architecture (DeepLORI). We demonstrated that DeepLORI can predict six common comorbidities of epilepsy patients with higher C-index than several competing methods. We performed a rigorous cross-validation plus an external validation to assess our model, demonstrating that DeepLORI allows for reliable predictions of comorbidity risks up to six years in advance. We showed that with the help of SHAP and our data augmentation approach it is possible to make DeepLORI based predictions explainable, even on the level of individual patients. From our perspective this is of great importance for generating confidence in ML based solutions in medicine.

From a medical perspective we see the value of our work in the potential for much earlier identification of epilepsy patients at risk of developing different comorbidities. For example, a patient at high risk of developing diabetes type 2 should consider losing weight and regularly check insulin levels. A patient at high risk of developing psychiatric disorders might consider early consultation with a psychiatrist. Hence, risk models could be a way to eventually move towards preventive medicine.

Further applications of our work could lie in addressing the high subject-to-subject variability in epilepsy: Based on the comorbidity risk profile learned by DeepLORI one might be able to identify subgroups of patients with more homogenous disease progression, potentially opening up opportunities for developing more personalized therapies in the future.

7 Abbreviations

AED:	Anti Epileptic Drug
SDA:	Stacked Denoising Autoencoder
MRMR:	Minimum Redundance Maximum Relevance
RSF:	Random Survival Forest

8 Acknowledgements

We like to thank Linda Kalilani, Babak Boroerjerdi for helpful discussions and support of the entire project.

9 Conflict of Interest

All authors have received salaries from UCB Pharma S.A. during the runtime of this project. UCB had no influence on the scientific content of this work.

10 Author Contributions

Designed and supervised the project: HF; extracted the data: CL, HK, KH; analyzed the data: TL, JdJ; performed experiments: TL; interpreted the data and results: TL, HF; drafted the manuscript: TL, HF. All authors have read the manuscript and agreed to its content.

11 References

An, S. *et al.* (2018) ‘Early Prediction of Drug Resistant Epilepsy using Claims Data’, *Epilepsy & behavior : E&B*, 89, pp. 118–125. doi: 10.1016/j.yebeh.2018.10.013.

Bergstra, J., Yamins, D. and Cox, D. D. (2013) ‘Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures’, in *Proceedings of the 30th International Conference on Machine Learning - Volume 28*. Atlanta, GA, USA: JMLR.org (ICML’13), p. I–115–I–123.

Boersma, E. *et al.* (2005) ‘Perioperative cardiovascular mortality in noncardiac surgery: validation of the Lee cardiac risk index’, *The American Journal of Medicine*, 118(10), pp. 1134–1141. doi: 10.1016/j.amjmed.2005.01.064.

Carroll, R. J., Bastarache, L. and Denny, J. C. (2014) ‘R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment’, *Bioinformatics*, 30(16), pp. 2375–2376. doi: 10.1093/bioinformatics/btu197.

Charlson, M. E. *et al.* (1987) ‘A new method of classifying prognostic comorbidity in longitudinal studies: development and validation’, *Journal of Chronic Diseases*, 40(5), pp. 373–383. doi: 10.1016/0021-9681(87)90171-8.

Choi, E. *et al.* (2016) ‘Doctor AI: Predicting Clinical Events via Recurrent Neural Networks’, *arXiv:1511.05942 [cs]*. Available at: <http://arxiv.org/abs/1511.05942>.

Cox, D. R. (1972) ‘Regression Models and Life-Tables’, *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2), pp. 187–220.

Desai, R. J. *et al.* (2020) ‘Comparison of Machine Learning Methods With Traditional Models for Use of Administrative Claims With Electronic Medical Records to Predict Heart Failure Outcomes’, *JAMA Network Open*, 3(1), pp. e1918962–e1918962. doi: 10.1001/jamanetworkopen.2019.18962.

