Published online: 2019-08-16

Appendix: Content Summaries of Selected Best Papers for the IMIA Yearbook 2019 Section "Clinical Information Systems"

Eichstaedt JC, Smith RJ, Merchant RM, Ungar LH, Crutchley P, Preoţiuc-Pietro D Asch DA, Schwartz HA

Facebook language predicts depression in medical records

Proc Natl Acad Sci U S A 2018;115(44):11203-8

The paper deals with depression, one of the most prevalent mental illnesses. The authors state that each year around 7%-26% of the US population experience a depression or depression related symptoms. The number of patients that receive minimally adequate treatment is according to their investigation at most 49%. Therefore, the authors conclude that such high rates of underdiagnosis and undertreatment suggest that existing procedures for screening and identifying depressed patients are inadequate. The authors introduce a novel approach by using Facebook language data from a sample of consenting patients to detect/predict depression from that data. The study involved the analysis of the Facebook posts of 114 patients that prior had a diagnosis of depression. For each of the patients with such a diagnose, five random patients without a diagnose of depression in the same period of time were added to the analysis to simulate the prevalence of the disease. The authors built a prediction model using the textual content of the Facebook posts, post length, frequency of posting, temporal posting patterns, and demographics. A 10-fold cross-validation was applied to avoid overfitting. To yield interpretable and fine-grained language variables, 200 topics were extracted using latent dirichlet allocation. The results suggest that the closer in time the Facebook data are to the documentation of depression in the EMR, the better their predictive power. Within 6 months preceding the documentation of depression an accuracy of 0.72(AUC) was achieved. The authors conclude that Facebook language-based prediction models perform similarly to screening surveys in identifying patients with depression when using diagnostic codes in the EMR to identify diagnoses of depression.

Parr SK, Shotwell MS, Jeffery AD, Lasko TA, Matheny ME

Automated mapping of laboratory tests to LOINC codes using noisy labels in a national electronic health record system database

J Am Med Inform Assoc 2018;25(10):1292-300

Aggregated data from multiple data-sources can be a valuable source for different fields of research and other domains such as public health or the creation of clinical evidence. A very common problem related to the use of data from different sources is their different coding and the use of non-standardized terminology. An important aspect in the assessment of patient outcomes are laboratory findings. In this context Logical Observation Identifiers Names and Codes (LOINC) can be seen as an important standard. However, mapping laboratory findings to LOINC codes manually can be time consuming. As electronic health records (EHR) are a rich source of data accumulated through routine clinical care the authors aim to develop a machine learning algorithm to automate mapping of unlabeled data and reclassification of incorrect mappings within labeled data. For this purpose, inpatient and outpatient laboratory data (6.6 billion laboratory results) from 130 Veterans Affairs (VA) hospitals was collected. The dataset used for training contained an unknown number of labelling errors and was not manually cleaned (noisy labelling approach). They implemented logistic regression, a random forest multiclass classifier, and a 1-versus-rest ensemble of binary random forest classifiers. All models were refined with a 5-fold cross-validation. Although the mapping results of the models are not in all cases convincing, the authors investigated important reasons for this matter. In addition, they were able to prove that their results are similar in accuracy to the best reported automated methods for laboratory test mapping. This is a remarkable result as the model was built using noisy data. The approach is one of the first that can be applied fully automated with no need of manual intervention.

Xiao C, Ma T, Dieng AB, Blei DM, Wang F

Readmission prediction via deep contextual embedding of clinical concepts

PLoS One 2018;13(4):e0195024

Hospital readmission is a critical figure in the assessment of the quality of a treatment process. Indeed, many hospital readmissions are avoidable and pose a certain risk for the patient. Although it is not an easy task to predict hospital readmission as it is not only related to the disease and the treatment but also a complex set of risk factors which are interrelated. The current paper aims at presenting a hybrid deep learning model structure that combines topic modelling and Recurrent Neural Network (RNN) to distill the complex knowledge hidden in those contexts and perform accurate readmission prediction. The proposed 'CONTENT' model covers both the global and local contexts within the patient journey from an EHR through a hybrid Topic Recurrent Neural Network (TopicRNN) model. It transforms patients' complicated event structures into deep clinical concept embedding, which can be viewed as a novel form of patient representation encoding the patient clinical conditions from both long and short terms. In order to build that model the EHR data (including disease, lab findings and medication codes) from a cohort of 5,393 patients with congestive heart failure was used. The model outputs a context vector for each patient, which characterizes his/her overall condition. The proposed model outperforms existing methods in readmission prediction (e.g. 0.6103 ± 0.0130 vs. second best 0.5998 ± 0.0124 in terms of ROC-AUC). The derived patient representations were further utilized for patient phenotyping. The subgroups allow a better understanding of different readmission risks in the cohort.