[Note: All submissions must be in PDF format. Failure adhere to the format requirements may result in rejection of your submission without review]

Name:	Jenna Marie Reps
Affiliation:	Janssen
Email:	jreps@its.jnj.com
Presentation type (select one):	Poster

A framework to efficiently identify potential prognostic factors

Jenna M. Reps, PhD, M. Soledad Cepeda, MD, PhD
Janssen Research and Development, Raritan, NJ

Abstract

One of the main usages of prognostic models is to gain insight into an illness by helping identify causal prognostic factors which can aid the understanding of the mechanisms behind illness development. Methods such as regularized logistic regression can filter through thousands of variables and pick out those that are most informative in predicting the outcome. Combining data driven methods that pick out the informative variables with clinical review may help identify new prognostic factors. In this paper a data-driven framework that can use the OHDSI data network to combine multiple database perspectives to identify potential new prognostic factors for a specific illness is proposed and implemented for the outcome of rapid osteoarthritis development.

Introduction

Prognostic models are models that predict the risk of some future outcome during a certain time period for a cohort of people. The primary use of prognostic models is to calculate the personalized risk of the outcome for each individual, but these models can be an excellent way to learn about the illness[1] as they can highlight the variables that are associated to the outcome (prognostic factors). If clinicians become aware of causal prognostic factors, then they can use this knowledge to develop hypotheses that might result in new insights about the illness mechanisms. In addition, knowledge of causal prognostic factors could be used for disease interception[1].

Conventional prognostic models only consider a small number of expert pre-specified variables, so rather than identifying new prognostic factors, these models just give insight into how useful each variable is in predicting the risk. With more advanced methods, such as regularized logistic regression[2], it is now possible to include thousands of variables and let the data/model pick the most informative variables[3]. By implementing this approach, it may be possible to identify new prognostic factors. Unfortunately, many variables selected by the regularized regression models may be selected due to overfitting and will not necessarily be predictive in new data. Clinically useful prognostic factors are likely to be found consistently across datasets.

In this paper a multi-step framework for identifying key prognostic factors is proposed. The first step is to perform the prediction analyses across multiple datasets. This required applying lasso logistic regression using very broad variables to predict the illness across a number of datasets and then combining the models' variable importances to find variables (the prognostic factors) that are consistently picked as being informative in terms of predicting the illness. Broad variables are used as these are likely to be more stable across datasets (e.g., if specific concepts were used then these concepts may not be consistently used across all the datasets). Finally, the refined variables consistently selected by the logistic regressions are then presented to a clinician for review and the clinician can use his or her expertise to pick out which variables should be further studied.

Prediction problem to evaluate framework:

The prediction problem chosen to test the framework is predicting the 2-year risk of intra-articular injection of the knee. The at risk cohort is people who have an outpatient visit in 2008 or after (inclusion: age >=18, min 180 days prior observation and min 730 days post observation, exclusion: prior knee injection/replacement or inflammatory conditions), the outcome cohort is people who have a procedure of intra-articular injection of the knee. The risk period is 2-years within the first outpatient visit that satisfies the inclusion/exclusion.

For each dataset prediction model we train a lasso regularized logistic regression model and perform two fold cross validation for hyper-parameter selection. We split the data into 30% test data and 70% train data and report the area under the ROC curve and number of variables selected by the model, see Table 1.

Proposed Framework:

- **1) Exploration**: (find important variables in each dataset) Develop lasso logistic regression prediction models on numerous datasets using a large number of broad variables (e.g. Medra 'HLGT' level concepts and ATC 2nd) and determine variable importance.
- **2) Replication:** (check for variable importance consistency) find all the variables that are select by all datasets' models and have a consistent direction (coefficient value is positive or negative consistently).
- **3) Clinical Review:** clinician reviews variables and picks some to be further evaluated

Results & Discussion

The performance of each model trained in step 1 and number of variables selected are presented in Table 1.

Database	*AUC*	*Variables with positive coefficient*	*Variables with negative coefficient*
Truven CCAE	0.74	177	235
Truven Medicare	0.62	169	244
Truven Medicaid	0.76	105	175
Optum	0.75	166	258

Table 1. The prediction performance on the test set.

There were 41 positive coefficient variables and 55 negative coefficient broad term variables selected consistently across datasets. A selection of these variables is presented in Table 2.

Variable	*Direction*
Appetite and general nutritional disorders (Obesity)	**Increased risk**
Malabsorption conditions	Increased risk
Procedure -Cardiac and vascular investigations (excl enzyme tests)	Increased risk
Venous varices	Increased risk
Mood disorders and disturbances NEC	Increased risk
Ingredient group with parent: ANTIGOUT PREPARATIONS	**Increased risk**
Iron and trace metal metabolism disorders	Increased risk
Gastrointestinal infections	Increased risk
Bone disorders (excl congenital and fractures)	Decreased risk
Autoimmune disorders	Decreased risk
Musculoskeletal and soft tissue investigations (excl enzyme tests)	Decreased risk
Psychiatric disorders NEC	Decreased risk
Bone, calcium, magnesium and phosphorus metabolism disorders	**Decreased risk**
Gender = MALE	**Decreased risk**
Age group: 35-39	**Decreased risk**
Thyroid gland disorders	Decreased risk
Lipid metabolism disorders	Decreased risk
Aneurysms and artery dissections	Decreased risk

Table 2. The prognostic factors identified by applying the framework, known prognostic factors are in bold.

The results show that many of the known prognostic factors, such as obesity, were correctly identified by the framework but many variables that are not known to be prognostic factors were also identified.

Conclusion

In this paper we presented a novel framework that can utilize a network of datasets to efficiently identify potential prognostic factors. These new prognostic factors can be reviewed by a clinical expert to gain insight into the mechanism behind the illness progression/development. In future work the identified potential risk factors could be evaluated using estimation methods to determine whether they are causal or not.

References

1. Riley, Richard D., et al. Prognosis Research Strategy (PROGRESS) 2: prognostic factor research. *PLoS Med* 2013:10.2;e1001380.

2. Tibshirani, Robert. Regression shrinkage and selection via the lasso.*Journal of the Royal Statistical Society. Series B (Methodological)* 1996: 267-288.
3. Steyerberg, Ewout W., et al. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Statistics in medicine* 2000:19.8;1059-1079.