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

Thomas Gerlach, Chao Lu, Holger Fröhlich

Feature Extraction from SIDER

For each potential side effect a vector encoding describing the likelihood of that effect in terms of 5 categories (very rare, rare, uncommon, common, very common) was created. This was done as follows: SIDER provides side effect frequencies for drugs in three different formats:

A) intervals (e. g. *0 to 17%*, *0-2%*)

B) percentages

C) frequency categories (e. g. *common*, *rare*, ...)

For robustness reasons our aim was to map all side effects to ordinal categories. Obviously cases A) and B) impose a challenge in that context, which we addressed as follows: Intervals (case A) were parsed and mean effect frequencies extracted. We then applied the WHO definition of frequency of adverse drug reactions associated intervals[1] to map these frequencies to one of the defined categories:

1) *Very rare*: <0.001%

2) *Rare*: [0.01%,0.1%)

3) *Uncommon*: [0.1%,1%)

4) *Common*: [1%,10%)

5) *Very common*: ≥ 10%

Finally we counted for each particular side effect the number of prescriptions within a quarter that could induce a rare, common, etc. effect and generated a corresponding vector, for example:

Headache common	Headache rare	Headache very common
3	1	2

Feature Extraction from DisGeNET

DisGeNET contains about 161,000 confidence scored gene-disease-associations for about 12,400 genes and 1,100 diseases based on various manually curated databases and text-mining derived associations. Since we were only interested in strong disease associations, we took only results into account, which were either manually curated or, if based on text-mining, had a score greater or equal than the smallest confidence score of any manually curated result AND was mentioned in at least 10 publications. This approach resulted into 13,168 associations between 4,512 genes and 537 diseases.

[1] http://www.who.int/medicines/areas/quality_safety/safety_efficacy/trainingcourses/definitions.pdf

Feature Extraction from MEDI

Table MEDI_HPS.csv downloaded from MEDI[2] contains about 148,000 disease-symptoms associations for about 4,220 diseases and 320 symptoms. We only considered those which occurred in at least 10 PubMed articles. This results in 20,086 associations for 2,597 diseases and 309 symptoms.

Features Derived from EHR Data

Total AED quantity per quarter

The aggregated physiological effect of a particular AED depends on the prescribed quantity (e. g. number of tablets or drops), the route of administration (e. g. oral, sublingual, rectal) and the dose (e.g. 50mg). Claims data from Truven contains for each prescription start and stop dates, day supply and prescribed quantity, which allows to derive the daily drug quantity as the ratio between prescribed quantity and day supply. Hence, for a given drug with defined route of administration and dose it is possible to aggregate the daily quantity to a quarterly quantity. Hence, there was one numeric feature per drug, route of administration, dose and quarter.

Table S1: Definition of comorbidities according to PheWAS terms:

Comorbidity	**PheWAS**
Anxiety	Anxiety disorder / Generalized anxiety disorder / Anxiety, phobic and dissociative disorders / Agorophobia, social phobia, and panic disorder
bipolar disorder & schizophrenia	bipolar / schizophrenia and other psychotic disorders
depression	Depression / Major depressive disorder
diabetes	Type 2 diabetes / Diabetes mellitus
hyperlipidemia	Hyperlipidemia / Mixed hyperlipidemia
hypertension	Essential hypertension
migraine	Migraine / Migrain with aura
overweight	Overweight / Obesity
stroke & ischemic attack	Ischemic stroke / Transient cerebral ischemia

[2] https://medschool.vanderbilt.edu/cpm/center-precision-medicine-blog/medi-ensemble-medication-indication-resource

Table S2: Overview about extracted features

id	Type	Description	quantitative feature	Source	no. of features per quarter	total no. of features
1	general	age at index date	x	MarketScan™	1	1
2	general	gender		MarketScan™	1	1
3	general	geographic region		MarketScan™	1	1
4	general	type of insurance and enrollment		MarketScan™	2	2
5	general	coverage of prescriptions yes/no		MarketScan™	1	1
6	general	number of days in hospital	x	MarketScan™	1	9
7	diagnosis	diagnosis		MeSH	1,080	9,720
8	diagnosis	diagnosis group (PheWAS + high level PheWAS)		PheWAS	1,681 + 570	15,129 + 5,130
9	diagnosis	disease pathways		DisGeNET, KEGG, Wiki	487	4,383
10	diagnosis	disease related biological processes		DisGeNET, GO	4,875	43,875
11	diagnosis	biomarkers		TTD	1736	15,624
12	diagnosis	symptoms		human disease-symptoms network	252	2,268
13	drug	prescribed AED total quantity X route of administration X dose	x	MarketScan™	90	715
14	drug	substance class and group		RED BOOK™	179 + 26	1,611 + 234
15	drug	targeted Wiki and KEGG pathways		TTD, DrugBank	671	6,039
16	drug	targeted GO		TTD, DrugBank, GO	2,570	23,130
17	drug	tissue expression of drug targets		Human Protein Atlas	35	315
18	drug	sideeffects X frequency	x	SIDER	2,840	25,560
19	drug	likely drug indication areas (PheWAS + high level PheWAS)		MEDI	877 + 392	7,893 +3,528
					18,368	165,169

Table S3: Number of patients with reported comorbidities 180 days after index date

focused comorbidity	n
Anxiety	2090
Bipolar, Schizophrenia	883
Depression	1799
Diabetes	704
Hyperlipidemia	1503
Hypertension	1177
Migraine	968
Overweight	1338
Stroke, Ischemic Attack	521

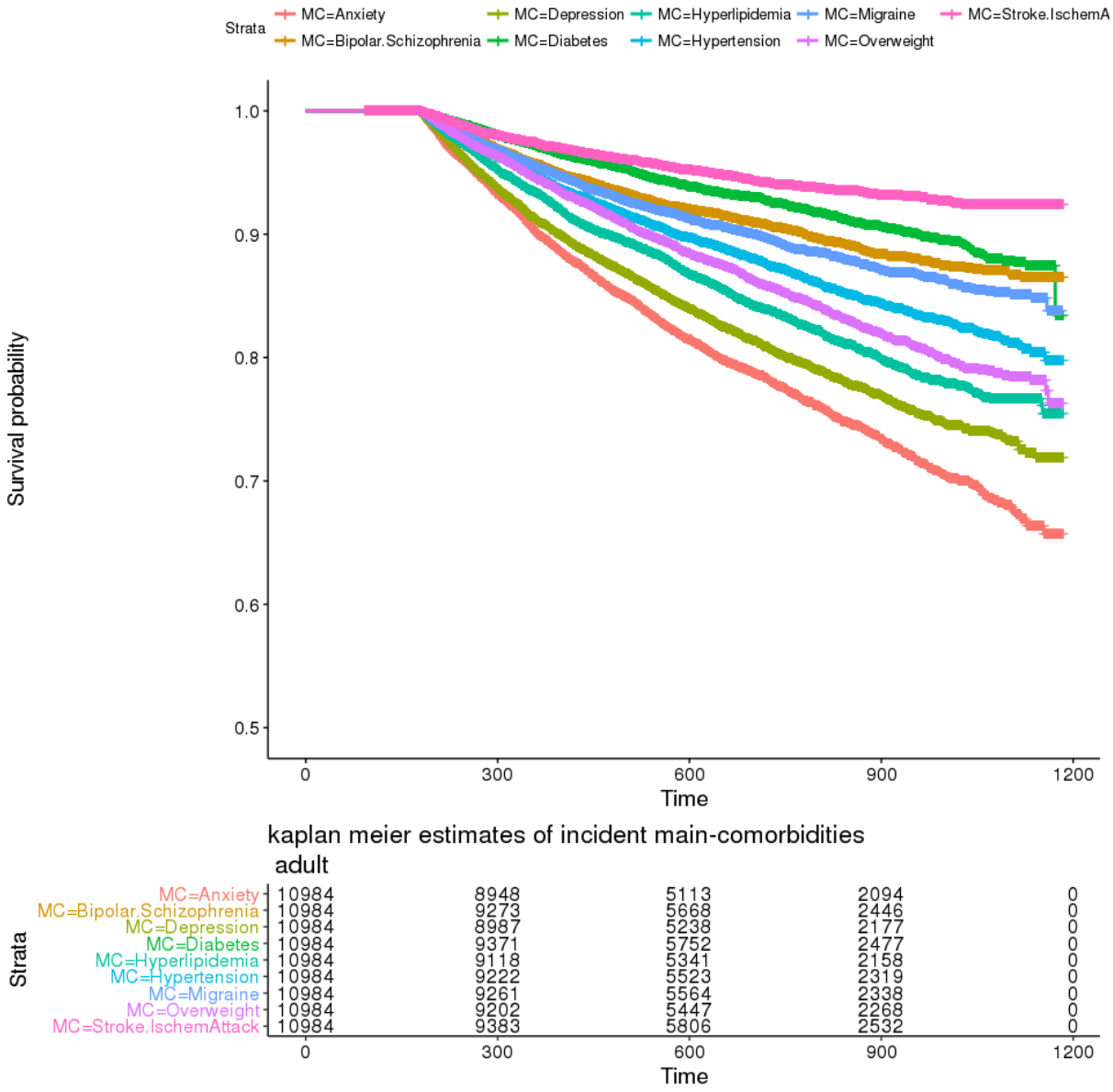


Figure S1: Kaplan Meier curves of focused comorbidities. Event probability as a function of time after index date. First predictions were made 180 days after index date (time 0). The total population consists of incident (at least one incident focused comorbidity after 180) and censored patients (none of the focused comorbidity observed at any time).

