Predicting Comorbidities of Epilepsy Patients Using Big Data from Electronic Health Records Combined With Biomedical Knowledge

Thomas Gerlach[1], Chao Lu[3], and Holger Fröhlich[*1,2]

[1]**UCB Biosciences GmbH, Monheim, Germany**
[2]**University of Bonn, Bonn-Aachen International Center for IT, Life Sciene Data Analytics and Algorithmic Bioinformatics, Germany**
[3]**UCB Ltd., Raleigh, USA**

ABSTRACT

Epilepsy is a complex brain disorder characterized by repetitive seizure events. Epilepsy patients often suffer from various and severe physical and psychological comorbidities. While general comorbidity prevalence 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 individual comorbidities. To address these needs we used Big Data from electronic health records (˜100 Million raw observations), which provide a time resolved view on an individual's disease and medication history. A specific contribution of this work is an integration of these data with information from 14 biomedical sources (DisGeNET, TTD, KEGG, Wiki Pathways, DrugBank, SIDER, Gene Ontology, Human Protein Atlas, ...) to capture putative biological effects of observed diseases and applied medications. In consequence we extracted >165,000 features describing the longitudinal patient journey of >10,000 adult epilepsy patients. We used maximum-relevance-minimum-redundancy feature selection in combination with Random Survival Forests (RSF) for predicting the risk of 9 major comorbidities after first epilepsy diagnosis with high cross-validated C-indices of 76 - 89% and analyzed the influence of medications on the risk to develop specific comorbidities. Altogether we see our work as a first step towards earlier detection and better prevention of common comorbidities of epilepsy patients.
Supplementary material: https://drive.google.com/file/d/0B4OhgVPeWvGTeUNFQVJLai1HRlk/view?usp=sharing, code: https://github.com/thomasmooon/GCB2017

Keywords: electronic health records, machine learning, comorbidity prediction

1 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, anxiety, depression, bipolar disorder and cardiovascular diseases (Seidenberg et al., 2009; Ottman et al., 2011; Keezer et al., 2016). Some comorbidities confer a poor disease prognosis, because they complicate pharmacological treatment owing to possible drug-drug interactions and adverse events (Verrotti and Mazzocchetti, 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 Mazzocchetti, 2016).

The aim of this work was to start addressing these needs by building a machine learning model that allows for predicting comorbidity risks of epilepsy patients based on their past time resolved clinical history. More specifically, we here employed Big Data from Electronic Health Records (EHR) of more than 10,000 adult epilepsy patients. The overall potential of EHR data for biomedical research has been e.g. discussed in Jensen et al. (2012). Population level statistical associations of comorbidities to epilepsy using EHR data have been investigated in Sajatovic et al. (2015). Applications of machine learning techniques to EHR data for different medical questions can e.g. be found in Weiss et al.

[*]corresponding author: holger.froehlich@ucb.com

(2012); Peissig et al. (2014); Miotto et al. (2016); Choi et al. (2016). However, to our knowledge nothing exists so far to predict epilepsy comorbidities. Furthermore, previous work purely relies on the phenotypic information that is extract-able from EHR data, hence missing relevant aspects on the molecular level. A specific contribution of our work is thus to integrate patient medication and diagnosis information extracted from insurance claims based EHRs with various biomedical databases that capture putative biological effects of observed diseases and applied drugs. To achieve this goal we downloaded and combined data from 14 data sources and constructed more than 165,000 features out of the 2 years medical history of each individual patient and then used maximum-relevance-minimum-redundancy feature selection in combination with Random Survival Forests (RSF) (Ishwaran et al., 2008) for predicting the risk of 9 major comorbidities after first epilepsy diagnosis. Our work demonstrates the principal feasibility of using claims based EHRs covering a limited portion of patient history to predict comorbidity risks of epilepsy patients and to better understand the influence of individual medications in that context.

2 METHODS

2.1 Claims Based Electronic Health Records

US commercial inpatient and outpatient data covering the years 2011 - 2015 were obtained from the TruvenHealth MarketScan™ databases[1]. 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. Within these data epilepsy patients were identified according to the fulfillment of at least one of the following criteria:

- an occurrence of $\geq$2 ICD-9-CM codes of 345.xx (except 345.3) among separate medical encounters (separate dates in any care venue)
- an occurrence of $\geq$1 ICD-9-CM code of 345.xx (except for 345.3) AND $\geq$1 ICD-9-CM code of 780.39 among separate medical encounters
- 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
- an occurrence of $\geq$ 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
- Individuals with ICD-9-CM code 345.3 will be required to have an occurrence of $\geq$2 ICD-9-CM codes of 345.3 separated by at least 30 days, or an occurrence of the 345.3 code and $\geq$1 ICD-9-CM code 780.39 separated by at least 30 days, or $\geq$1 ICD-9-CM code 345.3 and $\geq$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 resulting dataset had 97,999,822 records from 526,923 patients. The data was further filtered by requiring for each patient a) at least 2 years of medical history before and 180 days follow up after index date; b) age between 18 and 66; c) any AED treatment during observation period. Altogether this yielded 12,225,388 records from 30,807 patients. For parts of these patients diagnoses after index date were coded in ICD10, which we mapped to ICD-9-CM via the Thomas Reuters™ public web resource [2] and manual curation.

