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

doi:10.3233/SHTI200195

Supervised Bayesian Statistical Learning to Identify Prognostic Risk Factor Patterns from Population Data

Colin J. CROOKS[a,1]
[a] *University of Nottingham, UK*

Abstract. Current methods for building risk models assume averaged uniform effects across populations. They use weighted sums of individual risk factors from regression models with only a few interactions, such as age. This does not allow risk factor effects to vary in different morbidity contexts. This study modified a supervised Bayesian statistical learning method of topic modelling, allowing individual factors to have different effects depending on a patient's other comorbidity. This study used topic modelling to assess more than 71,000 unique risk factors in a population cohort of 1.4 million adults within routine data. The model learnt prognostically important risk factor patterns that predicted 5 year survival, and the resulting model achieved excellent calibration and discrimination with a C statistic of 0.9 in a held out validation cohort. The model explained 92\% of the observed variation in 5 year survival in the population. This paper validates using survival supervised Bayesian topic modelling within large routine electronic population health data to identify prognostically important risk factor patterns.

Keywords. Latent Dirichlet allocation, topic modelling, Bayesian modelling, electronic health records, co-morbidity

1. Introduction

To understand the effects of large numbers of individual factors, and how these vary in different combinations, requires a different approach from weighted linear summations of small numbers of individual factors common in current prediction models. Methods are needed in which risk factors are modelled by their observed patterns, rather than as individual factors. Bayesian statistical learning methods using topic modelling provide a more interpretable alternative to current machine learning methods for identifying unknown patterns in data[1] By using mixtures of risk factor patterns, risk factor effects can vary in different patients depending on the co-occurrence of other factors. This reflects the heterogeneity observed in the general population. Importantly, it has an advantage of transparency through the interpretable intermediate step of topics constructed from risk factor patterns. Topic modelling also has computational advantages through Bayesian learning, as it can process larger numbers of codes, whilst it incorporates priors to reduce over fitting where the data is sparse. Importantly, topic

[1] Corresponding Author: Colin J. Crooks, NIHR Nottingham Biomedical Research Centre at the Nottingham University Hospitals NHS Trust and the University of Nottingham, and Nottingham Digestive Diseases Centre, School of Medicine, University of Nottingham, NG7 2UH, UK; E-mail: colin.crooks@nottingham.ac.uk

modelling can be supervised by outcomes, including survival time [2, 3]. The aim of this study was to assess the feasibility of using survival supervised topic modelling to predict mortality from unselected population based routine health data through learning prognostically important morbidity patterns.

2. Methods

The study cohort consisted of all adults in the English population over 18 years who were alive and registered on 1st January 2010 to the Clinical Practice Research Datalink (CPRD, protocol number 16_269R was approved by the Independent Scientific Advisory Committee (ISAC)). All their routine primary care (diagnostic Read codes from the CPRD clinical file) and secondary care (ICD 10 diagnostic codes from linked Hospital Episodes Statistics (HES)) data recorded prior to 1st November 2009 were included in the study with no filtering. The two month exclusion window avoided including codes related to a final illness event. All deaths from this cohort were identified for the study outcome between 1st January 2010 and 1st January 2015. Each diagnostic code was weighted with the frequency with which it was recorded in the data for the year prior to 1st November 2010. Age was adjusted within the baseline hazard by setting age as the origin of follow up time. The cohort was divided randomly into a 50% training cohort, 25% testing cohort and 25% validation cohort using simple random sampling.

2.1. Bayesian statistical learning model - Survival Supervised Topic Modelling

Blei *et al* developed the initial Latent Dirichlet allocation method for categorising documents based on associating word frequency combinations with unobserved (or latent) topics[1], and observed topics (supervised allocation)[2]. For this study topic modelling was supervised by censored survival data, using cumulative risks predicted from survival models as previously described [3]. However, to scale this up to large but noisy and sparse electronic health record data, with tens of thousands of unique codes, this study integrated 2 further modifications into the algorithm: 1) An asymmetric Dirichlet prior that was learnt from the data, promoting uninformative but frequently occurring codes to be allocated into uninformative baseline topics, 2) A penalised Cox proportional hazards model fitted via a coordinate descent algorithm to improve stability, with automated centering of topics based on the current topic allocations.