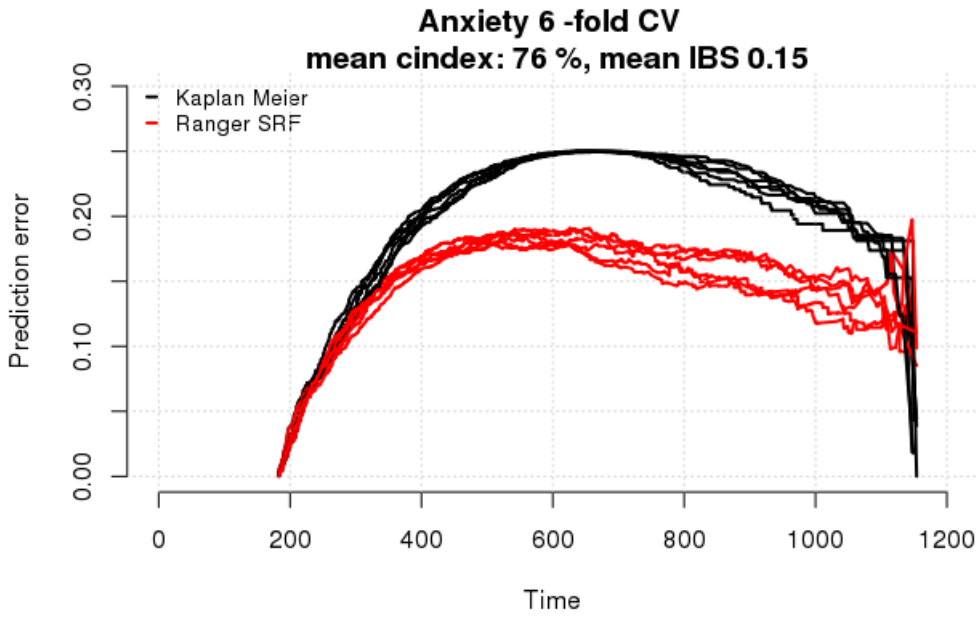


Figure S2: Prediction error (Brier score) as a function of time after index date. First predictions were made 180 days after index date (time 0). Prediction errors based on each individual validation set during the 6-fold cross-validation procedure are shown as separate curves.

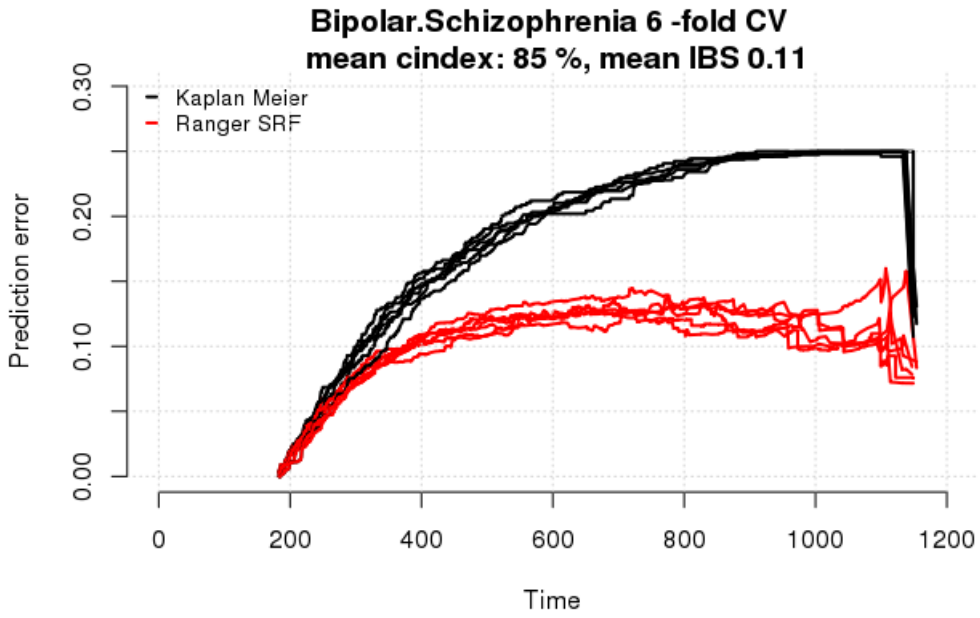


Figure S3: Prediction error (Brier score) as a function of time after index date. First predictions were made 180 days after index date (time 0). Prediction errors based on each individual validation set during the 6-fold cross-validation procedure are shown as separate curves.

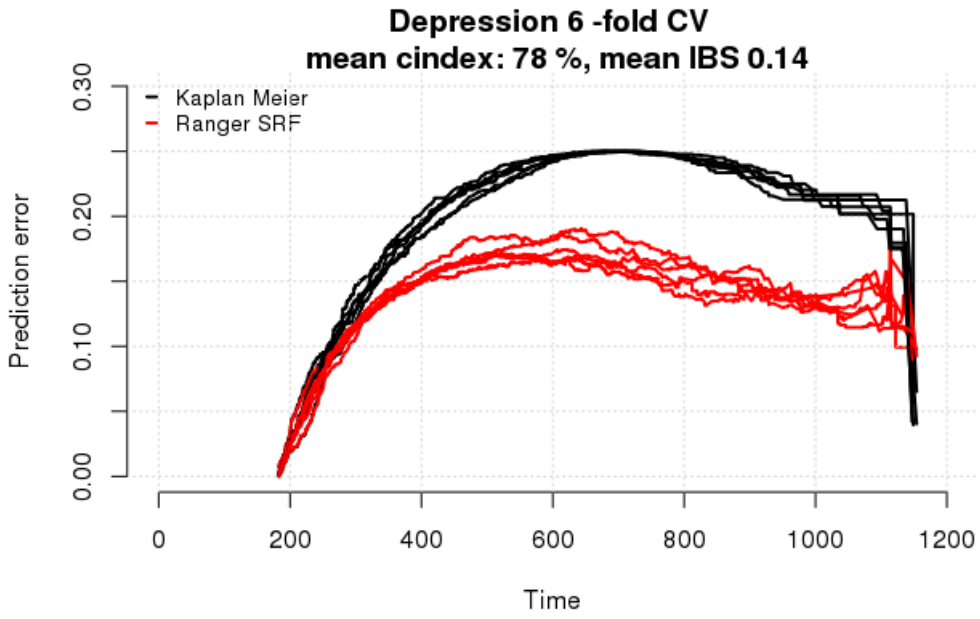


Figure S4: Prediction error (Brier score) as a function of time after index date. First predictions were made 180 days after index date (time 0). Prediction errors based on each individual validation set during the 6-fold cross-validation procedure are shown as separate curves.

