Digital Personalized Health and Medicine
L.B. Pape-Haugaard et al. (Eds.)

doi:10.3233/SHTI200176

Modulation of Medical Condition Likelihood by Patient History Similarity

Jonathan TURNER[a,1], Dympna O'SULLIVAN [b] and Jon BIRD[c]
[a] *City, University of London, London, UK*
[b] *Technological University Dublin, Dublin, Ireland*
[c] *University of Bristol, Bristol, UK*

Abstract. Introduction: We describe an analysis that modulates the simple population prevalence derived likelihood of a particular condition occurring in an individual by matching the individual with other individuals with similar clinical histories and determining the prevalence of the condition within the matched group. **Methods:** We have taken clinical event codes and dates from anonymised longitudinal primary care records for 25,979 patients with 749,053 recorded clinical events. Using a nearest neighbour approach, for each patient, the likelihood of a condition occurring was adjusted from the population prevalence to the prevalence of the condition within those patients with the closest matching clinical history. **Results:** For conditions investigated, the nearest method performed well in comparison with standard logistic regression. **Conclusions:** Results indicate that it may be possible to use histories to identify 'similar' patients and thus to modulate future likelihoods of a condition occurring.

Keywords. EHR; machine learning; clinical terminologies

1. Introduction

The advent of electronic information systems in recent decades has enabled a rapid increase in data creation and collection in many disciplines, not least in healthcare, where we now have large amounts of data stored in individuals' electronic health records. These data have been acquired during the care of individuals with a variety of aims: to improve the care of the individuals; to improve communications between individuals' care givers; to maintain records of care; and to enable appropriate charging to payers [1]. Electronic records generally now hold data dating back 30 years or more and exist in various healthcare enterprises and departments for a variety of purposes – e.g. booking systems, GP systems, lab test results, imaging – and contain a number of data items, such as demographic information, test results, diagnoses, treatments, clinic appointments and charging information. In the work presented here, a set of longitudinal primary care records was analyzed to see if patients with similar recorded histories of symptoms and diagnoses were likely to share similar future diagnoses. The analysis of Electronic Health Records (EHR) data for risk prediction is an active area of research [2] however scant work using solely primary care records data for general future conditions likelihood was found. Miotto et al [3] used secondary care EHR data to make such predictions for a number of conditions using deep learning, achieving an

[1] Corresponding author: Jonathan Turner, Centre for Health Informatics, City, University of London, Northampton Square, London EC1V 0HB, United Kingdom; E-mail: jonathan.turner@city.ac.uk.

F-score of 0.236 for their predictions. Weng et al [4] tested four machine learning algorithms (random forest, logistic regression, gradient boosting machines, neural networks) to predict cardiovascular risk. McCormick et al [5] describe a Bayesian hierarchical model for the selection of association rules, predicting future medical histories for individuals on the basis of common clinical histories. Wang et al [6] developed a multilinear sparse logistic regression for risk prediction with patient EHRs for joint risk prediction (two or more diseases that are related to each other in terms of sharing common comorbidities, symptoms, risk factors). In summary, other work to date has primarily been on secondary care records for patients being treated in hospital settings as part of single episodes; the work we describe here investigates the use of primary care records that record events over many years.

2. Objective

With many symptoms being common to multiple diseases, there is a challenge in producing an initial diagnosis or recommendation for diagnostic tests from a set of symptoms that could have been produced by a number of diseases. Often the initial choice of diagnosis or testing is based on a clinician's impression of the likelihood of that condition in a general population; however the opportunity may exist for modification of these likelihoods based on individuals' recorded histories. In this work we aim to investigate the potential of applying machine learning methods to medical histories, in order to see if we can predict whether an individual's chances of being diagnosed with a particular condition are modulated from general population prevalence by comparison of their previous diagnosis and symptom histories with others' diagnosis and symptom histories. Jenssen and Kenyon note that 'medical decision-making is about assessing risk and weighing competing probabilities in the setting of imperfect information' [7]. We used coded histories from primary care records as the basis for this work. Primary care records, in general, utilize coding systems to store diagnosis and other information about patients. Common coding systems used for these records include ICD9, the dominant system in primary care in the USA and now being replaced by ICD10; and the Read Code system, the dominant system in the UK but now being replaced by SNOMED CT. Data used in this study were derived from UK patients, coded using Version 3 of the Read Codes (CTV3), and US patients, coded using ICD9. We chose CTV3 as the primary coding system. Some data required translation from ICD9 to CTV3 using a look-up table created from existing look-up tables for ICD9 to SNOMED CT from NIH [8] and for SNOMED CT to Read CTV3 from NHS Technology Reference Data Update Distribution Service [9].

