Future Generation Computer Systems 59 (2016) 105–113

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

A provenance model based on declarative specifications for intensive data analyses in hemotherapy information systems

Fernanda Nascimento Almeida [a,b], Gisela Tunes [c,*], Julio Cezar Brettas da Costa [c], Ester Cerdeira Sabino [d,e], Alfredo Mendrone-Júnior [e], João Eduardo Ferreira [a]

[a] Institute of Mathematics and Statistics, University of São Paulo, São Paulo, SP, Brazil
[b] Department of Computer Science, Federal University of Juiz de Fora, Juiz de Fora, MG, Brazil
[c] Department of Statistics, Institute of Mathematics and Statistics, São Paulo University, São Paulo, SP, Brazil
[d] Fundação Pró-Sangue/Hemocentro de São Paulo, São Paulo, SP, Brazil
[e] Department Infectious Disease, University of São Paulo, São Paulo, SP, Brazil

HIGHLIGHTS

- We study the connection between blood donation and decrease in Hct level.
- Our study is based on large unnormalized database produced by São Paulo Blood Center.
- Specialist knowledge was crucial for provenance description and data analysis.
- For repeated donors, young women are more likely to become LHct.
- Donors with Hct levels close to lower boundary, should take care in next donations.

ARTICLE INFO

Article history:
Received 5 March 2015
Received in revised form
21 August 2015
Accepted 16 September 2015
Available online 8 October 2015

Keywords:
Provenance
Inclusion criteria
Hemotherapy
Anemia

ABSTRACT

During the donation process, blood donors are screened for their hemoglobin or hematocrit level to protect them from developing anemia. Nevertheless, there is no standard procedure to predict anemia development after blood donation. The São Paulo Blood Center is responsible for maintaining a database with information on each donation. However, this database does not have good quality, and consequently, it is difficult to establish systematic analyses using the donation database without previously validating the data. To provide better quality donation data, this paper presents a provenance description based on classification criteria defined by specialists. More concretely, this paper answers the following main question: is there a connection between blood donation and a decrease in hematocrit levels?" This question was addressed to prevent undesirable outcomes to blood donors. In this paper, we show that it is possible to provide detailed investigations to answer this main question using the data description without the need to impose changes in the current database system structure sponsored by the São Paulo Blood Center.

1. Introduction

The great dissemination of tasks executed by *in silico trials* [1] increased the search for computing resources capable of handling the demands of the newly generated data. Since then, data management has become essential for efficient analyses, especially those resulting from heterogeneous sources or those that have variable quality. The increasing interest in heterogeneous database applications and the absence of standards have resulted in multiple representations of the same data resources.

In medicine (and particularly in hospital database systems), multiple representations of data can occur and cause serious problems. Another common source of errors is human mistakes, because these databases are filled in manually. Regular procedures such as data correction, normalization, and validation are not performed very frequently. As a consequence, the extracted information and results are not useful for experts, because they will not be able to properly assess the reliability of the generated results.

* Corresponding author.
E-mail address: tunes@ime.usp.br (G. Tunes).

http://dx.doi.org/10.1016/j.future.2015.09.019
0167-739X/

Generally, this problem occurs because the information and rules used to generate specific datasets are not preserved.

Provenance is an important concept for the data generation process [2–4]. A more common definition for provenance is the documentation of the history of the data [5–7]. This documentation contains each step of the data transformation process of the data source. The recording of the data details used during an *in silico* experiment is also known as provenance [1,8–10]. This provenance is essential for the verification and validation of an experiment [1,11] and is commonly used in scientific databases to aid in the identification and processing of images [4]. Provenance is sometimes based on scientific workflows such as VisTrails, Panda, Taverna, and Kepler among many others [12–15]. In these cases, provenance is used as a resource of traceability to store the execution log history of specific data [7,11,16,17].

In this sense, we propose that expert knowledge together with the outline of an *in silico* experiment can enable the extraction of more reliable information. Then, the provenance would be used to trace each event or decision made by the experts during the *in silico* experiment. An *in silico* trial can be defined as a scientific experiment whose performance requires intensive computing support [18]. We aim to assist the analysis process of a specific problem: the prevention of donation due to the risk of anemia development among regular blood donors. We intend to record the rule set adopted for each trial step. The data description model presented in this work could help researchers reproduce the same type of trial and achieve similar analyses. Finally, we validated and analyzed the results using descriptive statistical methods.