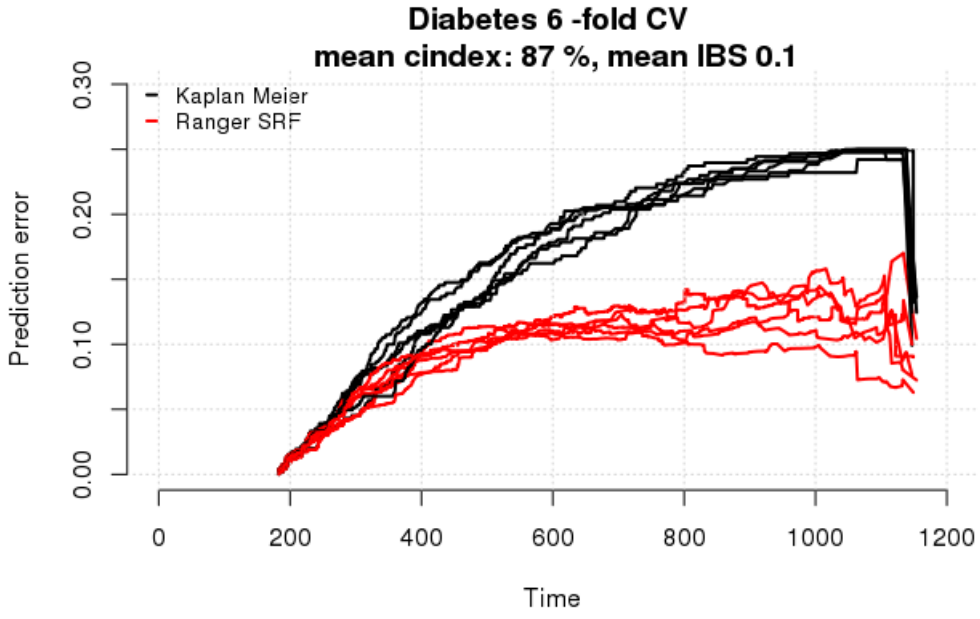


Figure S5: Prediction error (Brier score) as a function of time after index date. First predictions were made 180 days after index date (time 0). Prediction errors based on each individual validation set during the 6-fold cross-validation procedure are shown as separate curves.

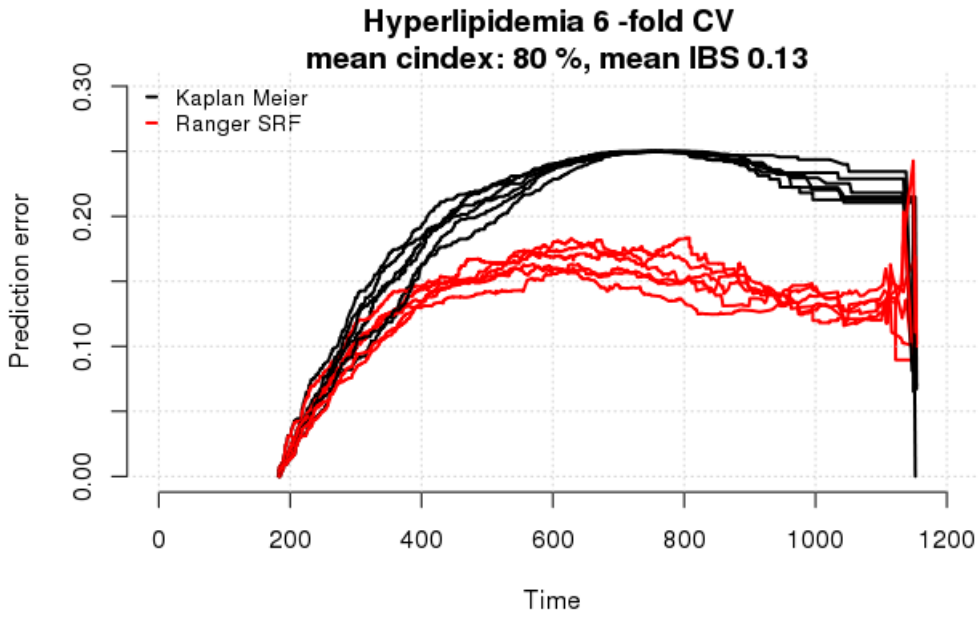


Figure S6: Prediction error (Brier score) as a function of time after index date. First predictions were made 180 days after index date (time 0). Prediction errors based on each individual validation set during the 6-fold cross-validation procedure are shown as separate curves.

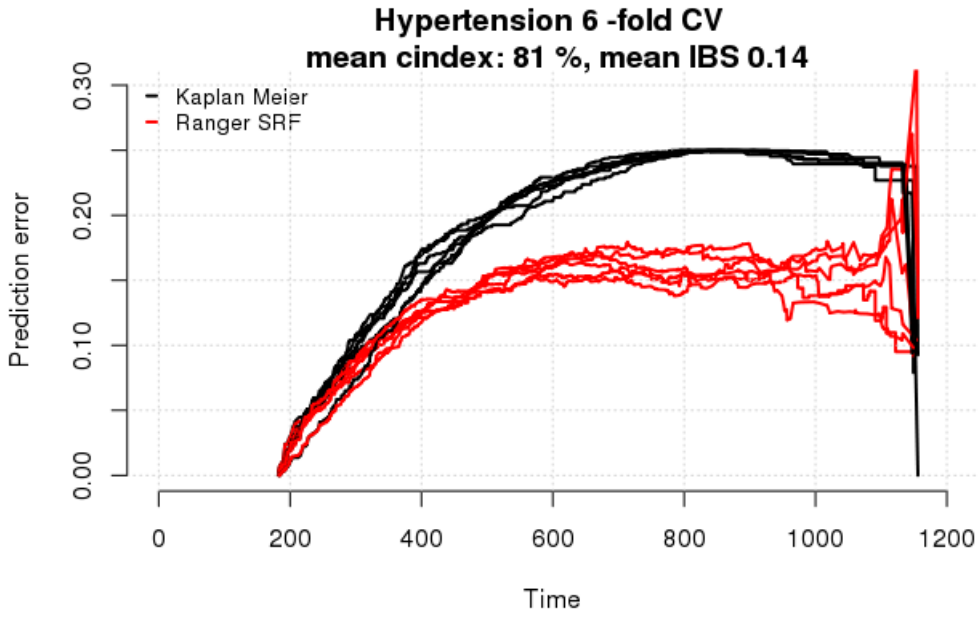


Figure S7: Prediction error (Brier score) as a function of time after index date. First predictions were made 180 days after index date (time 0). Prediction errors based on each individual validation set during the 6-fold cross-validation procedure are shown as separate curves.

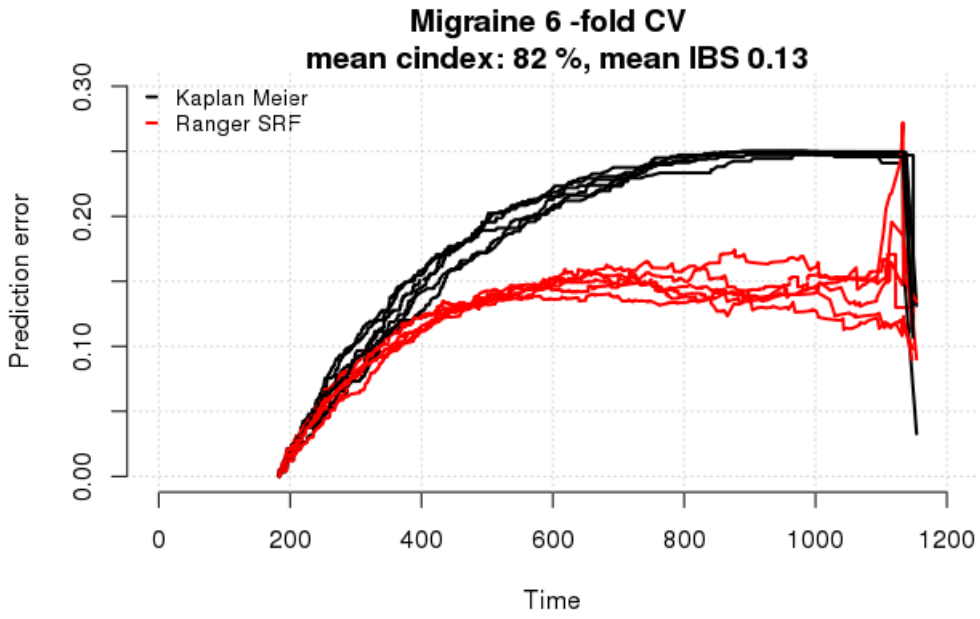


Figure S8: Prediction error (Brier score) as a function of time after index date. First predictions were made 180 days after index date (time 0). Prediction errors based on each individual validation set during the 6-fold cross-validation procedure are shown as separate curves.