2.2. An asymmetric Dirichlet Prior learnt from the data

High frequency but uninformative terms can, by their association with other terms, become allocated to informative topics. To reduce this an asymmetric prior for document level topic distributions was learnt from the word topic allocations [5]. This allowed some topics within the model to automatically form baseline topics with a higher prior Dirichlet probability, thus promoting patients with few or uninformative risk factors to be allocated to these baseline topics along with their uninformative risk factor combinations. This topic prior was given a Dirichlet distribution learnt using the Newton-Raphson method [5]. The baseline topic with the highest prevalence, i.e. the highest number of words assigned to it was allocated a zero coefficient in the Cox model for each iteration. This aided the convergence of the Cox model by keeping the topic simplex centered.

2.3. A penalised proportional hazards model via a coordinate descent algorithm

To improve stability a penalty was introduced to the Cox model, shrinking coefficients towards the null, reducing instability, and preventing over-fitting where the data are sparse. Within the learning algorithm each topic is required to retain some level of probability, therefore L2 regularisation (or ridge regression) rather than L1 (Lasso) weighting was used [4]. Cyclic coordinate descent was used to fit the Cox model and the latent Dirichlet algorithm, as the large number of parameters was prohibitive for matrix inversion [4]. This also allowed coarse parallelisation of the document level updates.

3. Results

The topic model algorithm described above was applied to an unselected population cohort of training 712,868 patients derived from linked primary and secondary cared data from the English Clinical Practice Research Datalink for a range of numbers of topics from 10 to 200. The model was fitted to predict five year survival from patients alive on the 1st January 2010 using over 71,141 unique diagnostic codes available prior to 1st November 2009. The model converged on average in under 20 iterations. In addition to varying the topic number, a range of L2 penalties was assessed from 10^{-4} to10^{4}, with the best fitting selected by out of sample likelihood in cross validation during the maximisation step.

3.1. Model parameter selection

The model learnt in the training cohort was applied to the testing cohort of 356,482 patients, and the perplexity calculated as a measure of model fit perplexity is shown in figure 1. Perplexity is an average of the inverse likelihood per code, and the lower the perplexity the better the model fit. Figure 1 shows that model fit increased rapidly with an increase in topic number, but after 90 topics this plateaued. A fixed penalty of 0.01 was the minimal value that aided convergence in the test data and was selected as the L2 penalty. Fifty three of the topics were significantly associated with survival (p<0.05 likelihood ratio test), and the remainder were null and uninformative.

The topics generated by the latent Dirichlet allocation demonstrated recognisable combinations of risk factors that cut across different disease categories to define clinically recognisable frailty and multi-morbidity types. Summaries of three of selected topics that were strongly associated with survival are shown in (table 1).

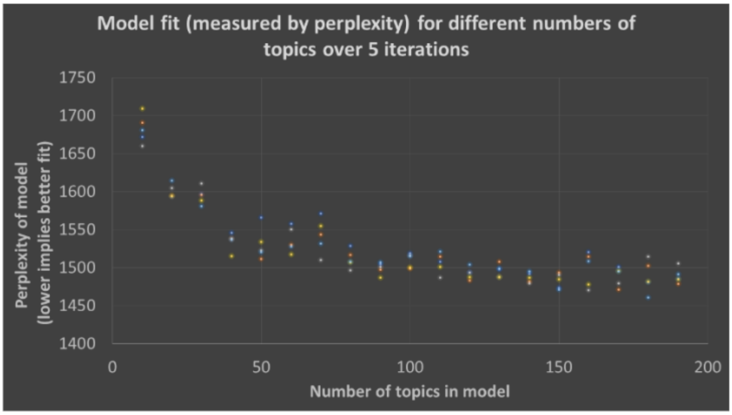

Figure 1. Model fit by number of topics in the testing dataset (lower is better)

Table 1. ICD and Read code descriptors with high probabilities for three example learnt topics that were strongly associated with survival

Summary	Cardiac disease, housebound, limited mobility and constipation	Cataract, falls, osteoporosis, fractures and UTI,	Chronic leg ulcers and nursing care
Read and ICD 10 code descriptors	• Congestive heart failure	• Hypertension	• Dressing of wound
	• Syncope and collapse	• Cataract	• Nursing care - dressing
	• AF & flutter	• High BMI	• Leg ulcer NOS
	• COPD	• Fall-accidental	• Cellulitis NOS
	• Senility	• Dementia	• Dressing of ulcer
	• Living alone	• UTI	• Wound healing
	• Constipation	• Osteoporosis	• Change of dressing
	• Housebound	• Fracture neck of femur	• Wound observation

