Big Data Research ••• (••••) •••–•••

Contents lists available at ScienceDirect

Big Data Research

www.elsevier.com/locate/bdr

Scalable Machine Learning for Predicting At-Risk Profiles Upon Hospital Admission ☆,☆☆

Pierre Genevès [a,*], Thomas Calmant [a], Nabil Layaïda [a], Marion Lepelley [b], Svetlana Artemova [b], Jean-Luc Bosson [b]

[a] Univ. Grenoble Alpes, CNRS, Grenoble INP, Inria, LIG, F-38000 Grenoble, France
[b] Univ. Grenoble Alpes, CNRS, Public Health Department CHU Grenoble Alpes, Grenoble INP, TIMC-IMAG, 38000 Grenoble, France

ARTICLE INFO

Article history:
Received 12 October 2017
Received in revised form 14 February 2018
Accepted 25 February 2018
Available online xxxx

Keywords:
Application
Big prescription data
Volume
Variety
Experiments

ABSTRACT

We show how the analysis of very large amounts of drug prescription data make it possible to detect, on the day of hospital admission, patients at risk of developing complications during their hospital stay. We explore, for the first time, to which extent volume and variety of big prescription data help in constructing predictive models for the automatic detection of at-risk profiles.

Our methodology is designed to validate our claims that: (1) drug prescription data on the day of admission contain rich information about the patient's situation and perspectives of evolution, and (2) the various perspectives of big medical data (such as veracity, volume, variety) help in extracting this information. We build binary classification models to identify at-risk patient profiles. We use a distributed architecture to ensure scalability of model construction with large volumes of medical records and clinical data.

We report on practical experiments with real data of millions of patients and hundreds of hospitals. We demonstrate how the fine-grained analysis of such big data can improve the detection of at-risk patients, making it possible to construct more accurate predictive models that significantly benefit from volume and variety, while satisfying important criteria to be deployed in hospitals.

1. Introduction

A major challenge in healthcare is the prevention of complications and adverse effects during hospitalization. A complication is an unfavorable evolution or consequence of a disease, a health condition or a therapy; and an adverse effect is an undesired harmful effect resulting from a medication or other intervention. Typical examples include for instance pressure ulcers, hospital–acquired infections (HAI), admissions in Intensive Care Unit (ICU), and death.

From the perspective of complications, healthcare establishments can be considered as risky environments. For instance, in the USA, an estimated 13.5% of hospitalized Medicare beneficiaries experienced adverse effects during their hospital stays; and an additional 13.5% experienced temporary harm events during their stays[1] [1]. However, physician reviewers determined that 44% of adverse and temporary harm events were clearly or likely preventable [1]. Preventable events are often linked to the lack of patient monitoring and assessment.

One challenging and very interesting goal is to be able to predict the patients' outcomes and tailor the care that certain patients receive if it is believed that they will do poorly without additional intervention. In doing so, hospitals could prevent unnecessary readmissions, adverse events, or other delays in getting well [2]. For instance, if we can precisely identify groups of patients associated with a very high risk of requiring ICU treatment during their stay, then we can optimize their placement as soon as they are admitted, by affecting them e.g. to rooms closer to ICU, thereby drastically reducing transportation delay in life-critical situations in large hospitals. More generally, many complications could be avoided by immediate identification of at-risk patients upon admission and adapted prevention. A crucial prerequisite to any adapted and meaningful prevention is the precise identification of at-risk profiles.

☆ This article belongs to Big Data for Healthcare.

☆☆ This research was partially supported by the ANR project CLEAR (ANR-16-CE25-0010).

* Corresponding author.
E-mail address: pierre.geneves@cnrs.fr (P. Genevès).
URL: http://pierre.geneves.net (P. Genevès).

[1] Temporary harm events are those that require intervention but do not cause lasting harm.

https://doi.org/10.1016/j.bdr.2018.02.004
2214-5796/

The widespread adoption of Electronic Health Records (EHR) makes it possible to benefit from quality information provided by healthcare professionals [3]. This opens the way for applying AI techniques in building helpful analytics systems for big medical data in which we can have a high level of trust – since drug prescriptions engage the responsibilities of healthcare professionals.

This paper aims to develop an automatic prediction system for identifying at-risk patients, based on a fine-grained analysis of large volumes of electronic health record data. This has long been viewed as a more challenging task than conventional prediction approaches with summary statistics and EHR-based scores [2,4].