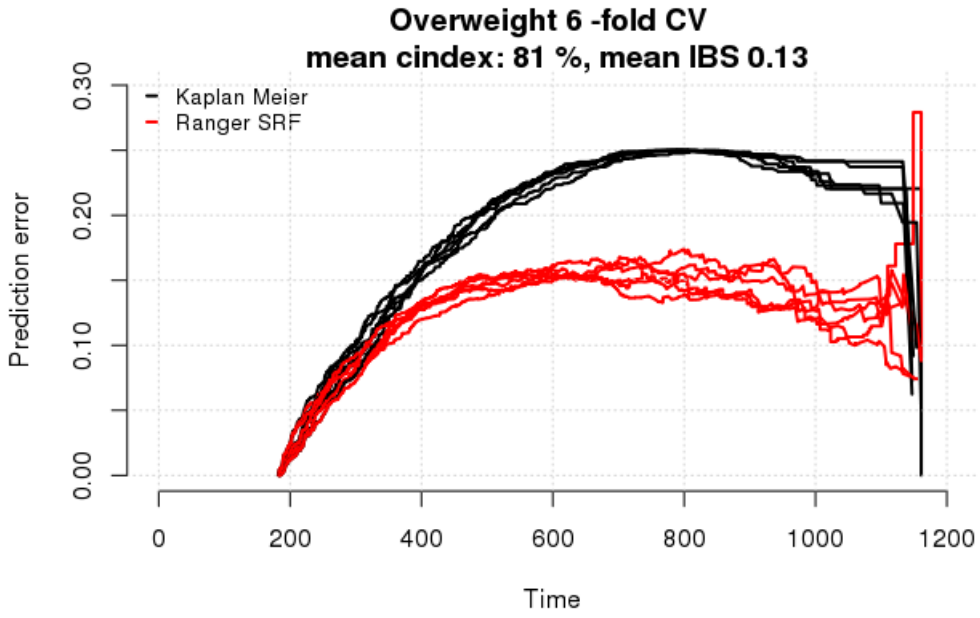


Figure S9: Prediction error (Brier score) as a function of time after index date. First predictions were made 180 days after index date (time 0). Prediction errors based on each individual validation set during the 6-fold cross-validation procedure are shown as separate curves.

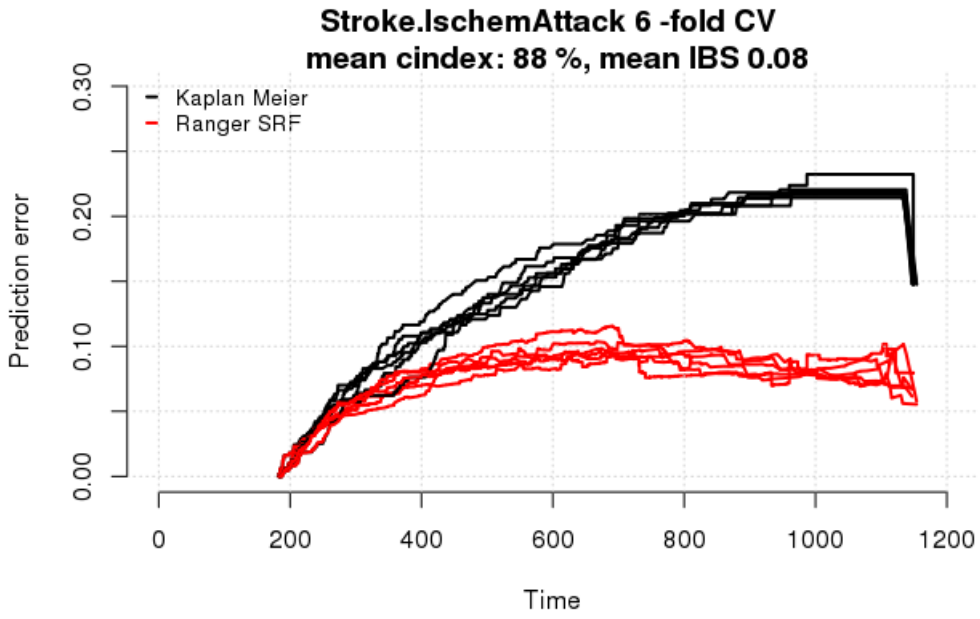


Figure S10: Prediction error (Brier score) as a function of time after index date. First predictions were made 180 days after index date (time 0). Prediction errors based on each individual validation set during the 6-fold cross-validation procedure are shown as separate curves.

Table S4: Anxiety - cumulated importance, stability and overrepresentation of feature domains. 472 features had a positive importance value.

nCV = number of times that a feature of the corresponding domain / sub-domain was selected during 6-fold cross-validation

domain	# features in RSF	cum. Imp	#features in data	pval	fdr	nCV
DISEASE.SYMPTOMS	78	816	2268	0.0E+00	0.0E+00	6
insurance, plantype	1	13	1	0.0E+00	0.0E+00	6
region	1	4	1	0.0E+00	0.0E+00	6
gender	1	4	1	0.0E+00	0.0E+00	5
insurance, healthplan	1	3	1	0.0E+00	0.0E+00	4
COVERAGE_OF_PRESCRIPTIONS	1	2	1	0.0E+00	0.0E+00	3
drug sideeffects	131	1091	25560	2.4E-10	4.7E-09	6
DISEASE.PHEWAS_HL	44	914	5130	3.4E-10	6.5E-09	6
HospDays	2	41	9	2.3E-06	3.5E-05	6
AED	11	108	715	2.6E-06	3.9E-05	6
indication PHEWAS	43	269	7893	9.2E-05	1.1E-03	6
therapeutic class	13	118	1611	5.0E-04	5.3E-03	6
indication PHEWAS HL	19	132	3528	6.2E-03	6.1E-02	6
therapeutic group	2	11	234	3.5E-02	3.2E-01	6
DRUG.TISSUE	1	2	315	2.5E-01	1.0E+00	4
drug pathway	15	84	6039	7.4E-01	1.0E+00	6
DISEASE.PATH	8	119	4383	9.2E-01	1.0E+00	6
DISEASE.PHEWAS	28	373	15129	1.0E+00	1.0E+00	6
DISEASE.BIOMARKER	26	164	15624	1.0E+00	1.0E+00	6
DISEASE.MESHDISEASE	3	84	9720	1.0E+00	1.0E+00	6
DRUG.GO	15	106	23130	1.0E+00	1.0E+00	6
DISEASE.GO	28	391	43875	1.0E+00	1.0E+00	6

Table S5: Bipolar & Schizophrenia - cumulated importance, stability and overrepresentation of feature domains. 459 features had a positive importance value.

nCV = number of times that a feature of the corresponding domain / sub-domain was selected during 6-fold cross-validation

domain	# features in RSF	cum. Imp	#features in data	pval	fdr	nCV
DISEASE.SYMPTOMS	81	629	2268	0.0E+00	0.0E+00	6
AGE	1	18	1	0.0E+00	0.0E+00	4
insurance, plantype	1	10	1	0.0E+00	0.0E+00	6
region	1	4	1	0.0E+00	0.0E+00	6
insurance, healthplan	1	2	1	0.0E+00	0.0E+00	4
COVERAGE_OF_PRESCRIPTIONS	1	2	1	0.0E+00	0.0E+00	5
gender	1	1	1	0.0E+00	0.0E+00	5
drug sideeffects	129	634	25560	9.3E-10	1.7E-08	6
HospDays	2	37	9	2.3E-06	3.5E-05	6
AED	11	33	715	2.6E-06	3.9E-05	6
DISEASE.PHEWAS_HL	35	499	5130	3.9E-06	5.8E-05	6
indication PHEWAS HL	27	76	3528	5.3E-06	7.5E-05	6
therapeutic class	14	63	1611	1.5E-04	1.8E-03	6
indication PHEWAS	39	156	7893	1.2E-03	1.2E-02	6
therapeutic group	2	6	234	3.5E-02	3.2E-01	6
DRUG.TISSUE	2	14	315	7.2E-02	6.4E-01	6
drug pathway	17	82	6039	5.6E-01	1.0E+00	6
DISEASE.PATH	4	50	4383	1.0E+00	1.0E+00	4
DISEASE.BIOMARKER	24	143	15624	1.0E+00	1.0E+00	6
DISEASE.PHEWAS	21	196	15129	1.0E+00	1.0E+00	6
DISEASE.MESHDISEASE	2	73	9720	1.0E+00	1.0E+00	6
DRUG.GO	15	64	23130	1.0E+00	1.0E+00	6
DISEASE.GO	28	323	43875	1.0E+00	1.0E+00	5

Table S6: Depression - cumulated importance, stability and overrepresentation of feature domains. 474 features had a positive importance value.

