






# A Method for Collecting Provenance Data: A Case Study in a Brazilian Hemotherapy Center

Márcio José Sembay<sup>(✉)</sup> , Douglas Dyllon Jeronimo de Macedo ,  
and Moisés Lima Dutra 

Federal University of Santa Catarina, Florianópolis, Brazil  
marcio.sembay@posgrad.ufsc.br,  
{douglas.macedo,moises.dutra}@ufsc.br

**Abstract.** Data provenance is a process that aims to provide an overview of the origin of data used by information systems. It focuses on the origin of the data, especially on identifying the data sources and the transformations the data has undergone over time. This paper proposes a method for data collection based on the Provenance Model (PROV-DM), to be applied on Brazilian hemotherapy centers. Storing data on anemia indices using data provenance is the overall purpose of it. This work uses concepts of data provenance, knowledge provenance and scientific workflow techniques. It is an exploratory research, of practical and deductive nature, with application of a case study. Actual data was extracted from reports generated by a Brazilian hemotherapy center, provided from 2000 to 2018. People unsuitable for blood donation, who had favorable anemia rates to be rejected, were quantified and analyzed. A total of 197,551 blood donor candidates who attended the hemotherapy center in 19 years were analyzed. In the end, it was possible to quantify the unfit candidates with the highest index of anemia. A total of 1,011 male and 4,039 female candidates were accounted for, totaling 4.02% and 16.09% respectively of donors unfit for blood donations.

**Keywords:** Data Provenance · Anemia · Hemotherapy Center

## 1 Introduction

Information Science has its own scientific status as a social science. By its interdisciplinary nature, it presents interfaces with Mathematics, Logic, Linguistics, Psychology, Computer Science, Production Engineering, Graphic Arts, Communication, Librarianship, Administration, and other similar scientific fields [3]. Regarding the use of Data Provenance, both the Information Science and Computer Science use the structures of scientific workflows, which are abstractions related to the source of data, used as a support in the modeling of scientific experiments. Provenance is related to the audit, screening, lineage, and source of data. It can also be considered a metadata that describes the origin and all path taken to achieve the results of an experiment [10, 17].

This paper proposes a method for collecting Provenance data related to anemia indices. According to specialists of a particular hemotherapy center in Brazil (which due to privacy conditions, will be reported here as “X Hemotherapy Center”), anemia is a generic name for a series of conditions characterized by deficiency in hemoglobin concentration or in the production of red blood cells. Hemoglobin is a blood element with the function of carrying oxygen in the lungs to nourish all cells in the body.

A current study shows that 30% of the world’s population is anemic, especially children under 2 years old and women of different age groups, although it can also occur in men and the elderly. In addition, it is estimated that 27% to 50% of the population is affected by iron deficiency, especially in lower income and developing populations. In Brazil, the data may vary according to the study and the population group analyzed. But overall, it is estimated that 40% to 50% of children have anemia [25]. In this sense, it is important to emphasize that the data contained in the original database of the health institution under study in this paper do not establish systematic relationships between the stored variables for possible analysis of the statements generated by the specialists in the process of refusal of blood donors, regarding anemia rates.

There is no computational analysis of stored variables to uncover anemia index donations to chart possible future preventions, only expert-generated statements. These statements recorded by biomedical specialists do not always agree with database variables for possible broader analysis. Incorporating expertly defined statements regarding anemia rates, the possibility for obtaining a higher quality reduced dataset is evident. The analysis performed on the reduced dataset provided more reliable answers about a given biological phenomenon.

However, it was important to create a framework, which was able to facilitate the proposition of the method for data provenance activities and the storage of the expert statements, through an auxiliary database. Thus, the proposed method can ensure that statements made by experts during the process of blood donor refusal for anemia are reliable for new information flows and for the generation of new knowledge.

