Title: Analytics to predict future International Classification of Diseases (ICD) codes based on Bayesian Probabilities and Network analyses.

Summary: Predictive analytics can be used to anticipate future medical expenses, resource management, and assessing patient needs. To address this issue, novel software and a database was developed that maps ICD10 (International Statistical Classification of Diseases) networks of patients using Bayesian probabilities. The database was based on multiple visits of 91000+ patients and enables the prediction of disease progression/risk.

Background on how to calculate Bayesian Probabilities using an example dataset. Each row is a patient health care medical record. Column 1 is Patient ID, and columns 2 to 4 are ICD codes of the patient.

A00100008	ICD10_STROKE	**ICD10_F11**	ICD10_B96
A00100012	ICD10_STROKE	ICD10_B96	**ICD10_F11**
A00100043	ICD10_I61	ICD10_E78	ICD10_F17
A00100045	ICD10_D64	ICD10_E86	ICD10_N39
A00100052	ICD10_A04	ICD10_B97	ICD10_D63
A00100054	ICD10_B35	ICD10_D70	ICD10_E03
A00100059	ICD10_N13	ICD10_N28	ICD10_F03
A00100060	ICD10_D64	ICD10_E66	ICD10_F15
A00100069	ICD10_H01	ICD10_H35	ICD10_H43
A00100073	ICD10_E11	**ICD10_F11**	ICD10_F32

Let us determine the Bayesian Probability of a patient with ICD10_F11 (Opioid related disorders) having a stroke.

The Bayesian Probability formula is:
$P(A|B)=P(B|A)*P(A)/P(B)$

B = ICD10_F11
A = STROKE

P(B)=number of occurrences in data set=3/10
P(A)=number of occurrences in data set =2/10
P(B|A)=number of times B occurs in A=**2/2**
P(A|B)=(**2/2**)*(**2/10**)/(3/10)
P(A|B)=(**1**)*(.2)/(.3)
P(A|B)=0.66667 or 67%

The dataset suggests a patient with ICD10_F11 has a 67% of having a stroke.

Establishing a Bayesian Probability Dataset for all ICDs.
The probabilities of the 3105 unique ICD codes, based on all possible priors in the dataset of 91,222 patient records, were determined using the C++ program 'make_bay_ICD_dataset.cpp'. The input files for this program were 'nr_report.txt' and 'codes.txt'. The nr_report.txt contains a list of all the non-redundant ICD codes for each patient in the raw data file and the codes.txt contains the 3105 unique ICD codes. The output file, bay_prob_ICD.txt, consists of 656,047 rows and four columns: (i) the Bayesian probability, (ii) the number of records used in the prediction, (iii) the target ICD, and (iv) the prior ICD (i.e., the given). Of note, the program took 3 days to determine all 656,047 probabilities.

The new dataset (bay_prob_ICD.txt) was used to (i) predict disease progression/risk in individual patients and populations of patients with similar ICDs (disease states), and (ii) generate network diagrams using the program: 'predict_icds2.cpp'. The C++ program yields two files: 'edges.csv' and 'nodes.csv'. These files served as inputs to the R program called 'R_plot_networks.txt', which was used to generate network plots. Below are three examples and corresponding network diagrams.

Target	Prob	Prior
ICD10_N17	0.48	ICD10_E11
ICD10_N17	0.57	ICD10_E87
ICD10_N17	0.53	ICD10_I10
ICD10_N17	0.44	ICD10_N18
ICD10_N17	0.50	ICD10_Z79
ICD10_N18	0.53	ICD10_E11
ICD10_N18	0.40	ICD10_I10
ICD10_N18	0.42	ICD10_I12
ICD10_N19	0.46	ICD10_E11
ICD10_N19	0.46	ICD10_I10
ICD10_N19	0.40	ICD10_N18
ICD10_N19	0.44	ICD10_Z79