nCV = number of times that a feature of the corresponding domain / sub-domain was selected during 6-fold cross-validation

domain	# features in RSF	cum. Imp	#features in data	pval	fdr	nCV
DISEASE.SYMPTOMS	84	784	2268	0.0E+00	0.0E+00	6
gender	1	6	1	0.0E+00	0.0E+00	3
insurance, plantype	1	5	1	0.0E+00	0.0E+00	6
insurance, healthplan	1	2	1	0.0E+00	0.0E+00	4
region	1	1	1	0.0E+00	0.0E+00	6
DISEASE.PHEWAS_HL	44	622	5130	3.4E-10	6.5E-09	6
drug sideeffects	123	993	25560	4.2E-08	7.3E-07	6
HospDays	2	36	9	2.3E-06	3.5E-05	6
AED	11	88	715	2.6E-06	3.9E-05	6
therapeutic class	14	159	1611	1.5E-04	1.8E-03	6
indication PHEWAS HL	23	112	3528	2.5E-04	2.9E-03	6
indication PHEWAS	40	214	7893	6.5E-04	6.8E-03	6
therapeutic group	3	14	234	5.8E-03	5.7E-02	6
DRUG.TISSUE	1	12	315	2.5E-01	1.0E+00	4
drug pathway	17	109	6039	5.6E-01	1.0E+00	6
DISEASE.PATH	9	145	4383	8.6E-01	1.0E+00	6
DISEASE.PHEWAS	30	477	15129	9.9E-01	1.0E+00	6
DISEASE.BIOMARKER	24	148	15624	1.0E+00	1.0E+00	6
DISEASE.MESHDISEASE	3	92	9720	1.0E+00	1.0E+00	6
DRUG.GO	15	121	23130	1.0E+00	1.0E+00	6
DISEASE.GO	27	385	43875	1.0E+00	1.0E+00	6

Table S7: Diabetes - cumulated importance, stability and overrepresentation of feature domains. 450 features had a positive importance value.

nCV = number of times that a feature of the corresponding domain / sub-domain was selected during 6-fold cross-validation

domain	# features in RSF	cum. Imp	#features in data	pval	fdr	nCV
DISEASE.SYMPTOMS	77	569	2268	0.0E+00	0.0E+00	6
AGE	1	35	1	0.0E+00	0.0E+00	6
insurance, plantype	1	12	1	0.0E+00	0.0E+00	6
region	1	3	1	0.0E+00	0.0E+00	6
gender	1	2	1	0.0E+00	0.0E+00	5
COVERAGE_OF_PRESCRIPTIONS	1	1	1	0.0E+00	0.0E+00	4
insurance, healthplan	1	0	1	0.0E+00	0.0E+00	4
drug sideeffects	136	706	25560	6.8E-12	1.4E-10	6
indication PHEWAS	53	124	7893	3.1E-08	5.4E-07	6
indication PHEWAS HL	26	104	3528	1.5E-05	2.0E-04	6
AED	10	27	715	1.5E-05	2.0E-04	6
HospDays	1	13	9	3.2E-04	3.5E-03	6
DRUG.TISSUE	3	42	315	1.6E-02	1.5E-01	6
DISEASE.PHEWAS_HL	23	216	5130	2.5E-02	2.4E-01	6
therapeutic class	9	38	1611	2.7E-02	2.5E-01	6
therapeutic group	1	2	234	1.6E-01	1.0E+00	5
drug pathway	21	82	6039	2.2E-01	1.0E+00	6
DISEASE.PATH	12	85	4383	5.7E-01	1.0E+00	5
DISEASE.BIOMARKER	21	144	15624	1.0E+00	1.0E+00	6
DISEASE.PHEWAS	16	184	15129	1.0E+00	1.0E+00	6
DISEASE.GO	22	255	43875	1.0E+00	1.0E+00	6
DRUG.GO	13	77	23130	1.0E+00	1.0E+00	6

Table S8: Hyperlipidemia - cumulated importance, stability and overrepresentation of feature domains. 453 features had a positive importance value.

nCV = number of times that a feature of the corresponding domain / sub-domain was selected during 6-fold cross-validation

domain	# features in RSF	cum. Imp	#features in data	pval	fdr	nCV
DISEASE.SYMPTOMS	81	670	2268	0.0E+00	0.0E+00	6
insurance, plantype	1	6	1	0.0E+00	0.0E+00	6
region	1	5	1	0.0E+00	0.0E+00	6
drug sideeffects	131	964	25560	2.4E-10	4.7E-09	6
DISEASE.PHEWAS_HL	37	601	5130	5.9E-07	1.0E-05	6
HospDays	2	26	9	2.3E-06	3.5E-05	6
indication PHEWAS HL	27	214	3528	5.3E-06	7.5E-05	6
therapeutic class	13	108	1611	5.0E-04	5.3E-03	6
AED	7	53	715	1.7E-03	1.7E-02	6
therapeutic group	3	19	234	5.8E-03	5.7E-02	6
indication PHEWAS	32	121	7893	4.0E-02	3.7E-01	6
DISEASE.PATH	10	184	4383	7.7E-01	1.0E+00	4
drug pathway	13	83	6039	8.8E-01	1.0E+00	6
DISEASE.BIOMARKER	25	145	15624	1.0E+00	1.0E+00	6
DISEASE.PHEWAS	17	259	15129	1.0E+00	1.0E+00	6
DISEASE.MESHDISEASE	2	98	9720	1.0E+00	1.0E+00	6
DRUG.GO	20	138	23130	1.0E+00	1.0E+00	6
DISEASE.GO	31	399	43875	1.0E+00	1.0E+00	6

Table S9: Hypertension - cumulated importance, stability and overrepresentation of feature domains. 457 features had a positive importance value.

nCV = number of times that a feature of the corresponding domain / sub-domain was selected during 6-fold cross-validation

domain	# features in RSF	cum. Imp	#features in data	pval	fdr	nCV
DISEASE.SYMPTOMS	80	720	2268	0.0E+00	0.0E+00	6
AGE	1	46	1	0.0E+00	0.0E+00	6
insurance, plantype	1	10	1	0.0E+00	0.0E+00	6
region	1	4	1	0.0E+00	0.0E+00	6
gender	1	4	1	0.0E+00	0.0E+00	4
COVERAGE_OF_PRESCRIPTIONS	1	3	1	0.0E+00	0.0E+00	5
insurance, healthplan	1	1	1	0.0E+00	0.0E+00	3
drug sideeffects	128	845	25560	1.8E-09	3.3E-08	6
therapeutic group	6	122	234	8.6E-06	1.2E-04	6
indication PHEWAS HL	26	154	3528	1.5E-05	2.0E-04	6
DRUG.TISSUE	6	47	315	5.7E-05	7.3E-04	6
AED	9	52	715	8.0E-05	9.9E-04	6
DISEASE.PHEWAS_HL	31	483	5130	1.2E-04	1.5E-03	6
therapeutic class	14	73	1611	1.5E-04	1.8E-03	6
HospDays	1	21	9	3.2E-04	3.5E-03	6
indication PHEWAS	41	210	7893	3.5E-04	3.8E-03	6
DISEASE.PATH	14	189	4383	3.5E-01	1.0E+00	5
drug pathway	14	67	6039	8.2E-01	1.0E+00	6
DISEASE.BIOMARKER	21	78	15624	1.0E+00	1.0E+00	6
DISEASE.PHEWAS	16	70	15129	1.0E+00	1.0E+00	6
DISEASE.MESHDISEASE	1	2	9720	1.0E+00	1.0E+00	4
DRUG.GO	15	108	23130	1.0E+00	1.0E+00	6
DISEASE.GO	28	364	43875	1.0E+00	1.0E+00	6

Table S10: Migraine - cumulated importance, stability and overrepresentation of feature domains. 454 features had a positive importance value.

nCV = number of times that a feature of the corresponding domain / sub-domain was selected during 6-fold cross-validation

domain	# features in RSF	cum. Imp	#features in data	pval	fdr	nCV
DISEASE.SYMPTOMS	88	682	2268	0.0E+00	0.0E+00	6
gender	1	20	1	0.0E+00	0.0E+00	4
AGE	1	10	1	0.0E+00	0.0E+00	6
region	1	3	1	0.0E+00	0.0E+00	6
COVERAGE_OF_PRESCRIPTIONS	1	2	1	0.0E+00	0.0E+00	3
insurance, healthplan	1	2	1	0.0E+00	0.0E+00	4
insurance, plantype	1	1	1	0.0E+00	0.0E+00	6
drug sideeffects	127	695	25560	3.4E-09	6.2E-08	6
indication PHEWAS HL	26	118	3528	1.5E-05	2.0E-04	6
indication PHEWAS	45	183	7893	2.2E-05	2.8E-04	6
AED	9	40	715	8.0E-05	9.9E-04	6
DISEASE.PHEWAS_HL	30	407	5130	2.7E-04	3.0E-03	6
HospDays	1	8	9	3.2E-04	3.5E-03	6
therapeutic class	11	40	1611	4.2E-03	4.2E-02	6
therapeutic group	2	9	234	3.5E-02	3.2E-01	5
DRUG.TISSUE	2	25	315	7.2E-02	6.4E-01	6
drug pathway	18	89	6039	4.6E-01	1.0E+00	6
DISEASE.PATH	5	33	4383	9.9E-01	1.0E+00	5
DISEASE.BIOMARKER	25	202	15624	1.0E+00	1.0E+00	6
DISEASE.PHEWAS	16	128	15129	1.0E+00	1.0E+00	6
DISEASE.MESHDISEASE	1	51	9720	1.0E+00	1.0E+00	6
DRUG.GO	16	92	23130	1.0E+00	1.0E+00	6
DISEASE.GO	26	151	43875	1.0E+00	1.0E+00	6

