Computers in Biology and Medicine 54 (2014) 199–210

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Risk factors and prediction of very short term versus short/intermediate term post-stroke mortality: A data mining approach

Jonathan F. Easton [a], Christopher R. Stephens [b,c], Maia Angelova [a,*]

[a] Mathematical Modelling Lab, Department of Mathematics and Information Sciences, Northumbria University, Newcastle upon Tyne NE1 8ST, UK
[b] Instituto de Ciencias Nucleares, Universidad Nacional Autonoma de Mexico, Mexico City 04510, DF, Mexico
[c] C3 - Centro de Ciencias de la Complejidad, Universidad Nacional Autonoma de Mexico, Mexico City 04510, DF, Mexico

ARTICLE INFO

Article history:
Received 26 March 2014
Accepted 5 September 2014

Keywords:
Data mining
Naïve Bayes analysis
Risk factors
Prediction
Stroke
Mortality
Medical relevance

ABSTRACT

Data mining and knowledge discovery as an approach to examining medical data can limit some of the inherent bias in the hypothesis assumptions that can be found in traditional clinical data analysis. In this paper we illustrate the benefits of a data mining inspired approach to statistically analysing a bespoke data set, the academic multicentre randomised control trial, UK Glucose Insulin in Stroke Trial (GIST-UK), with a view to discovering new insights distinct from the original hypotheses of the trial. We consider post-stroke mortality prediction as a function of days since stroke onset, showing that the time scales that best characterise changes in mortality risk are most naturally defined by examination of the mortality curve. We show that certain risk factors differentiate between very short term and intermediate term mortality. In particular, we show that age is highly relevant for intermediate term risk but not for very short or short term mortality. We suggest that this is due to the concept of frailty. Other risk factors are highlighted across a range of variable types including socio-demographics, past medical histories and admission medication. Using the most statistically significant risk factors we build predictive classification models for very short term and short/intermediate term mortality.

1. Introduction

Healthcare and medicine are two important areas where the revolution in data storage and analysis has permitted access to large quantities of data which can in principle be analysed for multiple purposes. Two important scenarios in which data are collected are (i) data associated with the clinical histories of patients – this is of relevance, for instance, to healthcare providers, insurers and healthcare authorities; (ii) data associated with clinical trials, where an intervention, such as a new drug or treatment regime, is tested on a particular group and outcomes measured. In case (i) the focus is on the individual, or a cohort of similar individuals, whereas in (ii) it is on the cohort itself or, rather, the intervention.

The existence of large electronic data bases is of course not unique to medicine. The data is typically of large volume and of very high dimension. This presents several complicating factors for traditional statistical modelling, and so, data mining, or knowledge discovery in data bases, as a general research area [1], has developed as a useful form of statistical analysis alongside the exponential explosion of data and is now being used extensively. However, its impact in medicine and healthcare [2] has been less, and slower, than one might expect or hope. Part of the reason for this is associated with privacy issues and also the large data fragmentation often inherent in medical data. However, another potential reason, as hinted at in [3], is that medicine is largely hypothesis driven, such as in testing an intervention or treatment. Indeed, this is the case with the data we examine in this paper. On the other hand, much of data mining is associated with knowledge discovery, where an a priori specific hypothesis to be examined does not necessarily exist, allowing exploration of the data to provide interesting results. In this sense, the present study is an observational study, where it is not necessarily feasible or the goal to infer causality. Rather, by revealing interesting, previously unknown, patterns in the data one can establish more detailed hypotheses about causal relationships that can be further tested, using controlled trials for example.

In this paper we will illustrate the potential benefits of a more data mining inspired approach, examining and analysing a bespoke

[*] Corresponding author. Tel.: +44 191 243 7611.
E-mail addresses: jonathan.easton@northumbria.ac.uk (J.F. Easton), stephens@nucleares.unam.mx (C.R. Stephens), maia.angelova@northumbria.ac.uk (M. Angelova).

http://dx.doi.org/10.1016/j.compbiomed.2014.09.003

data set, the academic multicentre randomised control trial, UK Glucose Insulin in Stroke Trial (GIST-UK) [4], with a view to discovering potentially new insights distinct from the original hypotheses associated with the trial. A discussion of the data can be found in Section 2. Due to the controlled nature of the data set used here, this case study will highlight the use of data mining techniques in this field of healthcare and the potential it offers. It is important to note however, that, due to the inclusion/exclusion criteria associated with GIST-UK, it is not necessarily the case that our conclusions can be taken over, as is, to a wider stroke population. Rather, we would see this paper as advocating the data mining approach and statistical methodology, with a view to applying it to other analogous data sets and, particularly, more representative stroke data sets.

In general terms, we will concentrate on post-stroke mortality and the corresponding risk factors, developing predictive models for stroke mortality over different time horizons relative to admission – very short term, short term and intermediate term. In this sense, our probabilistic models can be considered to be clinical outcomes algorithms which would aid the decision making process of healthcare professionals and data mining analysis is proven to be very useful in work of this nature [5]. Probabilistic models for outcomes have been developed in other circumstances, for instance, in [6,7] for mortality among critically ill hospitalised adults, in [8] for kidney dialysis patients and also in the case of stroke [9]. Methodologically, we will take a classification type approach and use a naïve Bayesian classifier [1,10] which will be discussed in Section 6.