One of the main issues with EHR 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 et al., 2014). In addition a mapping to MeSH (Rogers, 1963) was performed to allow for integration with other data sources (see below).

[1] https://truvenhealth.com/markets/life-sciences/products/data-tools/marketscan-databases
[2] http://www.tdrdata.com/ipd/ipd_ICD10ToICD9List

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3228v1 | CC BY 4.0 Open Access | rec: 6 Sep 2017, publ: 6 Sep 2017

2.2 Definition of Focused Comorbidities

Based on the indicated prevalence in epilepsy patients in the medical literature (Seidenberg et al., 2009; Ottman et al., 2011; Tellez-Zenteno et al., 2007) as well as frequency in our data we considered the following 9 comorbidities with exact PheWAS term definitions given in Table S1:

1. anxiety
2. bipolar disorder & schizophrenia
3. depression
4. diabetes
5. hyperlipidemia
6. hypertension
7. migraine
8. overweight
9. stroke & ischemic attack

Our aim was to predict the time till first diagnosis of these comorbidities after first epilepsy diagnosis. To reduce the risk that first diagnoses of these comorbidities would correspond to an actual situation that existed before index date (but was not reported there in our data) we required first comorbidity diagnoses to appear at least 180 days after index date.

2.3 Integration with Biomedical Knowledge Sources

We used DisGeNET (Piñero et al., 2017) to retrieve for each reported diagnosis in our data disease associated genes. Enrichment of Gene Ontology (GO) biological processes (The Gene Ontology Consortium, 2004), KEGG (Kanehisa et al., 2008) and Wiki pathways was then estimated via a conditional hyper-geometric test using GOstats (Falcon and Gentleman, 2007) with a tail area based false discovery rate (FDR - Strimmer, 2008) cutoff of 5% for GO and 20% for pathways. Furthermore, known disease biomarkers and symptoms were obtained from the Therapeutic Target Database (TTD) (Yang et al., 2016) and the human symptoms-disease network (Zhou et al., 2014). Similarly, medications in the EHR data were mapped to known targets via TTD and DrugBank (Wishart et al., 2006) via text matching of substance names, and information about tissue expression of drug targets was obtained from the Human Protein Atlas (Uhlén et al., 2015). Potential side effects of drugs and their likelihoods were retrieved from SIDER (Kuhn et al., 2016), again by application of text matching of substance names. We obtained information about likely indication areas of drugs from MEDI (Wei et al., 2013). Notably, only indication areas that were marked as "high confidence" by the MEDI authors and were mentioned in at least 10 PubMed articles were used. Finally, medications were mapped to substance classes and groups via the RED BOOK™ database. Figure 1 shows an overview about the integration process. Further details are described in the supplements.

2.4 Feature Matrix Construction Using Longitudinal Information

Applicaton of machine learning algorithms usually require the specification of a numeric feature matrix. In our case these features one one hand described general patient characteristics (e.g. age, gender, region, days in hospital). On the other hand our aim was also to capture - at least partially - the longitudinal nature of the medical history of each patient. We therefore constructed for each individual patient a feature vector for every 3 months in the 2 years medical history before index date plus 3 months after index date. The reason for including information from the 3 months after index date was that patients at the time of their first epilepsy diagnosis (and possibly directly after) typically visit their physician very frequently, and thus diagnoses within that time period can be assumed to be highly informative about the actual disease state of the patient at index date. According to our integration with several biological data sources features constructed for every 3 months time period covered diagnosis related features (MeSH term, PheWAS diagnosis group, enriched disease KEGG and Wiki pathways and GO biological processes, biomarkers, symptoms) as well as medication properties (total quantity of prescribed substance, substance class, putative targeted pathways, biological processes and tissues, potential side effects, likely indication areas). Table S2 shows an overview about extracted features together with their source of information. Altogether there were about 18,400 features for each of the 9 quarters. Hence, there were about 165,000 extracted features per patient in total.