The proposed method is based on an adaptation of the Provenance Data Model (PROV-DM), which consists of a computational strategy capable of ensuring that expert-generated statements are passed on from the original X Hemotherapy Center database to an auxiliary database. Its main goal is to provide for a broader analysis and improve the quality of blood donations. In the end, the proposed method was able to manage the statements generated by specialists during the process of blood donor refusal for anemia indices, which were obtained from reports generated by the X Hemotherapy Center database from 2000 to 2018. The structuring of the method is based on data provenance stages, the use of scientific workflows and the needs found throughout the research for the treatment of digital data.

This research was developed in the scope of the Doctoral Program in Information Science from the Federal University of Santa Catarina, Brazil.

## 2 Literature Review

### 2.1 Information Science

Information Science is an area that is directly or indirectly linked to information technologies in the use of its methods of organization and representation of information in research development. Information technologies are key elements in the development of Information Science, as the creation of technological tools promotes the development of theories in order to achieve the goals set by this science in relation to the problems it is dedicated to solving [19].

The focus of Information Science implies both sociological and epistemological approaches, focused on phases such as: generation, collection, organization, interpretation, storage, retrieval, dissemination, transformation, and use of information. Information Science is an interdisciplinary science that assumes several disciplines of technological knowledge and try to contribute to the generation of new scientific knowledge [3, 7].

### 2.2 Anemia in Blood Donors

In 1999, members of the United Nations International Children's Emergency Fund (UNICEF), the United Nations University (UNU), the World Health Organization (WHO) and the Micronutrients Initiative (MI) showed that 3.5 billion people worldwide have iron deficiency anemia and that iron deficiency may be present in 80% of the world's population [23].

One of the most frequently observed factors in assessing the presence of anemia in blood donor candidates in hemotherapy centers throughout Brazil is hematocrit levels, the percentage of volume occupied by red blood cells and hemoglobin, which are the main items extracted from the statements in the original X Hemotherapy Center database, along with the fit and unsuitable candidates for blood donations and their respective screening. In Brazil, Ordinance RDC 153 of July 2004, enacted by the Ministry of Health, establishes the minimum acceptable hemoglobin and hematocrit values for a blood donation. These values are: 13 g/dl hemoglobin and 39% hematocrit for men and 12.5 g/dl hemoglobin and 38% hematocrit for women [13].

In the same Ordinance, the Ministry of Health also determined the minimum interval that must be respected between blood donations. This interval should be eight weeks for men and twelve weeks for women, respectively, because the shorter the interval, the greater the chance of developing anemia [13].

### 2.3 Scientific Workflows

Scientific experiments consist of observing a phenomenon through data analysis, and using the results obtained to prove or disprove a hypothesis. Due to the need to organize, process, control, and analyze the experiment, its representation is made through a cycle whose steps are composition, execution, and analysis. A scientific workflow is an abstraction of this process, which allows the formal specification of the steps to be performed in a scientific experiment [11].

An example of using scientific workflows would be to capture the steps taken to create a new drug, i.e. the source of the data that led to the creation of such a formula. In this sense, whenever this formula was improved, we would have the original data for reuse and replication of the experiment.

To benefit from provenance data, this data has to be captured, modeled, and stored for future reference. Information on the provenance of stored data can be managed by various Scientific Workflow Management Systems (SGWfCs) [15, 16]. Some SGWfC, such as Taverna [18], Kepler [2] and Pegasus [11] allow you to capture workflow steps during their execution. However, these systems often adopt proprietary models to capture the provenance generated in executions [8].

In this paper, a specific workflow was developed to demonstrate data capture using an auxiliary database without the need to change the original database.

## 2.4 Data Provenance

Data provenance is the complementary documentation of a given data that contains the description of “how”, “when”, “where”, and “why” it was obtained and “who” obtained it [5]. When buying a work of art, it is important to know its origin from its inception, including all former owners, i.e. this information will be essential to establish the value of this work of art. The same is true of data where data provision makes it possible to ensure data quality and accuracy [21].

In this sense, whenever provenance is automatically captured, it can be divided into levels [10]: a) workflow: involves the execution description of a process, i.e. the tasks that are part of it, is used by the vast majority of solutions with SGWfC and, in this case, must be adapted to capture the data from the different processes executed; b) activity: can occur in two ways. In the first, each executed process/program changes to capture the provenance data. In the second, specific programs can be created to monitor the execution of a given process and capture the provenance data; c) operating system: uses the data provided by the system, storing it in a specific database for provenance analysis.