The paper is organised as follows, the data used in this study will be discussed in Section 2. In Section 3, by examining the overall mortality curve, we look at the classification of the distinct time intervals – very short, short, and intermediate term and offer our own classification to be used in examining the impact of risk factors. Age, in particular, as a risk factor will be seen to play a different role over different time horizons and is analysed in Section 4. Other risk factors are identified and discussed in Section 5 which are used in Section 6 to build predictive models with the potential to categorise the risk of mortality for stroke patients upon admission to hospital. A comparison to other data mining models including regression and decision trees is made in Section 7. Finally, conclusions are made in Section 8.

2. Data

The data used in this case study is associated with the academic multicentre randomised control trial, UK Glucose Insulin in Stroke Trial (GIST-UK) [4]. Patient confidentiality makes any simple, large scale analysis of clinical records difficult and so most studies involve analysis of controlled groups where there is a specific goal to the analysis. In this paper we will study a cohort of patients who have suffered an acute stroke under the conditions specified by the GIST-UK protocol.

The inclusion and exclusion criteria for patients in this study are highlighted here. Eligible patients had a fixed neurological deficit and no evidence of rapid improvement during a 60 min period after stroke onset. Patients were excluded from the trial with subarachnoid haemorrhage, isolated posterior circulation syndromes with no physical disability, pure language disorders, renal failure (urea >20 mmol/L or creatinine >200 μmol/L), anaemia (haemoglobin <9 g/dL), or coma (motor response ≤ 3 on the Glasgow coma scale). Also excluded were patients with an established history of insulin-treated diabetes mellitus, previous disabling stroke (modified Rankin scale [mRS] score >3), established history of dementia or abbreviated mental test score $<7/10$, or symptomatic cardiac failure (New York Heart Association grade III or IV) [4]. There were 933 patients in the study where mortality was determined over a 3 month period.

3. Stroke mortality as a function of days from stroke onset

As there is an elevated risk of mortality for an extended period after a stroke it is of interest to ask if the risk factors for mortality are constant across time. A relevant question is then: what are the appropriate time horizons for considering mortality? What might be considered short, intermediate or long term? Many analyses of stroke mortality as a function of number of days from stroke onset or days since admission have considered short term to be associated with periods of less than a month, with typical categorisations being short term <1 month, intermediate term 1–3 months and long term 3–12 months [11–14]. However, typical mortality curves as a function of days since admission show that between approximately 50% (ischaemic) and 70% (haemorragic) of stroke patients who will die in the intermediate term (up to 93 days) are already dead within 30 days [11]. With this in mind, we suggest that a more appropriate characterisation of short, intermediate and long term comes from an examination of the cumulative mortality curve itself. The curve for this data set can be seen in Fig. 1, where we see the cumulative proportion of deaths as a function of all deaths through the 3 month period in which mortality was measured. We would argue that the form of the curve suggests that immediacy of death should be more appropriately categorised by dividing the cumulative curve for death related to days from stroke onset into n time bins, such that an equal proportion $1/n$ of deaths is included in each bin.

First, we will consider four time bins, $n=4$, so that from Fig. 1 the number of days since stroke onset which corresponds to 25%, 50%, 75% and 100% of the total number of deaths will be observed. As can be deduced from the data of Fig. 1, the corresponding time frames for the four bins are 5 days (25%), 12 days (50%) and 33 days (75%). Four bins are used for the investigation in Section 4 as we are testing the hypothesis on a specific variable and are interested in the change in risk over small time scales. The second division will be mortality over 1–7 days and 8–93 days to be used for building predictive models. In this case, the motivation for choosing two mortality periods was so that the statistical significance of the results would be higher when investigating a wide range of variables. Also a week has an extra significance as a time period as opposed to 8 or 9 days. Given that short term is usually defined in the literature as <1 month, we refer to the periods 1–5 days as *very short* term, 6–33 days *short* term and 34–93 days *intermediate* term for the analysis of age alone. For the predictive model analysis 1–7 days will be termed *very short* term and 8–93 days as *short/intermediate* term.

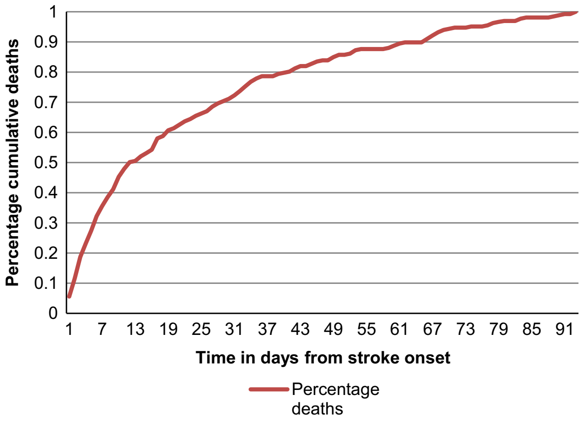

Fig. 1. Cumulative proportion of deaths as a function of days from stroke onset.