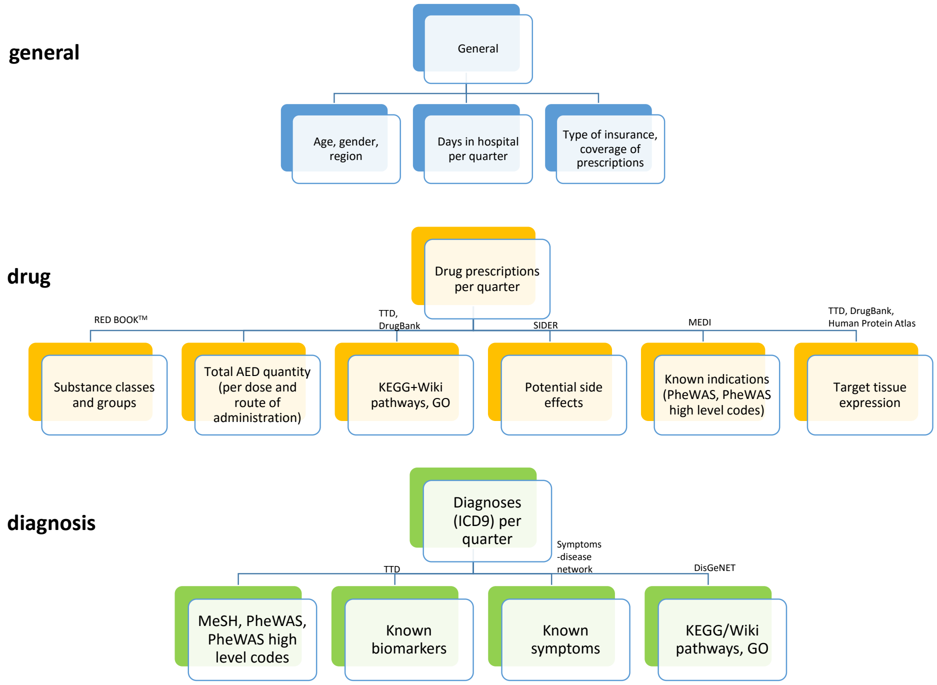

Figure 1. EHR data contains general patient characteristics, prescriptions and diagnosis codes. With the help of external information sources from prescriptions and diagnosis codes a number of interpret-able features can be extracted.

2.5 Machine Learning

Our aim was to predict for each individual patient the time till first comorbidity occurrence at least 180 days after index date. This can be seen as an instance of a time-to-event modeling problem with censoring, because patients may develop the comorbidity after the end of the observation period. The reason for us to start predicting comorbidities 180 days after index date and not before was that reported events in the database very close to the index date are possibly a result of an intensive medical examination of patients at time of their epilepsy diagnosis. Predicting comorbidities during that time period is thus likely to be overoptimistic and not very helpful for medical decision support.

We used Random Survival Forests (RSF) (Ishwaran et al., 2008) for predicting comorbidities based on the implementation in the R-package "ranger" (Wright and Ziegler, 2017). RSF are an extension of the well known Random Forest algorithm (Breiman, 2001) for time-to-event models and are particularly well suited for applications to large scale, highly heterogenous data, as in our situation. Furthermore, RSF are a non-parametric approach and do not, for example, rely on the proportional hazards assumption. Finally, the RSF implementation provided in "ranger" is scalable to Big Data due to the possibility of parallelization of the RSF algorithm.

The data in our application is extremely high dimensional. Off-the-shelf application of RSF to these data would not only yield severe computational issues, but also result into overfitting. Therefore, we first removed features appearing in less than 5% of the patients as non-zero. Afterwards, we reduced the number of input features further via the maximum-relevance-minimum-redundancy (MRMR) feature selection algorithm before RSF training (Ding and Peng, 2005). MRMR feature selection was chosen, because the algorithm accounts for the high level of redundancy among the entire set of 165,000 features that we expect in our application. MRMR selects only those features for which the correlation with respect to the clinical outcome (measured via Harrol's C-index – Harrell et al. 1982) is higher than the average correlation among all features. The maximum number of selected features was restricted to 500 here in order to make RSF training computationally feasible and to reduce overfitting risk.

The number of decision trees for the RSF was set to 5000, and the log-rank statistic was used as a split rule for nodes. Note that in general more decision trees provide a higher robustness due to lower variance of the model. However, the maximum number of decision trees is at the same time limited by RAM and computation time - specifically in this application with Big Data. The number of variables to possibly split at each node (mtry) was set to the square root of the total number of

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3228v1 | CC BY 4.0 Open Access | rec: 6 Sep 2017, publ: 6 Sep 2017

focused comorbidity	C-index (%)	C-index (SD)	IBS	IBS (SD)
Anxiety	76.0	0.9	14.9	0.6
Bipolar, Schizophrenia	85.4	1.0	10.6	0.5
Depression	78.2	1.0	14.1	0.4
Diabetes	86.6	1.4	9.9	0.7
Hyperlipidemia	79.8	1.0	13.4	0.2
Hypertension	80.7	1.9	13.6	0.8
Migraine	81.6	1.2	12.6	0.5
Overweight	80.6	1.0	12.8	0.5
Stroke, Ischemic Attack	88.4	1.4	7.8	0.4

Table 1. Cross-validated C-indices and integrated Brier scores (mean and standard deviation) for predicting time till comorbidity diagnosis after index date.

features (which is the typical default choice). Training of one single MRMR plus RSF combination took around 10 - 20 hours on an Intel Xeon E5-2697 v2 machine with 2.70GHz (which could be reduced via parallel computing to around 1 hour). Considering this aspect together with the large number of patients in our data we decided to run a single, stratified 6-fold cross-validation to evaluate prediction performance. The stratification was done such that within the cross-validation loop always approximately the same fraction of patients had any reported comorbidity diagnosis. Afterwards feature selection (filtering plus MRMR) and RSF training was performed within the cross-validation loop for each individual comorbidity, i.e. there were 9 different RSF models trained within each cross-validation loop (i.e. 54 RSF models in total).

To investigate relevance of individual features at the end we trained a final feature selection plus RSF combination for each of the 9 focused comorbidities based on all available data and investigated feature importance according to a permutation test (Ishwaran et al., 2008). Note that this step particularly allowed to select features from a subset of the MRMR selected features.

3 RESULTS

3.1 Overview about Patient Cohort

7,868 patients had PheWAS terms associated to any of our focused comorbidities at least 180 days after index date. The number of patients with these comorbidities varied from 2090 (anxiety) to 521 (stroke & ischemic attacks), see Table S2 for more details. 2783 patients had no reported comorbidity within at least 180 days of follow up after index date and were thus viewed as censored. Figure S1 depicts Kaplan-Meier curves for each of these comorbidities, indicating largely differing risks: While, for example, 1200 days after index date ˜35% of patients suffered from an anxiety disorder, less than 10% had a stroke or ischemic attack.