Each CTV3 term is associated with a single concept and consists of a 5 byte code with an optional 2 byte. Codes comprise letters, numbers and trailing decimal points and fall into one of three categories: findings and diagnoses; processes; and medications. All codes are arranged in a hierarchy with a table listing all parent-child relations.

Figure 1. An example of codes in the CTV3 hierarchy.

Figure 1 shows a small extract from the

CTV3 hierarchy, illustrating the relationship between codes relating to bacterial chest infections. This Figure shows that Acute haemophilus influenza bronchitis is a Haemophilus influenzae infection; there are four other codes that are child codes of Haemophilus influenzae infection. Haemophilis influenzae infection itself is a Haemophilus infection, which in turn is a Bacterial disease. Bacterial disease is an Infective disorder, a child code of Clinical findings, a category coming under the root node of the CTV3 tree.

3. Methods

The data sources were sets of longitudinal clinical records derived from general practice data across the UK and USA, de-identified prior to being made available to us for this work. Data were presented as a set of records of events, including patient age, gender, and codes of events including diagnoses, symptoms, administrative events and prescriptions, together with the year of the event and history of smoking and alcohol consumption. There were 25,979 patients in the full data set (49.3%M, 50.7%F), with a total of 749,053 recorded items, of which 375,312 items were diagnostic events. Age distribution and prevalence of common conditions (as determined by [10]) in the full data set was not different to that present in the general population. For this work, only diagnosis and symptom codes and those codes giving information about smoking and alcohol consumption were retained from the records in the data set. The Konstanz Information Miner (KNIME) [11] was used to prepare the data and merge the data sets; statistical analysis was performed using R [12]. All software was run on an Asus N56VM laptop with Intel Core i7-3610QM 2.30 GHz CPU with 8GB RAM running 64-bit Windows 10.

For each record, the codes retained were used to determine its k nearest neighbour (KNN) records by use of the binomial method [13] using a program written in R for this work. Optimum values for k, the number of nearest neighbours, were established for each condition by experimentation on a training set of records removed from the complete data set, where the F-measure of accuracy was optimized over a range of values for k. The most effective depth in the CTV3 tree at which to aggregate codes from deeper levels was similarly determined by experiment. Calculations were repeated for a set of medical conditions. The process was run separately for each condition, with the condition of interest hidden from the information used for the distance calculation. For each record, the prevalence of the condition in the KNN was compared to the prevalence in the remainder of the data set; those records where the prevalence in the neighbours was significantly ($p < 0.05$) higher than in the remaining records, were assigned a prediction of positive for the condition of interest, otherwise the prediction was negative. Predictions were then compared to the actual presence in each record.

4. Results

An example is given for a single condition, allergic rhinitis. A Fagan nomogram [14] of the changed probability of an individual having the condition is illustrated in Figure 2. The prior probability of an individual having the condition, i.e. the population prevalence, is 0.178. After implementing the nearest neighbours algorithm, (the 'test') those predicted to have the condition had a probability of having the condition of 0.252.

The sensitivity of the test was 0.599; the specificity was 0.614. Results for six conditions are shown in Table 1. The results include comparison with a simple logistic regression (LR) method, chosen as an established prediction method [2].

KNN generally performed better than LR when judged by F-score, sensitivity and specificity, for all conditions. Low prevalence conditions performed poorly with LR, predicting no positive cases. Because of LR's better performance in positive likelihood ratio, it also gave higher results for odds ratio than KNN, suggesting that if LR predicts positive for a condition then there is greater confidence in the record containing that condition than if KNN predicts positive. However, if LR predicts negative for a record then there is less confidence that the record does truly not contain the condition than if the KNN predicted negative. F1 scores achieved for the more common conditions compared well with those achieved by Miotto et al [3]. Additionally, the KNN method presents an explainable method that may be more convincing than other machine learning methods in a clinical setting [4].

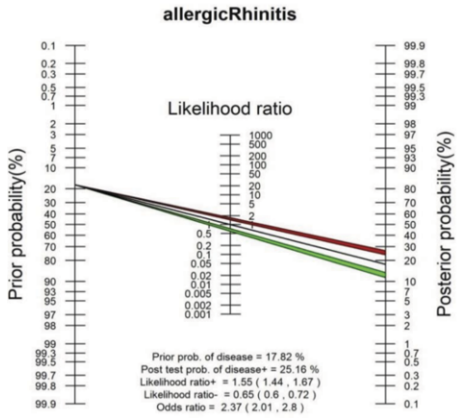

Figure 2. Results for allergic rhinitis using k-nearest neighbours.

Table 1. Comparison of KNN and LR prediction methods for six conditions.