- Ding, C. and Peng, H. (2005) ‘Minimum redundancy feature selection from microarray gene expression data’, *Journal of Bioinformatics and Computational Biology*, 3(2), pp. 185–205. doi: 10.1142/s0219720005001004.
- Dworzynski, P. *et al.* (2020) ‘Nationwide prediction of type 2 diabetes comorbidities’, *Scientific Reports*, 10(1), p. 1776. doi: 10.1038/s41598-020-58601-7.
- Engels, A. *et al.* (2020) ‘Osteoporotic hip fracture prediction from risk factors available in administrative claims data – A machine learning approach’, *PLOS ONE*, 15(5), p. e0232969. doi: 10.1371/journal.pone.0232969.
- Fröhlich, H. *et al.* (2018) ‘From hype to reality: data science enabling personalized medicine’, *BMC Medicine*, 16(1), p. 150. doi: 10.1186/s12916-018-1122-7.
- Gerlach, T., Lu, C. and Fröhlich, H. (2017) ‘Predicting comorbidities of epilepsy patients using big data from Electronic Health Records combined with biomedical knowledge’, in *Proceedings of the GCB German Conference on Bioinformatics 2017. GCB 2017*, Tübingen, Germany: PeerJ Inc. doi: 10.7287/peerj.preprints.3228v1.
- Glauser, T. *et al.* (2020) ‘Identifying epilepsy psychiatric comorbidities with machine learning’, *Acta Neurologica Scandinavica*, 141(5), pp. 388–396. doi: 10.1111/ane.13216.
- Harrell, F. E. *et al.* (1982) ‘Evaluating the yield of medical tests’, *JAMA*, 247(18), pp. 2543–2546.
- Harutyunyan, H. *et al.* (2019) ‘Multitask learning and benchmarking with clinical time series data’, *Scientific Data*, 6(1), p. 96. doi: 10.1038/s41597-019-0103-9.
- Hoerl, A. E. and Kennard, R. W. (1970) ‘Ridge Regression: Biased Estimation for Nonorthogonal Problems’, *Technometrics*, 12(1), pp. 55–67. doi: 10.1080/00401706.1970.10488634.
- Howard, G. *et al.* (2007) ‘Regional Differences in African Americans’ High Risk for Stroke: The Remarkable Burden of Stroke for Southern African Americans’, *Annals of epidemiology*, 17(9), pp. 689–696. doi: 10.1016/j.annepidem.2007.03.019.
- Ioffe, S. and Szegedy, C. (2015) ‘Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift’, *arXiv:1502.03167 [cs]*. Available at: <http://arxiv.org/abs/1502.03167>.
- Ishwaran, H. *et al.* (2008) ‘Random Survival Forests’, *The Annals of Applied Statistics*, 2. doi: 10.1214/08-AOAS169.
- Karp David N. *et al.* (2016) ‘Reassessing the Stroke Belt’, *Stroke*, 47(7), pp. 1939–1942. doi: 10.1161/STROKEAHA.116.012997.
- Katzman, J. *et al.* (2018) ‘DeepSurv: Personalized Treatment Recommender System Using A Cox Proportional Hazards Deep Neural Network’, *BMC Medical Research Methodology*, 18(1). doi: 10.1186/s12874-018-0482-1.
- Keezer, M. R., Sisodiya, S. M. and Sander, J. W. (2016) ‘Comorbidities of epilepsy: current concepts and future perspectives’, *The Lancet Neurology*, 15(1), pp. 106–115. doi: 10.1016/S1474-4422(15)00225-2.

Kwak, G. H. and Hui, P. (2020) ‘DeepHealth: Review and challenges of artificial intelligence in health informatics’, *arXiv:1909.00384 [cs, eess, stat]*. Available at: <http://arxiv.org/abs/1909.00384>.

Lee, T. H. *et al.* (1999) ‘Derivation and prospective validation of a simple index for prediction of cardiac risk of major noncardiac surgery’, *Circulation*, 100(10), pp. 1043–1049. doi: 10.1161/01.cir.100.10.1043.

Lundberg, S. and Lee, S.-I. (2017) ‘A Unified Approach to Interpreting Model Predictions’, *arXiv:1705.07874 [cs, stat]*. Available at: <http://arxiv.org/abs/1705.07874>.