3.2 Comorbidities Can be Predicted with High Accuracy

Table 1 shows the overall high prediction performance of our RSF models in terms of C-index (Harrell et al., 1982), a generalization of the AUC used in binary classification for time-to-event models, and integrated Brier score (IBS) for each of the 9 focused main comorbidities. The IBS is defined according to Graf et al. (1999):

$$IBS = \frac{1}{T}\int_0^T \sum_{i=1}^{n} (\delta_i(t) - p_i(t))^2$$

where T is the maximal event (i.e. comorbidity occurrence) time in the dataset, δ_i indicates a censoring indicator of event time t for patient i, and p_i is the predicted event probability. We used the implementation provided in R-package "pec" (Mogensen et al., 2012).

In addition we investigated the dependency of the prediction error (Brier score) as a function of time (Figures S2 - S10). The analysis demonstrates a low prediction error of our models, which was significantly below that of the Kaplan-Meier estimator (which does not use any feature information) and stayed almost constant until 1200 days (i.e. more than 3 years) after index date.

3.3 Most Influential Features Reflect Literature Known and Discussed Associations

One of the interesting aspects of RSF is the ability to extract feature importance via a permutation test of single variables (Ishwaran et al., 2008), reflecting the decrease or increase of out-of-bag

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3228v1 | CC BY 4.0 Open Access | rec: 6 Sep 2017, publ: 6 Sep 2017

error prediction accuracy. To aid further understanding we grouped features into different categories (disease, drug, general), domains (e.g. disease biomarkers) and sub-domains (e.g. fibrinogen) and analyzed the statistical over-representation of feature groups among all variables with a positive importance (i.e. probable association to the clinical outcome) via a hyper-geometric test (Tables S4 - S12). This generally demonstrated a high enrichment (according to false discovery rate control with statistical test dependency – Benjamini, Y. and Yekutieli, D. 2001) and cumulative importance of features that were derived from biomedical background knowledge, i.e. were not originally encoded in our EHR data, in all our models (Table S13 - S15). Moreover, AEDs, further prescribed substance features (including therapeutic group and/or class as well as potential side effects), diagnoses related features (including disease symptoms), the number of days spent in hospital, sex and region were over-represented feature groups in models for all comorbidities. Moreover, Tables S4 - S12 demonstrate that features falling into these respective domains were selected highly consistently during the cross-validation procedure.

When analyzing features of prescribed substances in more detail (Table S16 - S24) we found a relatively high cumulated influence of several AEDs in our models predicting anxiety and bipolar disorder. This included, for example, Gabapentin, which is the most frequently prescribed anticonvulsant in our data (Figure S11). While some authors have discussed a possible psychotropic effect of AEDs (Nadkarni and Devinsky, 2005), this finding could also be explained by the fact that Gabapentin is not only used as anticonvulsant, but also to treat various psychiatric disorders (Berlin et al., 2015). Treatment with Gabapentin may thus slightly increase the prior chance to be diagnosed later e.g. with a bipolar disorder (Figure 2), even if there is not necessarily a causal link. Gabapentin related features had also a relatively high cumulative importance in our models for hyperlipidemia, diabetes migraine and overweight. A link between Gabapentin and hyperlimidemia has been reported in Gore et al. (2007). Gabapentin has been suggest to treat diabetes induced neuropathic pain (Gorson et al., 1999) and has been recommended for migraine prophylaxis (Mathew et al., 2001). On the other hand Gabapentin has been reported to promote weight gain (Ness-Abramof and Apovian, 2005).

In general, we found that *before* the first diagnosis of a certain comorbidity patients had sometimes prescriptions of substances that are in clinical use to treat these comorbidities. For example, patients later diagnosed with anxiety occasionally had prescriptions of substances that can be used as anxiolytics before. Consequently features related these drugs were also over-represented in the corresponding RSF model. This finding has to be interpreted carefully, because many substances (including AEDs) are not only used for one specific indication area, but for several ones. Hence, prescription of a substance that has - among other effects - also an anxiolytic one does not necessarily imply an anxiety disorder at time of prescription.

For several comorbidities we also found over-representations of specific drug related GO terms and molecular pathways: For example, for diabetes we found *steroid biosynthesis*. Indeed it has been discussed that extensive steroid treatment could induce diabetes (Hwang and Weiss, 2014). *PPAR(γ) signaling* is known to play an important role in the type-2 diabetes disease mechanism (Ahmadian et al., 2013). *Adrenergic signaling in cardiomyocytes* has been suggested to cross-talk with insulin dependent signaling and contribute to the pathophysiology of insulin resistance in diabetes (Fu and Zhou, 2013). The involvement of *endocannabinoid signaling* in the pathophysiology of behavioral disorders (including bipolar disorder and schizophrenia) has been discussed in Karhson et al. (2016).

