Computers in Biology and Medicine 75 (2016) 203–216

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Handling missing data in large healthcare dataset: A case study of unknown trauma outcomes

E.M. Mirkes [a], T.J. Coats [b], J. Levesley [a], A.N. Gorban [a,*]

[a] Department of Mathematics, University of Leicester, Leicester LE1 7RH, UK
[b] Emergency Medicine Academic Group, Department of Cardiovascular Sciences, University of Leicester, Leicester LE1 7RH, UK

ARTICLE INFO

Article history:
Received 1 April 2016
Received in revised form
1 June 2016
Accepted 2 June 2016

Keywords:
Missed data
Big data
Data cleaning
Mortality
Markov models
Risk evaluation

ABSTRACT

Handling of missed data is one of the main tasks in data preprocessing especially in large public service datasets. We have analysed data from the Trauma Audit and Research Network (TARN) database, the largest trauma database in Europe. For the analysis we used 165,559 trauma cases. Among them, there are 19,289 cases (11.35%) with unknown outcome. We have demonstrated that these outcomes are not missed 'completely at random' and, hence, it is impossible just to exclude these cases from analysis despite the large amount of available data. We have developed a system of non-stationary Markov models for the handling of missed outcomes and validated these models on the data of 15,437 patients which arrived into TARN hospitals later than 24 h but within 30 days from injury. We used these Markov models for the analysis of mortality. In particular, we corrected the observed fraction of death. Two naïve approaches give 7.20% (available case study) or 6.36% (if we assume that all unknown outcomes are 'alive'). The corrected value is 6.78%. Following the seminal paper of Trunkey (1983 [15]) the multi-modality of mortality curves has become a much discussed idea. For the whole analysed TARN dataset the coefficient of mortality monotonically decreases in time but the stratified analysis of the mortality gives a different result: for lower severities the coefficient of mortality is a non-monotonic function of the time after injury and may have maxima at the second and third weeks. The approach developed here can be applied to various healthcare datasets which experience the problem of lost patients and missed outcomes.

1. Introduction

Enthusiasm for the use of big data in the improvement of health service is huge but there is a concern that without proper attention to some specific challenges the mountain of big data efforts will bring forth a mouse [1]. Now, there is no technical problem with 'big' in healthcare. Electronic health records include hundreds of millions of outpatient visits and tens of millions of hospitalisations, and these numbers grow exponentially. The main problem is in quality of data.

'Big data' very often means 'dirty data' and the fraction of *data inaccuracies* increases with data volume growth. Human inspection at the big data scale is impossible and there is a desperate need for intelligent tools for accuracy and believability control.

The second big challenge of big data in healthcare is *missed information*. There may be many reasons for data incompleteness. One of them is in health service 'fragmentation'. This problem can be solved partially by the national and international unification of the electronic health records (see, for example, Health Level Seven International (HL7) standards [2] or discussion of the template for uniform reporting of trauma data [3]). However, some fragmentation is unavoidable due to the diverse structure of the health service. In particular, the modern tendency for personalisation of medicine can lead to highly individualised sets of attributes for different patients or patient groups. There are several universal technologies for the handling of missing data [4–10]. Nevertheless, the problem of handling missed values in large healthcare datasets is certainly not completely solved. It continues to attract the efforts of many researchers (see, for example, [11]) because the popular universal tools can lead to bias or loss of statistical power [12,13]. For each system, it is desirable to combine various existing approaches for the handling of missing data (or to invent new ones) to minimise the damage to the results of data analysis. For the best possible solution, we have to take into account the peculiarities of each database and to specify the further use of the cleaned data (it is desirable to understand in advance how we will use the

[*] Corresponding author.
E-mail addresses: em322@le.ac.uk (E.M. Mirkes), tc61@le.ac.uk (T.J. Coats), jl1@le.ac.uk (J. Levesley), ag153@le.ac.uk (A.N. Gorban).

http://dx.doi.org/10.1016/j.compbiomed.2016.06.004
0010-4825/© 2016 Elsevier Ltd. All rights reserved.

preprocessed data).

In our work we analyse missed values in the TARN database [14]. We use the preprocessed data for:

- the evaluation of the risk of death,
- the identification of the patterns of mortality,
- approaching several old problems like the Trunkey hypothesis about the trimodal distribution of trauma mortality [15].

The 'two stage lottery' non-stationary Markov model developed in the sequel can be used for the analysis of missing outcomes in a much wider context than the TARN database and could be applied to the handling of data gaps in healthcare datasets which experience the problem of transferred and lost patients and missing outcomes.

In this paper we analyse the unknown outcomes. The next task will be the analysis of missed data in the most common 'input' attributes.

2. Data set

There are more than 200 hospitals which send information to TARN (TARN hospitals). This network is gradually increasing. Participation in TARN is recommended by the Royal College of Surgeons of England and the Department of Health. More than 93% of hospitals across England and Wales submit their data to TARN. TARN also receives data from Dublin, Waterford (Eire), Copenhagen, and Bern.