Mathew, P. G. and Klein, B. C. (2019) ‘Getting to the Heart of the Matter: Migraine, Triptans, DHE, Ditans, CGRP Antibodies, First/Second-Generation Gepants, and Cardiovascular Risk’, *Headache*. doi: 10.1111/head.13601.

McCormick, P. and Joseph, T. (2020) *medicalrisk: Medical Risk and Comorbidity Tools for ICD-9-CM Data*. Available at: <https://CRAN.R-project.org/package=medicalrisk>.

Miotto, R. *et al.* (2016) ‘Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records’, *Scientific Reports*, 6, p. 26094. doi: 10.1038/srep26094.

Miotto, R. *et al.* (2018) ‘Deep learning for healthcare: review, opportunities and challenges’, *Briefings in Bioinformatics*, 19(6), pp. 1236–1246. doi: 10.1093/bib/bbx044.

Noh, J. *et al.* (2020) ‘Prediction of the Mortality Risk in Peritoneal Dialysis Patients using Machine Learning Models: A Nation-wide Prospective Cohort in Korea’, *Scientific Reports*, 10(1), p. 7470. doi: 10.1038/s41598-020-64184-0.

Onderwater, G. L. J. *et al.* (2019) ‘Large-scale plasma metabolome analysis reveals alterations in HDL metabolism in migraine’, *Neurology*, 92(16), pp. e1899–e1911. doi: 10.1212/WNL.00000000000007313.

Ottman, R. *et al.* (2011) ‘Comorbidities of epilepsy: Results from the Epilepsy Comorbidities and Health (EPIC) Survey: Epilepsy Comorbidities: The EPIC Survey’, *Epilepsia*. doi: 10.1111/j.1528-1167.2010.02927.x.

Park, J. H. *et al.* (2020) ‘Machine learning prediction of incidence of Alzheimer’s disease using large-scale administrative health data’, *npj Digital Medicine*, 3(1), pp. 1–7. doi: 10.1038/s41746-020-0256-0.

Peissig, P. L. *et al.* (2014) ‘Relational Machine Learning for Electronic Health Record-Driven Phenotyping’, *Journal of biomedical informatics*, 52, pp. 260–270. doi: 10.1016/j.jbi.2014.07.007.

Quan, H. *et al.* (2005) ‘Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data’, *Medical Care*, 43(11), pp. 1130–1139. doi: 10.1097/01.mlr.0000182534.19832.83.

Rajkomar, A. *et al.* (2018) ‘Scalable and accurate deep learning with electronic health records’, *npj Digital Medicine*, 1(1), pp. 1–10. doi: 10.1038/s41746-018-0029-1.