To further understand the role of feature groups across different comorbidities Figure 3 highlights whether variables from a particular domain or sub-domain were found among the 10 most important features in each comorbidity (together with the stability of the selection): This demonstrates, for example, the literature known role of *fibrinogen* as a biomarker for cardiovascular diseases and ischemic stroke (van Holten et al., 2013) as well as its discussed association to depression (Martins-de Souza et al., 2014). The involvement of *MAPK signaling* into cardiovascular disease mechanisms and specifically hypertension is described in Muslin (2008) and reflected by our models. Our models in turn predict an association of essential hypertension with all remaining 8 comorbidities, which is e.g. discussed for the non-obvious link with anxiety in Pan et al. (2015) and other mood disorders (including depression and bipolar disorder) in Boal et al. (2016). Drug based intervention into the *carboxylic acid transport* is seen as a possible therapy for depression (Gao et al., 2013), which is again well reflected by our model. Also the non-obvious links between anorexia and migraine, and vertigo (dizziness) and diabetes have been discussed in the literature (D'Andrea et al., 2009; Walley et al., 2014).

Altogether our analysis demonstrates that our models essentially capture a multitude of literature known or discussed associations between comorbidities and previous diagnoses, potential side effects as well as treatments.

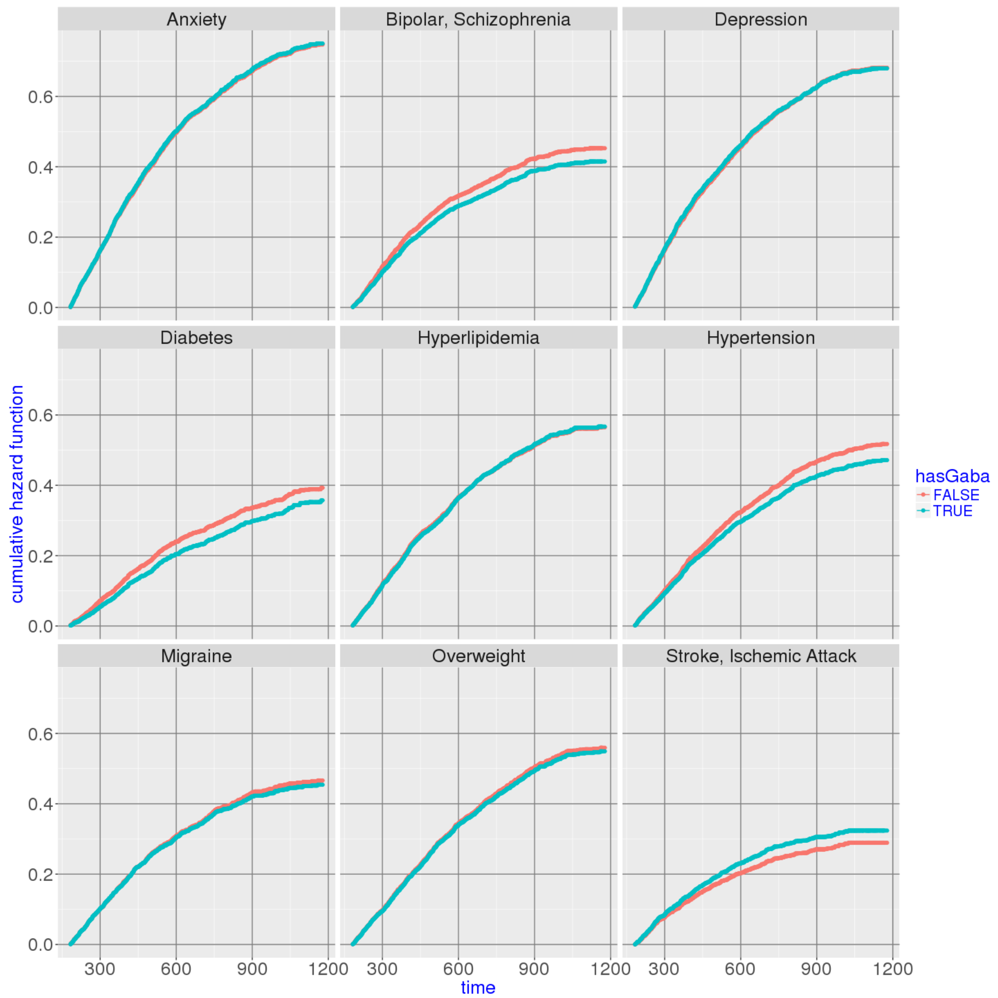

Figure 2. Predicted influence of Gabapentin on comorbidity risk: Shown are the cumulative hazard functions predicted by our RSF models for patients that have been treated with Gabapentin and those who were not.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3228v1 | CC BY 4.0 Open Access | rec: 6 Sep 2017, publ: 6 Sep 2017