Table S11: Overweight - cumulated importance, stability and overrepresentation of feature domains. 468 features had a positive importance value.

nCV = number of times that a feature of the corresponding domain / sub-domain was selected during 6-fold cross-validation

domain	# features in RSF	cum. Imp	#features in data	pval	fdr	nCV
DISEASE.SYMPTOMS	68	508	2268	0.0E+00	0.0E+00	6
COVERAGE_OF_PRESCRIPTIONS	1	10	1	0.0E+00	0.0E+00	3
insurance, plantype	1	5	1	0.0E+00	0.0E+00	6
region	1	4	1	0.0E+00	0.0E+00	6
gender	1	4	1	0.0E+00	0.0E+00	3
insurance, healthplan	1	0	1	0.0E+00	0.0E+00	3
drug sideeffects	133	868	25560	5.9E-11	1.2E-09	6
indication PHEWAS	49	235	7893	9.8E-07	1.6E-05	6
HospDays	2	37	9	2.3E-06	3.5E-05	6
indication PHEWAS HL	27	170	3528	5.3E-06	7.5E-05	6
DISEASE.PHEWAS_HL	34	489	5130	9.7E-06	1.3E-04	6
AED	9	68	715	8.0E-05	9.9E-04	6
therapeutic group	4	16	234	7.9E-04	8.2E-03	6
therapeutic class	12	92	1611	1.5E-03	1.5E-02	6
DRUG.TISSUE	2	28	315	7.2E-02	6.4E-01	5
drug pathway	19	106	6039	3.7E-01	1.0E+00	6
DISEASE.PATH	9	77	4383	8.6E-01	1.0E+00	4
DISEASE.PHEWAS	22	479	15129	1.0E+00	1.0E+00	6
DISEASE.BIOMARKER	23	128	15624	1.0E+00	1.0E+00	6
DISEASE.MESHDISEASE	2	53	9720	1.0E+00	1.0E+00	6
DRUG.GO	18	146	23130	1.0E+00	1.0E+00	6
DISEASE.GO	30	411	43875	1.0E+00	1.0E+00	6

Table S12: Stroke & Ischemic Attack - cumulated importance, stability and overrepresentation of feature domains. 425 features had a positive importance value.

nCV = number of times that a feature of the corresponding domain / sub-domain was selected during 6-fold cross-validation

domain	# features in RSF	cum. Imp	#features in data	pval	fdr	nCV
DISEASE.SYMPTOMS	74	481	2268	0.0E+00	0.0E+00	6
region	1	5	1	0.0E+00	0.0E+00	6
insurance, plantype	1	4	1	0.0E+00	0.0E+00	6
COVERAGE_OF_PRESCRIPTIONS	1	3	1	0.0E+00	0.0E+00	3
insurance, healthplan	1	2	1	0.0E+00	0.0E+00	4
gender	1	2	1	0.0E+00	0.0E+00	3
drug sideeffects	127	500	25560	3.4E-09	6.2E-08	6
indication PHEWAS HL	28	88	3528	1.8E-06	3.1E-05	6
AED	10	22	715	1.5E-05	2.0E-04	6
therapeutic group	5	50	234	8.9E-05	1.1E-03	6
HospDays	1	13	9	3.2E-04	3.5E-03	6
indication PHEWAS	40	107	7893	6.5E-04	6.8E-03	6
therapeutic class	8	31	1611	5.9E-02	5.4E-01	6
DISEASE.PHEWAS_HL	19	161	5130	1.5E-01	1.0E+00	6
drug pathway	22	64	6039	1.6E-01	1.0E+00	6
DRUG.TISSUE	1	16	315	2.5E-01	1.0E+00	4
DISEASE.PATH	6	60	4383	9.8E-01	1.0E+00	6
DISEASE.BIOMARKER	21	180	15624	1.0E+00	1.0E+00	6
DISEASE.PHEWAS	15	40	15129	1.0E+00	1.0E+00	6
DISEASE.MESHDISEASE	2	68	9720	1.0E+00	1.0E+00	6
DRUG.GO	16	86	23130	1.0E+00	1.0E+00	6
DISEASE.GO	25	304	43875	1.0E+00	1.0E+00	6

Table S13: Classification of features as derived from biomedical knowledge or not.

domain	derived?
insurance, plantype	no
region	no
gender	no
insurance, healthplan	no
COVERAGE_OF_PRESCRIPTIONS	no
DISEASE.PHEWAS_HL	no
HospDays	no
AED	no
therapeutic class	no
therapeutic group	no
AGE	no
DISEASE.PHEWAS	no
drug sideeffects	yes
indication PHEWAS	yes
indication PHEWAS HL	yes
DRUG.TISSUE	yes
drug pathway	yes
DISEASE.PATH	yes
DISEASE.BIOMARKER	yes
DISEASE.MESHDISEASE	yes
DRUG.GO	yes
DISEASE.GO	yes
DISEASE.SYMPTOMS	yes

Table S14: Cumulative importance of derived and original features in different comorbidity models.

	cumulative importance distribution		
focused comorbidity	**derived features**	**non-derived features**	**all**
Anxiety	67%	33%	100%
Bipolar	72%	28%	100%
Depression	69%	31%	100%
Diabetes	80%	20%	100%
Hyperlipidemia	74%	26%	100%
Hypertension	76%	24%	100%
Migraine	78%	22%	100%
Overweight	69%	31%	100%
Stroke	85%	15%	100%
all	**73%**	**27%**	**100%**

Table S15: Cumulative importance of feature domains, averaged over comorbidities. Feature domains are classified as derived from biomedical knowledge or not.

	cumulative importance distribution averaged by focused-comorbidity		
domain	**yes**	**no**	**all**
drug sideeffects	23%		23%
DISEASE.SYMPTOMS	18%		18%
DISEASE.PHEWAS_HL		14%	14%
DISEASE.GO	9%		9%
DISEASE.PHEWAS		7%	7%
indication PHEWAS	5%		5%
DISEASE.BIOMARKER	4%		4%
indication PHEWAS HL	4%		4%
DISEASE.PATH	3%		3%
DRUG.GO	3%		3%
drug pathway	2%		2%
therapeutic class		2%	2%
DISEASE.MESHDISEASE	2%		2%
AED		2%	2%
therapeutic group		1%	1%
AGE		1%	1%
HospDays		1%	1%
DRUG.TISSUE	1%		1%
insurance, plantype		0%	0%
gender		0%	0%
region		0%	0%
COVERAGE_OF_PRESCRIPTIONS		0%	0%
insurance, healthplan		0%	0%
all	**73%**	**27%**	**100%**

Table S16: Anxiety - cumulative importance, stability and overrepresentation of selected drug related feature sub-domains

nCV = number of times that a feature of the corresponding domain / sub-domain was selected during 6-fold cross-validation

domain	subdomain	# features in RSF	cum. Imp	#features in data	pval	fdr	nCV
therapeutic class	ANXIOLYTIC.SEDATIVE.HYPNOT.NEC	3	52	9	3.1E-08	3.5E-06	6
AED	Phenytoin.Sodium..Extended	2	9	36	4.0E-04	5.8E-03	6
therapeutic class	ANTICONVULSANT..BENZODIAZEPINE	1	11	9	5.6E-04	5.9E-03	5
AED	Gabapentin	2	41	53	1.3E-03	1.3E-02	6
AED	Clonazepam	1	24	45	1.4E-02	1.1E-01	6
AED	Carbamazepine	1	1	45	1.4E-02	1.1E-01	5
AED	Topiramate	1	20	46	1.5E-02	1.1E-01	6
AED	Lamotrigine	1	0	85	4.5E-02	3.4E-01	6