Contributions We show how the analysis of very large amounts of drug prescription data make it possible to detect, on the day of hospital admission, patients at risk of developing complications during their hospital stay. We explore, for the first time, to which extent volume and variety of big prescription data help in constructing predictive models for the automatic detection of at-risk profiles. We report on practical experiments with real data of millions of patients and hundreds of hospitals. We demonstrate how the fine-grained analysis of such big data can improve the detection of at-risk patients, making it possible to construct more accurate predictive models that significantly benefit from volume and variety, while satisfying important criteria to be deployed in hospitals.

2. Methodology

Our methodology is designed to validate our claims that: (1) drug prescription data on the day of admission contain rich information about the patient's situation and perspectives of evolution, and (2) the various perspectives of big medical data (such as veracity, volume, variety) help in extracting this information.

We thus focus on building binary classification models to identify at-risk patient profiles, using distributed supervised machine learning methods. Our approach involves a fully distributed architecture to ensure scalability of model construction with large volumes of medical records and clinical data. The machine learning models that we build yield predictions at hospital admission time.

2.1. Considered medical data and veracity

We consider real data from United States Hospitals. Our dataset features more than 33 million discharges from a representative group of 417 hospitals drawn by lot, as provided by the Premier Perspective database, which is the largest hospital clinical and financial database in the United States. Each individual drug prescription engages the responsibility of the prescriber. Each hospital submits quarterly updates of aggregated data. Patient-level data go through 95 quality assurance and data validation checks. Once the data have been validated, patient-level information is available, comprising data consistent with the standard hospital discharge file, demographic and disease state information, and information on all billed services, including date-specific logs of medications, laboratory, diagnostics, and therapeutic services.

The raw data for the year 2006 contains 33 048 852 admissions, and more than three billion patient charge records, representing 2.8 Tb of data.

For our study, we focused on basically two kinds of data: (1) population characteristics (age, gender, marital status, etc.) and (2) clinical data including all drug prescriptions (dosage, route of administration of each drug, etc.) for all admissions.

2.1.1. Filters

We selected adult and adolescent patients (between 15 and 89 years old[2]), hospitalized for more than 3 days. We chose this minimal length of stay of 3 days in order to ensure enough time for manifestation and detection of complications during the stay. Other exclusion criteria for the patients were:

- patients hospitalized in surgery, because in surgery medical prescription and its complexity varies considerably according to preoperative, operative and postoperative phase as described in Lepelley et al. [5] and this information was not available in the dataset);
- out-patients and consultations;
- those with no drug prescription at admission; without which we cannot apply our analysis.

These filters retain 1 487 867 admissions also studied in [5]. We further filter out elective admissions, and finally retain a total of 1 271 733 eligible admissions.

2.1.2. Considered complications and ground truth

To build the complication prediction system, we need labeled data for training and evaluation purposes. We consider four complications:

- death during hospital stay;
- admission to ICU on or after the second day (excluding patients directly admitted to ICU on the first day[3]);
- pressure ulcers that were not present at admission time but developed during the stay;
- hospital–acquired infections developed during the stay.

Labeling a posteriori the occurrence of deaths and admissions to ICU is trivial as this information can directly be inferred from the medical records. Labeling the occurrence of hospital–acquired infections is slightly more involved since one must basically distinguish secondary infections occurring during hospital stay from infections existing before admission. For this purpose, medical experts guided us to label complications in terms of the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) codes [6] that are used in medical records, inspired from the work of Roosan et al. [7]. We implemented complication labeling as a one-pass algorithm that labels each admission with the complication(s) that occurred a posteori (if any). This served to establish a ground truth, which we use for training models.

2.1.3. Participants and occurrence of complications

Fig. 1 illustrates the distribution of eligible admissions by age and gender. The gap in the number of admissions between genders for people aged between 15 and 40 years is due to pregnancies. The fact that females tend to live longer explains the gap in the number of admissions of older people.

Among the overall population, there were 39 988 cases of hospital death (3.14%), 34 076 cases of pressure ulcers complications (2.68%), 45 542 cases of ICU admission on or after the second day (3.58%), and 32 198 cases of hospital–acquired infections (2.53%). On average, the probability that a patient experiences during his hospital stay at least one of the considered complications is 10.43%.

Fig. 2 shows the percentage of occurrence of each complication for each of the 417 hospitals considered. The proportion of complications appears roughly similar between hospitals except for a few

[2] We filtered out other ages because this information was biased in the dataset, i.e. age 89 denoting in fact age category 89+.

[3] 171 892 people were admitted to ICU on the first day: they have been excluded from the train and test population for the ICU label.

Please cite this article in press as: P. Genevès et al., Scalable Machine Learning for Predicting At-Risk Profiles Upon Hospital Admission, Big Data Res. (2018), https://doi.org/10.1016/j.bdr.2018.02.004