domain.type	domain	subdomain	Anxiety	Bipolar, Schizophrenia	Depression	Diabetes	Hyperlipidemia	Hypertension	Migraine	Overweight	Stroke, Ischemic Attack
DISEASE	biomarker	B.CELL.CHRONIC.LYMPHOCYTIC.LEUKEMIA.LYMPHOMA.7B.BCL7B.				x (0)					
DISEASE	biomarker	CREATININE									x (1)
DISEASE	biomarker	CYSTATIN.C	x (1)								
DISEASE	biomarker	FIBRINOGEN			x (5)	x (6)				x (3)	x (4)
DISEASE	biomarker	HOMOCYSTEINE							x (3)		
DISEASE	biomarker	P.SELECTIN							x (5)		x (6)
DISEASE	biomarker	VASCULAR.CELL.ADHESION.MOLECULE							x (0)		
DISEASE	GO	CARBOXYLIC.ACID.TRANSPORT			x (3)						
DISEASE	GO	CELL.MIGRATION									x (2)
DISEASE	GO	CELLULAR.RESPONSE.TO.ENDOGENOUS.STIMULUS		x (5)							
DISEASE	GO	GLYCEROLIPID.METABOLIC.PROCESS									x (3)
DISEASE	GO	MORPHOGENESIS.OF.AN.EPITHELIUM				x (1)					
DISEASE	GO	NEGATIVE.REGULATION.OF.TRANSPORT						x (6)			
DISEASE	GO	NUCLEOSIDE.MONOPHOSPHATE.METABOLIC.PROCESS			x (3)						
DISEASE	GO	PURINE.RIBONUCLEOSIDE.TRIPHOSPHATE.METABOLIC.PROCESS		x (2)							
DISEASE	GO	STRIATED.MUSCLE.TISSUE.DEVELOPMENT				x (6)					
DISEASE	MeSH Code	ESSENTIAL HYPERTENSION	x (6)	x (3)	x (4)		x (6)		x (6)		x (5)
DISEASE	pathway	P38.MAPK.SIGNALING.PATHWAY						x (4)			
DISEASE	pathway	STATIN.PATHWAY					x (2)				
DISEASE	PheWAS Code	ANXIETY DISORDER		x (5)							
DISEASE	PheWAS Code	DEPRESSION		x (6)						x (6)	
DISEASE	PheWAS Code	ESSENTIAL HYPERTENSION	x (6)		x (5)	x (3)	x (6)		x (6)	x (6)	
DISEASE	PheWAS Code	HYPERLIPIDEMIA	x (6)		x (6)					x (6)	
DISEASE	PheWAS Code	MIGRAINE			x (4)						
DISEASE	PheWAS Code	ANXIETY, PHOBIC AND DISSOCIATIVE DISORDERS		x (6)	x (6)		x (6)	x (6)	x (6)	x (6)	
DISEASE	PheWAS Code	DISORDERS OF LIPOID METABOLISM	x (6)			x (6)		x (6)	x (6)	x (6)	x (6)
DISEASE	PheWAS Code	HYPERTENSION					x (6)				
DISEASE	PheWAS Code	MIGRAINE	x (5)								
DISEASE	PheWAS Code	MOOD DISORDERS	x (6)	x (6)		x (5)	x (6)	x (6)	x (6)		x (6)
DRUG	therapeutic group	CARDIOVASCULAR.AGENTS						x (6)			x (5)
DRUG	side effect	ANOREXIA							x (2)		
DRUG	side effect	VERTIGO				x (4)					

Figure 3. Analysis of 10 most relevant features for each comorbidity model: Variables were assigned to domains and sub-domains to aid understanding. An "x" indicates that at least one feature falling into the listed sub-domain was part of the 10 most relevant variables. In addition, the frequency (at most 6) is shown by which at least one feature of the respective sub-domain has been selected during the 6-fold cross-validation procedure.

4 CONCLUSION

Our work demonstrates the principal feasibility of combining claims based EHR data with biomedical background information to predict comorbidities of epilepsy patients. Our RSF models for 9 comorbidities obtained high prediction performances and low prediction errors over time. Our analysis of most influential features and feature groups reflected a number of literature known or discussed comorbidity associations, including medications. However, these associations do not necessarily correspond to causal mechanisms - as highlighted by the Gabapentin example, where previous treatment may just increase the prior chance be diagnosed with a psychiatric comorbidity later due to the indication areas covered by the drug. Further research is thus required to disect truely causal from non-causal associations.

The Gabapentin example also highlights that the primary EHR data used in this work has a number of inherent limitations and challenges: EHR data is purely phenotypic and only reflects causal mechanisms up to a limited degree. ICD diagnosis codes in these data can be ambiguous, may be incorrect, and may not correspond to the exact time of the disease occurrence. The period of observation for each single patient only covers a time period of at most 5 years, and provided data within that time period may not even be entirely complete. Compliance of patients to take a prescribed drug is principally unknown. Identification of patients with a complex disease such as epilepsy via ICD codes is by itself not trivial (Moura et al., 2017). We tried to address these challenges by a number of strategies:

1. by a conservative approach to identify epilepsy patients in our EHR data
2. by integrating knowledge from a large number of biomedical data sources in order to capture putative biological effects of diseases and treatments
3. by mapping ICD to the higher level PheWAS codes and aggregating features over certain time intervals
4. by using a relatively robust, ensemble based machine learning technique, such as RSF.

Notably, we investigated the time dependent prediction error of our models in order to ensure that the overall high prediction performance reflected by the C-index was not solely due to a precise prediction close to index date, where the risk would be higher that the diagnosis reported in the claims data reflected a disease state in the past of the patient.

Altogether we see our work as a first step towards earlier detection and better prevention of common comorbidities of epilepsy patients via machine learning techniques. Accordingly, there is a need for follow-up research to optimize the methodology that is presented here.

SUPPLEMENTARY MATERIAL

Supplementary material is accessible via https://drive.google.com/file/d/0B4OhgVPeWvGTeUNFQVJLai1HRlk/view?usp=sharing and code via https://github.com/thomasmooon/GCB2017.

ACKNOWLEDGMENTS

We thank Linda Kalilani and Babak Boroojerdi for helpful discussions and support of the entire research project.