Table S17: Bipolar disorder & schizophrenia - cumulative importance, stability and overrepresentation of selected drug related feature sub-domains

nCV = number of times that a feature of the corresponding domain / sub-domain was selected during 6-fold cross-validation

domain	subdomain	# features in RSF	cum. Imp	#features in data	pval	fdr	nCV
indication	IMMUNE DISORDERS	3	18	27	3.0E-06	1.6E-04	6
drug pathway	SEROTONERGIC.SYNAPSE	2	16	9	4.1E-06	1.8E-04	6
drug pathway	RETROGRADE.ENDOCANNABINOID.SIGNALING	2	6	9	4.1E-06	1.8E-04	6
AED	Phenytoin.Sodium..Extended	3	4	36	9.8E-06	3.7E-04	6
drug pathway	GAP.JUNCTION	1	20	9	4.8E-04	5.8E-03	5
drug pathway	FOLATE.METABOLISM	1	5	9	4.8E-04	5.8E-03	3
drug pathway	NICOTINE.METABOLISM	1	5	9	4.8E-04	5.8E-03	6
drug pathway	STEROID.BIOSYNTHESIS	1	4	9	4.8E-04	5.8E-03	6
drug pathway	PPAR.SIGNALING.PATHWAY	1	4	9	4.8E-04	5.8E-03	6
drug pathway	EFFECTS.OF.NITRIC.OXIDE	1	2	9	4.8E-04	5.8E-03	6
indication	TICS AND STUTTERING	1	17	18	2.0E-03	1.9E-02	2
indication	SYSTEMIC LUPUS ERYTHEMATOSUS	1	6	18	2.0E-03	1.9E-02	4
AED	Lamotrigine	2	5	85	3.9E-03	3.6E-02	6
AED	Divalproex.Sodium	1	3	27	4.5E-03	3.9E-02	4
AED	Gabapentin	1	12	53	1.7E-02	1.2E-01	6

Table S18: Depression - cumulative importance, stability and overrepresentation of selected drug related feature sub-domains

nCV = number of times that a feature of the corresponding domain / sub-domain was selected during 6-fold cross-validation

domain	subdomain	# features in RSF	cum. Imp	#features in data	pval	fdr	nCV
therapeutic class	ANTIINFLAM.S.MM.AGNTS.COMB.NEC	4	7	9	5.9E-11	1.7E-08	6
therapeutic class	ANXIOLYTIC.SEDATIVE.HYPNOT.NEC	2	22	9	3.3E-06	1.6E-04	6
therapeutic class	THY.ANTITHY..THYROID.HORMONES	2	13	9	3.3E-06	1.6E-04	6
DRUG.GO	BP.G.PROTEIN.COUPLED.RECEPTOR.SIGNALING.PATHWAY..COUPLED.TO.CYCLIC.NUCLEOTIDE.SECOND.MESSENGER	1	28	9	4.2E-04	5.8E-03	1
DRUG.GO	BP.CELLULAR.RESPONSE.TO.GROWTH.FACTOR.STIMULUS	1	23	9	4.2E-04	5.8E-03	4
DRUG.GO	BP.G.PROTEIN.COUPLED.RECEPTOR.SIGNALING.PATHWAY	1	6	9	4.2E-04	5.8E-03	5
DRUG.GO	BP.REGULATION.OF.ION.TRANSMEMBRANE.TRANSPORT	1	5	9	4.2E-04	5.8E-03	1
therapeutic class	ANTIBIOT..ERYTHROMYCN.MACROLID	1	2	9	4.2E-04	5.8E-03	5
DRUG.GO	BP.STEROID.BIOSYNTHETIC.PROCESS	1	5	9	4.2E-04	5.8E-03	6
therapeutic class	ASH..BENZODIAZEPINES	1	32	9	4.2E-04	5.8E-03	4
therapeutic class	ANTICONVULSANT..BENZODIAZEPINE	1	19	9	4.2E-04	5.8E-03	4
indication	ACUTE REACTION TO STRESS	1	10	18	1.7E-03	1.7E-02	4

Table S19: Diabetes - cumulative importance, stability and overrepresentation of selected drug related feature sub-domains

nCV = number of times that a feature of the corresponding domain / sub-domain was selected during 6-fold cross-validation

domain	subdomain	# features in RSF	cum. Imp	#features in data	pval	fdr	nCV
drug pathway	RETROGRADE.ENDOCANNABINOID.SIGNALING	3	7	9	2.7E-08	3.2E-06	6
indication	IMMUNE DISORDERS	4	19	27	6.2E-08	6.1E-06	6
AED	Phenytoin.Sodium..Extended	3	4	36	1.2E-05	4.3E-04	6
drug pathway	ADRENERGIC.SIGNALING.IN.CARDIOMYOCYTES	1	8	9	5.2E-04	5.8E-03	4
drug pathway	SEROTONERGIC.SYNAPSE	1	6	9	5.2E-04	5.8E-03	6
drug pathway	CGMP.PKG.SIGNALING.PATHWAY	1	6	9	5.2E-04	5.8E-03	5
drug pathway	PPAR.SIGNALING.PATHWAY	1	6	18	2.2E-03	2.0E-02	6
drug pathway	STEROID.BIOSYNTHESIS	1	6	18	2.2E-03	2.0E-02	6
AED	Lamotrigine	2	2	85	4.4E-03	3.9E-02	6
AED	Divalproex.Sodium	1	5	27	4.9E-03	4.2E-02	6
indication	OTHER FORMS OF CHRONIC HEART DISEASE	1	24	36	8.5E-03	6.8E-02	2
AED	Gabapentin	1	15	53	1.8E-02	1.3E-01	5

Table S20: Hyperlipidemia - cumulative importance, stability and overrepresentation of selected drug related feature sub-domains

nCV = number of times that a feature of the corresponding domain / sub-domain was selected during 6-fold cross-validation

domain	subdomain	# features in RSF	cum. Imp	#features in data	pval	fdr	nCV
DRUG.GO	BP.REGULATION.OF.ION.TRANSMEMBRANE.TRANSPORT	2	12	9	3.2E-06	1.6E-04	4
DRUG.GO	BP.G.PROTEIN.COUPLED.RECEPTOR.SIGNALING.PATHWAY..COUPLED.TO.CYCLIC.NUCLEOTIDE.SECOND.MESSENGER	1	8	9	4.1E-04	5.8E-03	2
DRUG.GO	BP.G.PROTEIN.COUPLED.RECEPTOR.SIGNALING.PATHWAY	1	7	9	4.1E-04	5.8E-03	6
indication	DIABETES MELLITUS	2	82	36	2.6E-04	5.8E-03	6
indication	ISCHEMIC STROKE	1	16	18	1.7E-03	1.7E-02	6
AED	Divalproex.Sodium	1	3	27	3.8E-03	3.5E-02	1
AED	Phenytoin.Sodium..Extended	1	7	36	6.7E-03	5.5E-02	6
AED	Carbamazepine	1	0	45	1.0E-02	8.0E-02	6
AED	Gabapentin	1	21	53	1.4E-02	1.1E-01	6

Table S21: Hypertension - cumulative importance, stability and overrepresentation of selected drug related feature sub-domains

nCV = number of times that a feature of the corresponding domain / sub-domain was selected during 6-fold cross-validation

domain	subdomain	# features in RSF	cum. Imp	#features in data	pval	fdr	nCV
therapeutic group	CARDIOVASCULAR.AGENTS	3	120	9	2.5E-08	3.2E-06	#NV
AED	Phenytoin.Sodium..Extended	3	14	36	1.1E-05	4.1E-04	6
indication	IMMUNE DISORDERS	2	16	18	4.2E-05	1.3E-03	5
DRUG.GO	BP.G.PROTEIN.COUPLED.RECEPTOR.SIGNALING.PATHWAY..COUPLED.TO.CYCLIC.NUCLEOTIDE.SECOND.MESSENGER	1	17	9	5.0E-04	5.8E-03	6
DRUG.GO	BP.G.PROTEIN.COUPLED.RECEPTOR.SIGNALING.PATHWAY	1	11	9	5.0E-04	5.8E-03	6
DRUG.GO	BP.CORTISOL.METABOLIC.PROCESS	1	6	9	5.0E-04	5.8E-03	6
indication	MYOCARDIAL INFARCTION	1	29	9	5.0E-04	5.8E-03	5
indication	ISCHEMIC STROKE	1	12	9	5.0E-04	5.8E-03	4
AED	Divalproex.Sodium	1	3	27	4.7E-03	4.1E-02	4
AED	Clonazepam	1	18	45	1.3E-02	9.8E-02	6
AED	Gabapentin	1	12	53	1.7E-02	1.3E-01	6
AED	Lamotrigine	1	1	85	4.1E-02	3.1E-01	6