There are few papers concerning blood donor behavior, and most of the existing works have focused on the attitudes and reasons for donation in a selected donor group [19–21]. Qualitative information from blood center databases is difficult to obtain because the available data volumes are large and are not filtered or normalized continuously.

A candidate may be deferred from donation to protect the donor or the recipient. To avoid anemia due to blood donation, a donor is tested for hemoglobin levels (a protein that carries O2 inside the red cells) or for hematocrit (a measurement of the proportion of red cells in the blood). Both measurements can be used to identify individuals with anemia. Iron stocks are highly controlled by organisms. Typically, anemia is the last consequence of iron deficiency. Therefore, it is important to understand whether there is a connection between the donor and the hemoglobin (Hb) level decrease that can lead to anemia to help blood centers avoid undesirable results among blood donors.

Hematocrit (Hct) and Hb levels are both used to determine anemia from a blood sample obtained prior to donation. Hematocrit represents the proportion of the blood that consists of packed red cells. The hematocrit is expressed as a percentage by volume. Hemoglobin is a protein inside the red cells. By measuring either one we can infer whether the person has anemia. This is a well-established practice, and its primary objective is to protect the donor's health. Evaluating Hb levels is also important to avoid blood collection from donors who have a high risk of becoming anemic. Therefore, abnormally low values should be further investigated [20].

The *Fundação Pró-Sangue* (FPS) is the largest blood center in Latin America. It is responsible for the analysis and processing of approximately 100,000 blood units/year of whole blood. The FPS has a database designed to store donations, blood donation attempts and procedures related to donation (screening). This database controlling blood collection began in 1994, but it was not fully operational until 1996. In 2007, a data warehouse was established for the Reds II study [22]. This same software was used to extract and analyze the data collected from 1996 to 2006. During this period, 1,469,505 donors were selected who were responsible for 2,886,131 screenings at the FPS. These screenings contain records of whole blood donations, blood platelet donations, and deferrals.

Table 1
Useful data and reliability of preselected attributes.

Attribute	Useful data (%)	Reliability
Sex	99.99	High
Race	52	Low
Education	4.5	Low
Donation type	80	High
Visit day	99.9	High
Visit month	99.9	High
Visit year	99.9	High
Weight	92.5	Low
Height	–	–
Hematocrit	89.9	High
Hemoglobin	81.89	High

We believe that to properly analyze such a huge data volume it is important to consider the scientific knowledge of researchers regarding data meaning and usage. Only through this knowledge will it be possible to perform more reliable analyses and obtain better data quality. In this context, the a priori usage of clustering and classification methods is no longer feasible, because they are not capable of detecting relationships in low quality and incomplete data sets [23].

2. Provenance model based on declarative specifications

2.1. Data collection and processing

The data collected by FPS is entered manually in free form. Therefore, we decided to evaluate the quality, consistency and reliability of the stored data. Additionally, the database was not standardized and was vulnerable to inconsistencies, as was common for this type of medical data. To begin the development of the data description, a set of interesting characteristics was selected; all other characteristics with no direct relationship to this work were discarded. Next, we established actions to quantify data for each attribute to verify those that were more suitable for analysis.

First, we defined *Useful Data* for each attribute as the fraction of valid entries for that attribute (i.e., entries that were not null). For example, an entry with a sex attribute set as 'X' would be invalid, as would an entry with a null value.

Because the data are entered in free form and some fields are declared by the donor, we rank the attributes according to their subjectivity. Therefore, we created a measurement called *Reliability* to consider the way that data were collected. Attributes that were solely declared by the subject and not verified were considered highly subjective (displayed with the keyword '*Low*' in Table 1). This indicates that subjective attributes are more unreliable and should contribute less to the analysis. With these two measurements, we achieved the results presented in Table 1 for the preselected attributes.

Second, we implemented routines for treatment, cleaning, normalization, and data extraction. The provenance process summarized in Fig. 1 shows the set of rules (*inclusion criteria*) and processes used to obtain the study groups for this paper. The transformation processes are modeled as filters, and their routines and their data products serve as inputs and outputs for these tasks.

The first process (DS#1, in Fig. 1) was used to select records for the period studied (from 1996 to 2006). The second process (DS# 2, in Fig. 1) is the dataset cleaning step that is composed of a set of routines that have data cleaning purposes related to the variables of interest. As the name suggests, the attribute "donation type" contains records for the type of donation, which can be either