REFERENCES

Ahmadian, M., Suh, J. M., Hah, N., Liddle, C., Atkins, A. R., Downes, M., and Evans, R. M. (2013). PPARγ signaling and metabolism: The good, the bad and the future. *Nature Medicine*, 99(5):557–566.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29:1165–1188.

Berlin, R. K., Butler, P. M., and Perloff, M. D. (2015). Gabapentin Therapy in Psychiatric Disorders: A Systematic Review. *The Primary Care Companion for CNS Disorders*, 17(5).

Boal, A. H., Smith, D. J., McCallum, L., Muir, S., Touyz, R. M., Dominiczak, A. F., and Padmanabhan, S. (2016). Monotherapy With Major Antihypertensive Drug Classes and Risk of Hospital Admissions for Mood Disorders. *Hypertension (Dallas, Tex.: 1979)*, 68(5):1132–1138.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

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 (Oxford, England)*, 30(16):2375–2376.

Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., and Sun, J. (2016). Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. In *PMLR*, pages 301–318.

D'Andrea, G., Ostuzzi, R., Francesconi, F., Musco, F., Bolner, A., d'Onofrio, F., and Colavito, D. (2009). Migraine prevalence in eating disorders and pathophysiological correlations. *Neurological Sciences: Official Journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology*, 30 Suppl 1:S55–59.

Ding, C. and Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology*, 3(2):185–205.

Falcon, S. and Gentleman, R. (2007). Using GOstats to test gene lists for GO term association. *Bioinformatics*, 23(2):257–258.

Fu, F. and Zhou, Q. (2013). Learning sparse causal Gaussian networks with experimental intervention: Regularization and coordinate descent. *Journal of the American Statistical Association*, 108(501):288–300.

Gao, Z., Hurst, W. J., Czechtizky, W., Hall, D., Moindrot, N., Nagorny, R., Pichat, P., Stefany, D., Hendrix, J. A., and George, P. G. (2013). Identification and profiling of 3,5-dimethyl-isoxazole-4-carboxylic acid [2-methyl-4-((2S,3'S)-2-methyl-[1,3']bipyrrolidinyl-1'-yl)phenyl] amide as histamine H(3) receptor antagonist for the treatment of depression. *Bioorganic & Medicinal Chemistry Letters*, 23(23):6269–6273.

Gore, M., Sadosky, A., Tai, K.-S., and Stacey, B. (2007). A retrospective evaluation of the use of gabapentin and pregabalin in patients with postherpetic neuralgia in usual-care settings. *Clinical Therapeutics*, 29(8):1655–1670.

Gorson, K. C., Schott, C., Herman, R., Ropper, A. H., and Rand, W. M. (1999). Gabapentin in the treatment of painful diabetic neuropathy: A placebo controlled, double blind, crossover trial. *Journal of Neurology, Neurosurgery & Psychiatry*, 66(2):251–252.

Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17-18):2529–2545.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. (1982). Evaluating the yield of medical tests. *JAMA*, 247(18):2543–2546.

Hwang, J. L. and Weiss, R. E. (2014). Steroid-induced diabetes: A clinical and molecular approach to understanding and treatment. *Diabetes/metabolism research and reviews*, 30(2):96–102.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860.

Jensen, P. B., Jensen, L. J., and Brunak, S. (2012). Mining electronic health records: Towards better research applications and clinical care. *Nature Reviews. Genetics*, 13(6):395–405.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., and Yamanishi, Y. (2008). KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, 36:480–484.

Karhson, D. S., Hardan, A. Y., and Parker, K. J. (2016). Endocannabinoid signaling in social functioning: An RDoC perspective. *Translational Psychiatry*, 6(9):e905.

Keezer, M. R., Sisodiya, S. M., and Sander, J. W. (2016). Comorbidities of epilepsy: Current concepts and future perspectives. *The Lancet. Neurology*, 15(1):106–115.

Kuhn, M., Letunic, I., Jensen, L. J., and Bork, P. (2016). The SIDER database of drugs and side effects. *Nucleic Acids Research*, 44(Database issue):D1075–D1079.

Martins-de Souza, D., Maccarrone, G., Ising, M., Kloiber, S., Lucae, S., Holsboer, F., and Turck, C. W. (2014). Plasma fibrinogen: Now also an antidepressant response marker? *Translational Psychiatry*, 4(1):e352.

Mathew, N. T., Rapoport, A., Saper, J., Magnus, L., Klapper, J., Ramadan, N., Stacey, B., and Tepper, S. (2001). Efficacy of gabapentin in migraine prophylaxis. *Headache*, 41(2):119–128.

Miotto, R., Li, L., Kidd, B. A., and Dudley, J. T. (2016). Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Scientific Reports*, 6:26094.

Mogensen, U. B., Ishwaran, H., and Gerds, T. A. (2012). Evaluating Random Forests for Survival Analysis using Prediction Error Curves. *Journal of statistical software*, 50(11):1–23.

Moura, L. M. V. R., Price, M., Cole, A. J., Hoch, D. B., and Hsu, J. (2017). Accuracy of claims-based algorithms for epilepsy research: Revealing the unseen performance of claims-based studies. *Epilepsia*, 58(4):683–691.

Muslin, A. J. (2008). MAPK Signaling in Cardiovascular Health and Disease: Molecular Mechanisms and Therapeutic Targets. *Clinical science (London, England : 1979)*, 115(7):203–218.