3.2. Validation and calibration

The 90 risk factor topics that were learnt by the model in the training cohort were applied to the held out validation cohort (n=356,473 patients) using the inference algorithm as described by Blei *et al* [1], adjusted for age (in 10 year age bands) and gender. This model demonstrated a 1 year discrimination with a C statistic of 0.925 (95% confidence interval 0.921-0.929), which reduced to 0.908 (0.906-0.910) by 5 years. The average observed survival for each centile of the predicted topic model is plotted in figure 2. This showed excellent calibration across the range of predicted survival. 92% of the observed variation in 5 year survival was explained by the model (R^2 adjusted for censoring).

4. Conclusion

Topic modelling provides a novel Bayesian statistical learning approach to learn prognostic patterns of risk factors from routine health data, and this study demonstrates it can be successfully adapted to censored electronic health records. This study confirmed the importance of using risk factor patterns to model risk rather than individual risk factors, and demonstrated how these different patterns of coding can be captured automatically. One limitation is the model assumed learnt patterns were independent from each other, and further work will incorporate a correlated topic model. However, the current method identified clinically recognisable risk factor patterns of multi-morbidity and frailty that would not be included in standard co-morbidity scores. This approach could be applied to identify cross cutting morbidity patterns that are necessary for prediction models with the increasing complexity of multi-morbidity in the population. The C++ code underlying this paper can be accessed at doi.org/10.5281/zenodo.1045521 or github.com/ColinCrooks/SurvivalSupervisedLDA.

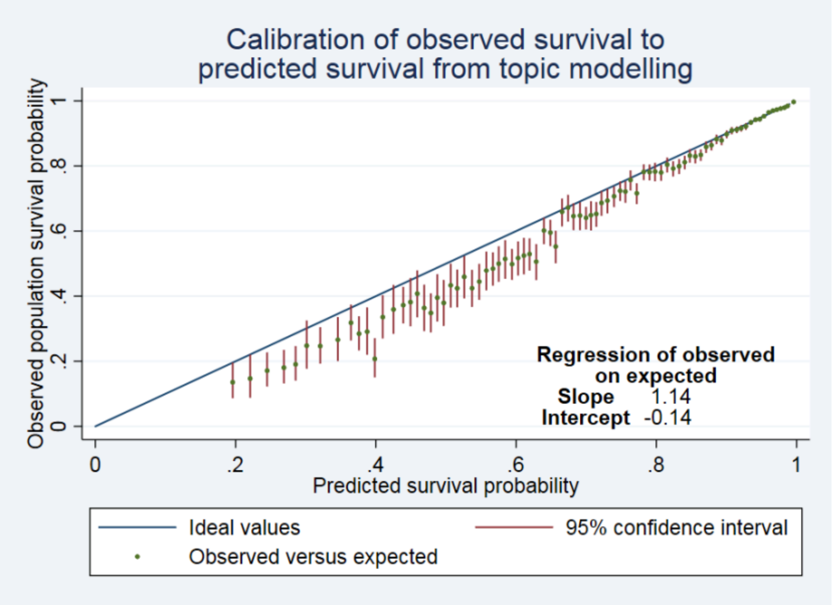

Figure 2: Calibration curve of the fitted model to the held out validation cohort

References

[1] Blei DM, Ng A, Jordan M. Latent Dirichlet Allocation. The Journal of Machine Learning Research. 3 (2003), 993–1022. doi:10.5555/944919.944937.

[2] Blei DM, McAuliffe JD. Supervised Topic Models. Advances in Neural Information Processing Systems. 2 (2007), 121–128.

[3] Ye S, Dawson JA, Kendziorski C. Extending information retrieval methods to personalized genomic-based studies of disease. Cancer informatics. 13(Suppl 7) (2014), 85–95. doi:10.4137/CIN.S16354.

[4] Mittal S, Madigan D, Burd RS, Suchard Ma. High-dimensional, massive sample-size Cox proportional hazards regression for survival analysis. Biostatistics. 15(2) (2014), 207–221. doi:10.1093/biostatistics/kxt043.

[5] Wallach HM, Mimno D, Mccallum A. Rethinking LDA : Why Priors Matter. Advances in Neural Information Processing Systems. 22(2) (2009) 1973–1981. doi:10.1007/s10708-008-9161-9.