We use TARN data collected from 01.01.2008 (start of treatment) to 05.05.2014 (date of discharge). The database contains 192,623 records and more than 200 attributes. Sometimes several records correspond to the same trauma case because the patients may be transferred between TARN hospitals. We join these records. The resulting database includes data of 182,252 different trauma cases with various injuries.

16,693 records correspond to patients, who arrived (transferred from other institutions) to TARN hospitals later than 24 h after injury. This sample is biased, for example the Fraction Of Dead (FOD) outcomes for this sample are 3.34% and FOD for all data is 6.05%. This difference is very significant for such a big sample. (If all the outcomes in a group of the trauma cases are known then we use the simple definition of FOD in the group: the ratio of the number of registered deaths in this group to the total number of patients there. Such a definition is not always applicable. The detailed and more sophisticated analysis of this notion follows in the next section.) We remove these 16,693 trauma cases from analysis but use them later for validation of the 'mortality after transfer' model. Among them, there are 15,437 patients who arrived at a TARN hospital within 30 days after injury. We call this group 'IN30' for short (Fig. 1).

As a result we have 165,559 records for analysis ('Main group'). This main group consists of two subgroups: 146,270 patients from this group approached TARN during the first day of injury and remained in TARN hospitals or discharged to a final destination during the first 30 days after injury. We call this group the 'Available within 30 days after injury' cases (or 'Available W30D' for short). The other 19,289 patients have been transferred within 30 days after injury to a hospital or institution (or unknown destination) who did not return data to the TARN system. We call them 'Transferred OUT OF TARN within 30 days after injury' or just 'OUT30' (Fig. 1).

The patients with the non-final discharge destinations 'Other Acute hospital' and 'Other institution' were transferred from a TARN hospital to a hospital (institution) outside TARN and did not return to the TARN hospitals within 30 days after injury.

The database includes several indicators for evaluation of the severity of the trauma case, in particular, Abbreviated Injury Scale (AIS), Injury Severity Score (ISS) and New Injury Severity Score (NISS). For a detailed description and comparison of the scores we refer readers to reviews [16,17]. The comparative study of predictive ability of different scores has a long history [18–21]. The scores are used for mortality predictions and are tested on different datasets [22–25]. In the database, there exist no gaps in AIS (and hence ISS and NISS) values even for patients rapidly dying. Most severely injured patients have a CT 'pan-scan' within the first hour or two of injury which is likely to define all life-threatening injuries. In addition the report from the post-mortem examination is used in the compilation of an injuries' list which is the basis of AIS, and hence ISS and NISS, scoring.

3. Definitions and distributions of outcomes

The widely used definition of the endpoint outcome in trauma research is survival or death within 30 days after injury [25–27].

A substantial number of TARN in-hospital deaths following trauma occur after 30 days: there are 957 such cases (or 8% of TARN in-hospital death) among 11,900 cases with 'Mortuary' discharge destination. This proportion is practically the same in the main group (165,559 cases): 894 deaths after 30 days in hospital (or 7.9%) among 11,347 cases with 'Mortuary' discharge destination.

Death later than 30 days after injury may be considered as caused by co-morbidity rather than the direct consequence of the injury [25]. These later deaths are not very interesting from the perspective of an acute trauma care system (as we cannot influence them), but they might be very interesting from the perspective of a geriatric rehabilitation centre or of an injury prevention program for elderly patients.

On the other hand, when 'end of acute care' is used as an outcome definition then a significant portion of deaths remains unnoticed. For example, in the 3332 trauma cases treated in the Ulleval University Hospital (Oslo, Norway, 2000–2004) 18% of deaths occurred after discharge from the hospital [27].

The question of whether it is possible to neglect trauma caused mortality within 30 days after trauma for the patients with the discharge destination 'Home', 'Rehabilitation' and other 'recovery' outcomes is not trivial [27]. Moreover, here are two questions:

- How do we collect all the necessary data after discharge within 30 days after trauma – a technical question?
- How do we classify the death cases after discharge within 30 days after trauma; are they consequences of the trauma or should they be considered as comorbidity with some additional reasons?

The best possible answer to the first question requires the special combination of technical and business process to integrate data from different sources. The recent linkage from TARN to the Office for National Statistics (ONS) gives the possibility to access the information about the dates of death in many cases. It is expected that the further data integration process will recover many gaps in the outcome data.

The last question is far beyond the scope of data management and analysis and may be approached from different perspectives. Whether or not the late deaths are important in a model depends on the question being asked. From the data management perspective, we have to give the formal definition of the outcome in terms of the available database fields. It is impossible to use the standard definition as survival or death within 30 days after injury because these data are absent. We define the outcome 'Alive