Table S22: Migraine - cumulative importance, stability and overrepresentation of selected drug related feature sub-domains

nCV = number of times that a feature of the corresponding domain / sub-domain was selected during 6-fold cross-validation

domain	subdomain	# features in RSF	cum. Imp	#features in data	pval	fdr	nCV
drug sideeffects	NERVOUS.SYSTEM.DISORDER	5	22	18	2.5E-11	9.9E-09	6
drug pathway	SEROTONERGIC.SYNAPSE	3	27	9	1.6E-08	2.4E-06	6
AED	Phenytoin.Sodium..Extended	4	14	36	1.4E-07	1.3E-05	6
drug pathway	RETROGRADE.ENDOCANNABINOID.SIGNALING	2	7	9	3.1E-06	1.6E-04	6
drug pathway	SALIVARY.SECRETION	1	6	9	4.0E-04	5.8E-03	2
drug pathway	STEROID.BIOSYNTHESIS	1	6	9	4.0E-04	5.8E-03	6
drug pathway	CGMP.PKG.SIGNALING.PATHWAY	1	6	9	4.0E-04	5.8E-03	3
DRUG.GO	BP.G.PROTEIN.COUPLED.RECEPTOR.SIGNALING.PATHWAY..COUPLED.TO.CYCLIC.NUCLEOTIDE.SECOND.MESSENGER	1	9	9	4.0E-04	5.8E-03	6
DRUG.GO	BP.CELLULAR.RESPONSE.TO.GROWTH.FACTOR.STIMULUS	1	6	9	4.0E-04	5.8E-03	6
DRUG.GO	BP.CELLULAR.RESPONSE.TO.HYPOXIA	1	5	9	4.0E-04	5.8E-03	4
DRUG.GO	BP.AXONOGENESIS	1	4	9	4.0E-04	5.8E-03	3
AED	Gabapentin	1	16	53	1.4E-02	1.1E-01	5
AED	Lamotrigine	1	3	85	3.3E-02	2.5E-01	6

Table S23: Overweight - cumulative importance, stability and overrepresentation of selected drug related feature sub-domains

nCV = number of times that a feature of the corresponding domain / sub-domain was selected during 6-fold cross-validation

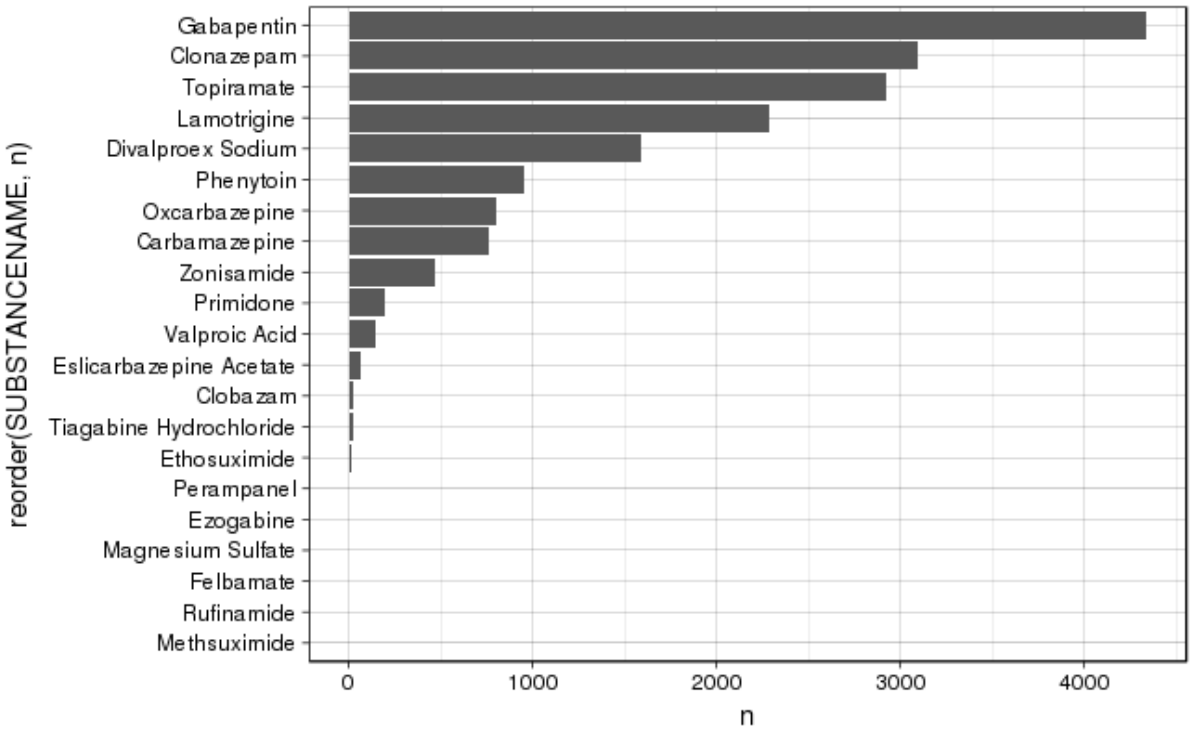
domain	subdomain	# features in RSF	cum. Imp	#features in data	pval	fdr	nCV
drug pathway	STEROID.BIOSYNTHESIS	2	13	9	3.1E-06	1.6E-04	6
drug pathway	EFFECTS.OF.NITRIC.OXIDE	2	6	9	3.1E-06	1.6E-04	6
drug pathway	FLUOROPYRIMIDINE.ACTIVITY	2	4	9	3.1E-06	1.6E-04	6
AED	Phenytoin.Sodium..Extended	2	7	36	2.5E-04	5.8E-03	6
drug pathway	PPAR.SIGNALING.PATHWAY	1	6	9	4.0E-04	5.8E-03	6
drug pathway	NICOTINE.METABOLISM	1	3	9	4.0E-04	5.8E-03	5
DRUG.GO	BP.ADENYLATE.CYCLASE.INHIBITING.G.PROTEIN.COUPLED.RECEPTOR.SIGNALING.PATHWAY	1	20	9	4.0E-04	5.8E-03	4
DRUG.GO	BP.G.PROTEIN.COUPLED.RECEPTOR.SIGNALING.PATHWAY	1	17	9	4.0E-04	5.8E-03	6
DRUG.GO	BP.RESPONSE.TO.GLUCOCORTICOID	1	14	9	4.0E-04	5.8E-03	2
DRUG.GO	BP.CELLULAR.KETONE.METABOLIC.PROCESS	1	7	9	4.0E-04	5.8E-03	3
AED	Gabapentin	2	41	53	7.8E-04	8.1E-03	6
indication	OTHER FORMS OF CHRONIC HEART DISEASE	1	33	36	6.6E-03	5.5E-02	6
AED	Clonazepam	1	11	45	1.0E-02	7.9E-02	3
AED	Lamotrigine	1	4	85	3.3E-02	2.5E-01	6

Table S24: Stroke & Ischemic Attack - cumulative importance, stability and overrepresentation of selected drug related feature sub-domains

nCV = number of times that a feature of the corresponding domain / sub-domain was selected during 6-fold cross-validation

domain	subdomain	# features in RSF	cum. Imp	#features in data	pval	fdr	nCV
drug pathway	RETROGRADE.ENDOCANNABINOID.SIGNALING	3	10	9	1.8E-08	2.4E-06	6
AED	Phenytoin.Sodium..Extended	2	0	36	2.8E-04	5.8E-03	6
drug pathway	PPAR.SIGNALING.PATHWAY	1	5	9	4.3E-04	5.8E-03	6
drug pathway	NICOTINE.METABOLISM	1	4	9	4.3E-04	5.8E-03	6
drug pathway	CALCIUM.REGULATION.IN.THE.CARDIAC.CELL	1	4	9	4.3E-04	5.8E-03	2
drug pathway	STEROID.BIOSYNTHESIS	1	3	9	4.3E-04	5.8E-03	5
drug pathway	CGMP.PKG.SIGNALING.PATHWAY	1	3	9	4.3E-04	5.8E-03	5
DRUG.GO	BP.CORTISOL.METABOLIC.PROCESS	1	16	9	4.3E-04	5.8E-03	4
DRUG.GO	BP.CHEMICAL.SYNAPTIC.TRANSMISSION	1	8	9	4.3E-04	5.8E-03	3
DRUG.GO	BP.G.PROTEIN.COUPLED.RECEPTOR.SIGNALING.PATHWAY	1	7	9	4.3E-04	5.8E-03	6
DRUG.GO	BP.CELLULAR.RESPONSE.TO.HYPOXIA	1	7	9	4.3E-04	5.8E-03	6
DRUG.GO	BP.RESPONSE.TO.GLUCOCORTICOID	1	3	9	4.3E-04	5.8E-03	4
DRUG.GO	BP.BLOOD.COAGULATION	1	3	9	4.3E-04	5.8E-03	6
AED	Divalproex.Sodium	1	1	27	4.0E-03	3.6E-02	6
AED	Carbamazepine	1	0	45	1.1E-02	8.4E-02	6
AED	Gabapentin	1	8	53	1.5E-02	1.1E-01	5

Figure S11: Most frequently prescribed AEDs[3]



[3] UCB compounds (listing after Gabapentin) not shown due to constraints by the company