- Rist, P. M., Tzourio, C. and Kurth, T. (2011) 'Associations between lipid levels and migraine: cross-sectional analysis in the epidemiology of vascular ageing study', *Cephalalgia: An International Journal of Headache*, 31(14), pp. 1459–1465. doi: 10.1177/0333102411421682.
- Rogers, F. B. (1963) 'Medical subject headings', *Bulletin of the Medical Library Association*, 51, pp. 114–116.
- Romano, P. S., Roos, L. L. and Jollis, J. G. (1993) 'Adapting a clinical comorbidity index for use with ICD-9-CM administrative data: differing perspectives', *Journal of Clinical Epidemiology*, 46(10), pp. 1075–1079; discussion 1081–1090. doi: 10.1016/0895-4356(93)90103-8.
- Sachdev, A. and Marmura, M. J. (2012) 'Metabolic Syndrome and Migraine', *Frontiers in Neurology*, 3. doi: 10.3389/fneur.2012.00161.
- Schneeweiss, S. *et al.* (2003) 'Improved comorbidity adjustment for predicting mortality in Medicare populations', *Health Services Research*, 38(4), pp. 1103–1120. doi: 10.1111/1475-6773.00165.
- Seidenberg, M., Pulsipher, D. T. and Hermann, B. (2009) 'Association of epilepsy and comorbid conditions', *Future neurology*, 4(5), pp. 663–668. doi: 10.2217/fnl.09.32.
- Sessler, D. I. *et al.* (2010) 'Broadly Applicable Risk Stratification System for Predicting Duration of Hospitalization and Mortality', *Anesthesiology: The Journal of the American Society of Anesthesiologists*, 113(5), pp. 1026–1037. doi: 10.1097/ALN.0b013e3181f79a8d.
- Sigakis, M. J. G., Bittner, E. A. and Wanderer, J. P. (2013) 'Validation of a risk stratification index and risk quantification index for predicting patient outcomes: in-hospital mortality, 30-day mortality, 1-year mortality, and length-of-stay', *Anesthesiology*, 119(3), pp. 525–540. doi: 10.1097/ALN.0b013e31829ce6e6.
- Silberstein, S. D. (2017) 'Topiramate in Migraine Prevention: A 2016 Perspective', *Headache*, 57(1), pp. 165–178. doi: 10.1111/head.12997.
- Spritzer, S. D., Bravo, T. P. and Drazkowski, J. F. (2016) 'Topiramate for Treatment in Patients With Migraine and Epilepsy', *Headache*, 56(6), pp. 1081–1085. doi: 10.1111/head.12826.
- Thesmar, D. *et al.* (2019) 'Combining the Power of Artificial Intelligence with the Richness of Healthcare Claims Data: Opportunities and Challenges', *Pharmacoeconomics*, 37(6), pp. 745–752. doi: 10.1007/s40273-019-00777-6.
- Tibshirani, R. (1996) 'Regression Shrinkage and Selection via the Lasso', *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1), pp. 267–288.
- Uno, H. *et al.* (2011) 'On the C-statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival Data', *Statistics in medicine*, 30(10), pp. 1105–1117. doi: 10.1002/sim.4154.
- Valade, D. (2019) 'New horizons for acute and prophylactic treatments of migraine', *Biologie Aujourd'hui*, 213(1–2), pp. 59–64. doi: 10.1051/jbio/2019021.

Verrotti, A. and Mazzocchetti, C. (2016) ‘Beyond seizures — the importance of comorbidities in epilepsy’, *Nature Reviews Neurology*, 12(10), pp. 559–560. doi: 10.1038/nrneurol.2016.135.

Victor, T. *et al.* (2010) ‘Migraine prevalence by age and sex in the United States: A life-span study’, *Cephalalgia*, 30(9), pp. 1065–1072. doi: 10.1177/0333102409355601.

Weiss, J. C. *et al.* (2012) ‘Machine Learning for Personalized Medicine: Predicting Primary Myocardial Infarction from Electronic Health Records’, *AI Magazine*, 33(4), pp. 33–33. doi: 10.1609/aimag.v33i4.2438.

Wright, M. N. and Ziegler, A. (2017) ‘ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R’, *Journal of Statistical Software*, 77(1). doi: 10.18637/jss.v077.i01.

Xiao, C., Choi, E. and Sun, J. (2018) ‘Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review’, *Journal of the American Medical Informatics Association*, 25(10), pp. 1419–1428. doi: 10.1093/jamia/ocy068.

Zaccara, G. (2009) ‘Neurological comorbidity and epilepsy: implications for treatment’, *Acta Neurologica Scandinavica*, 120(1), pp. 1–15. doi: 10.1111/j.1600-0404.2008.01146.x.

Zou, H. and Hastie, T. (2005) ‘Regularization and variable selection via the elastic net’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), pp. 301–320. doi: 10.1111/j.1467-9868.2005.00503.x.

Hierarchical Data Model

DeepLORI Model Architecture

DeepLORI model performance: original data VS external validation data

Time dependent prediction performance

SHAP relative importance by feature origin

comorbidity

SHAP relative importance by subdomain

domain

subdomain

feature importance, Migraine
SHAP values boxplot 30x bootstrap

Migraine

rank = 11

importance = 0.089, relative importance = 2.0 %

SHAP individual explanations for Stroke, Ischemic Attack

low risk patient

Predicted risk score: -17.29

Average risk score: -16.71

high risk patient

Predicted risk score: -15.63

Average risk score: -16.71