Condition	Number in data set	Method	F1 Score	F2 Score	Sensitivity	Specificity	Positive Likelihood Ratio	Negative Likelihood Ratio	Odds Ratio
Allergic Rhinitis	2297	KNN	0.354	0.469	0.599	0.614	1.550	0.654	2.37
		LR	0.113	0.080	0.067	0.974	2.518	0.959	2.626
Bronchitis	2004	KNN	0.328	0.433	0.552	0.661	1.625	0.679	2.394
		LR	0.139	0.981	0.223	0.164	7.240	0.881	5.146
Obesity	2000	KNN	0.334	0.443	0.566	0.670	1.713	0.648	2.643
		LR	0.409	0.326	0.287	0.979	13.422	0.729	18.420
Gastroparesis	26	KNN	0.017	0.039	0.286	0.944	5.06	0.757	6.684
		LR	0	0	0	0.999	0	0	0
Autism	25	KNN	0.016	0.034	0.167	0.970	5.493	0.854	6.392
		LR	0	0	0	1	0	0	0
Eczema	670	KNN	0.229	0.369	0.623	0.791	2.984	0.476	6.265
		LR	0.137	0.095	0.079	0.996	21.649	0.925	23.413

5. Conclusions and Future Work

The results of our analysis indicate that it is possible to apply nearest neighbour techniques using medical histories to modulate future likelihoods of a condition occurring in individual patients. The technique employed in this work showed

performance improvements over logistic regression methods particularly for conditions of low prevalence, although there were limitations due to the relatively small data set size. This method is suggested as a useful low-cost screening tool. Early discovery of a condition can often improve the chances of successful treatment or of mitigation of the effects of the condition and many conditions can be detected through screening tests on asymptomatic individuals. Disadvantages of conventional screening tests include risks of harm to the individual if the screening test is invasive; inconvenience and/or stress to the individual; and financial cost to the individual and/or the health care system. It is therefore advantageous to perform screening tests on subsets of the population selected to be at great risk of the condition being screened for [15].

We intend to extend our analysis in a number in a number of areas. We will repeat the analysis using the existing methods but translating event codes to SNOMED CT and comparing results to those described here. We intend to optimize the methods used to address the slow run time experienced in this analysis. Run time has implications for use in clinical practice: currently the system takes around 30 minutes to run on a sample of 5,000 records. This is likely to be too long for opportunistic likelihood predictions for an individual arriving in a clinic, although acceptable for use in identification of individuals appropriate for screening or invitation to consultation.

References

[1] Coiera E. *Guide to Health Informatics.* CRC Press, London, 2015.

[2] Christodouloua E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calsterad B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology* **110** (2019): 12-22.

[3] Miotto R, Li L, Kidd BA, Dudley JT. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Scientific Reports* **17** (2016),6:26094.

[4] Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can Machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS ONE* **12** (2017), e0174944.

[5] McCormick T, Rudin C, Madigan D. Bayesian Hierarchical Rule Modeling For Predicting Medical Conditions. *Annals of Applied Statistics*, **6** (2012), 652-668.

[6] Wang X, Wang F, Hu J, Sorrentino R. Exploring joint disease risk prediction. Proceedings of the Annual Symposium of American Medical Informatics Association (AMIA), 2014.

[7] Jenssen BP, Kenyon CC. Probability, Uncertainty, and Value in Inpatient Diagnosis: Connecting the Dots. *Hospital Pediatrics,* **5** (2015) 403-405

[8] US National Library of Medicine. *ICD-9-CM Diagnostic Codes to SNOMED CT Map, Version 2016* (2016). Available at: https://www.nlm.nih.gov/research/umls/mapping_projects/icd9cm_to_snomedct.html (accessed 3rd January 2018).

[9] Health and Social Care Information Centre. *NHS Data Migration Releases.* (2013). Available at https://isd.digital.nhs.uk/trud3/user/guest/group/0/pack/9/subpack/9/releases (accessed 13th January 2020).

[10] Practice Fusion. *The 25 Most Common Medical Conditions.*, 2016. Available at https://www.practicefusion.com/blog/25-most-common-diagnoses/) (accessed 13th January 2020).

[11] Berthold MR, Cebron N, Dill F, Gabriel TR, Kotter T, Meini T, et al. KNIME: The Konstanz Information Miner. In Preisach C., Burkhardt H., Schmidt-Thieme L., Decker R. (eds): Data Analysis, Machine Learning and Applications. *Studies in Classification, Data Analysis, and Knowledge Organization.* Berlin: Springer; 2008. p. 26-31.

[12] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available from http://www.R-project.org/ (accessed 10th October 2018).

[13] Beretta L, Santaniello A. Nearest neighbor imputation algorithms: a critical evaluation. *BMC Medical Informatics and Decision Making* **16** (2016), 197-208.

[14] Fagan TJ. Nomogram for Bayes theorem. *New England Journal of Medicine* **293** (1975), 257.

[15] Hobbs F. Cardiovascular disease: different strategies for primary and secondary prevention? *Heart* **90** (2010), 1217-1223.