Nadkarni, S. and Devinsky, O. (2005). Psychotropic Effects of Antiepileptic Drugs. *Epilepsy Currents*, 5(5):176–181.

Ness-Abramof, R. and Apovian, C. M. (2005). Drug-induced weight gain. *Drugs of Today (Barcelona, Spain: 1998)*, 41(8):547–555.

Ottman, R., Lipton, R. B., Ettinger, A. B., Cramer, J. A., Reed, M. L., Morrison, A., and Wan, G. J. (2011). Comorbidities of epilepsy: Results from the Epilepsy Comorbidities and Health (EPIC) survey. *Epilepsia*, 52(2):308–315.

Pan, Y., Cai, W., Cheng, Q., Dong, W., An, T., and Yan, J. (2015). Association between anxiety and hypertension: A systematic review and meta-analysis of epidemiological studies. *Neuropsychiatric Disease and Treatment*, 11:1121–1130.

Peissig, P. L., Costa, V. S., Caldwell, M. D., Rottscheit, C., Berg, R. L., Mendonca, E. A., and Page, D. (2014). Relational Machine Learning for Electronic Health Record-Driven Phenotyping. *Journal of biomedical informatics*, 52:260–270.

Piñero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., García-García, J., Sanz, F., and Furlong, L. I. (2017). DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research*, 45(D1):D833–D839.

Rogers, F. B. (1963). Medical subject headings. *Bulletin of the Medical Library Association*, 51:114–116.

Sajatovic, M., Welter, E., Tatsuoka, C., Perzynski, A. T., and Einstadter, D. (2015). Electronic medical record analysis of emergency room visits and hospitalizations in individuals with epilepsy and mental illness comorbidity. *Epilepsy & Behavior*, 50:55–60.

Seidenberg, M., Pulsipher, D. T., and Hermann, B. (2009). Association of epilepsy and comorbid conditions. *Future neurology*, 4(5):663–668.

Strimmer, K. (2008). A unified approach to false discovery rate estimation. *BMC Bioinformatics*, 9:303.

Tellez-Zenteno, J. F., Patten, S. B., Jetté, N., Williams, J., and Wiebe, S. (2007). Psychiatric Comorbidity in Epilepsy: A Population-Based Analysis. *Epilepsia*, 48(12):2336–2344.

The Gene Ontology Consortium (2004). The Gene Ontology ({GO}) database and informatics resource. *Nucleic Acids Research*, 32:D258–D261.

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C. A.-K., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., Alm, T., Edqvist, P.-H., Berling, H., Tegel, H., Mulder, J., Rockberg, J., Nilsson, P., Schwenk, J. M., Hamsten, M., von Feilitzen, K., Forsberg, M., Persson, L., Johansson, F., Zwahlen, M., von Heijne, G., Nielsen, J., and Pontén, F. (2015). Proteomics. Tissue-based map of the human proteome. *Science (New York, N.Y.)*, 347(6220):1260419.

van Holten, T. C., Waanders, L. F., de Groot, P. G., Vissers, J., Hoefer, I. E., Pasterkamp, G., Prins, M. W. J., and Roest, M. (2013). Circulating biomarkers for predicting cardiovascular disease risk; a systematic review and comprehensive overview of meta-analyses. *PloS One*, 8(4):e62080.

Verrotti, A. and Mazzocchetti, C. (2016). Epilepsy: Beyond seizures — the importance of comorbidities in epilepsy. *Nature Reviews Neurology*, 12(10):559–560.

Walley, M., Anderson, E., Pippen, M. W., and Maitland, G. (2014). Dizziness and Loss of Balance in Individuals With Diabetes: Relative Contribution of Vestibular Versus Somatosensory Dysfunction. *Clinical Diabetes : A Publication of the American Diabetes Association*, 32(2):76–77.

Wei, W.-Q., Cronin, R. M., Xu, H., Lasko, T. A., Bastarache, L., and Denny, J. C. (2013). Development and evaluation of an ensemble resource linking medications to their indications. *Journal of the American Medical Informatics Association: JAMIA*, 20(5):954–961.

Weiss, J. C., Natarajan, S., Peissig, P. L., McCarty, C. A., and Page, D. (2012). Machine Learning for Personalized Medicine: Predicting Primary Myocardial Infarction from Electronic Health Records. *AI Magazine*, 33(4):33.

Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., and Woolsey, J. (2006). DrugBank: A comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research*, 34(Database issue):D668–672.

Wright, M. and Ziegler, A. (2017). Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software, Articles*, 77(1):1–17.

Yang, H., Qin, C., Li, Y. H., Tao, L., Zhou, J., Yu, C. Y., Xu, F., Chen, Z., Zhu, F., and Chen, Y. Z. (2016). Therapeutic target database update 2016: Enriched resource for bench to clinical drug target and targeted pathway information. *Nucleic Acids Research*, 44(D1):D1069–1074.

Zaccara, G. (2009). Neurological comorbidity and epilepsy: Implications for treatment. *Acta Neurologica Scandinavica*, 120(1):1–15.

Zhou, X., Menche, J., Barabási, A.-L., and Sharma, A. (2014). Human symptoms–disease network. *Nature Communications*, 5:4212.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3228v1 | CC BY 4.0 Open Access | rec: 6 Sep 2017, publ: 6 Sep 2017