By using data provenance it is possible to keep a complete record of how the calculation or processing was performed and it is essential to [6]: (a) ensure repeatability, (b) catalog the result, (c) avoid duplication of effort, and (d) retrieve data sources from output data. The main benefits of provenance for data quality are [4]: a) communicates data quality: reliability, suitability, accuracy, timeliness, redundancy; b) improves data interpretation as a function of source recognition; c) contributes to the justification of the use of a given data; d) reduces the possibility of errors in judging the accuracy of the data; e) allows non-data expert users to understand the processing steps; (f) identify the process used to conduct the creation of scientific data; g) allows updating of data from relational views; h) allows modification of relational view schemas; i) allows the use of historical data sources.

In this sense, the application of data provenance can be observed in the most varied areas, such as digital libraries, food industry, journalism, the traceability of information in social networks and the transparency of commercial applications, among others [9].

### Provenance of Knowledge

The term provenance of knowledge includes the source of the so-called meta-information, which is based on obtaining a description of the origin of part of the knowledge, including a description of the reasoning method used to generate it. However, data provenance and knowledge provenance have the same concerns and motivations, differing as to the purpose of the record that will be captured [20].

The provenance of knowledge provides two aspects: a) a personal and more abstract view of a document and its derivations, specifically for the experiment and the person, with the direct contribution of the scientist; and b) a more specific understanding of the data processing domain or its execution process, and may receive contributions from both the scientist and the note-taking curators [22, 23].

In this work, provenance of knowledge is related to the context of observing the statements on anemia rates described by the experts in the reports provided by the X Hemotherapy Center on donors who have become unfit for blood donation, determining the reliability of the researcher's reasoning about a given dataset. The provenance of knowledge was a term used to demonstrate and record the rules and reasoning used in the sample derivation processes from the reduced dataset, obtained at the X Hemotherapy Center X, in relation to the data relating anemia rates in unfit blood donors.

## 3 Related Works

In a PhD thesis written in 2012 at the University of São Paulo, the author proposes a model for describing data provenance for knowledge extraction in hemotherapy information systems based on the Open Provenance Model (OPM), designed to manage provenance records. Other similar applications can be found in the paper entitled "Laboratory and clinical genomic data sharing is crucial to improving genetic health care: the position statement of the American College of Medical Genetics and Genomics" [12]. In this research, the institution responsible for the study presents clinical level patterns by which statements about gene/disease associations and the clinical significance of variants were captured, by means of data provenance techniques, in statements made by experts in shared genomic data systems.

## 4 Proposal

This paper proposes a method for collecting provenance data related to anemia indices, by adapting some components taken from PROV-DM. PROV-DM's main function is to describe people, entities, and activities involved in the production of data. In addition, the PROV-DM model provides the conditions for provenance to be demonstrated and exchanged between different systems. For this purpose, a data provenance application was created. This application uses an auxiliary database to store provenance data related to anemia indices. It abstracts some attributes from the original database through researcher analysis. When searching the database, the provenance process assists in the process of tracking and signaling from the data source, as well as their movement between different data sources [21].

The proposed method sought to store the statements related to amounts of blood donation candidates with anemia rates considered unfit for blood donations, taken from reports provided by the X Hemotherapy Center system, in Brazil, from 2000 to 2018. The X Hemotherapy Center board has provided 19 years of reports from its blood donation registration system, containing various attributes and at least 80 reasons for refusing blood donors. All information provided has preserved the confidentiality of blood donors.

These reports were transformed into a CSV file by a software application created with the Java Eclipse IDE. Also, a local PostgreSQL database was created to serve the auxiliary database, i.e. the local repository of provenance. This local repository was populated with records taken from the original database provided by the X Hemotherapy Center X. Without changing the structure of the original database, It was possible, to observe the predominance of anemia indices found in blood donors at the X Hemotherapy Center X during 19 years.

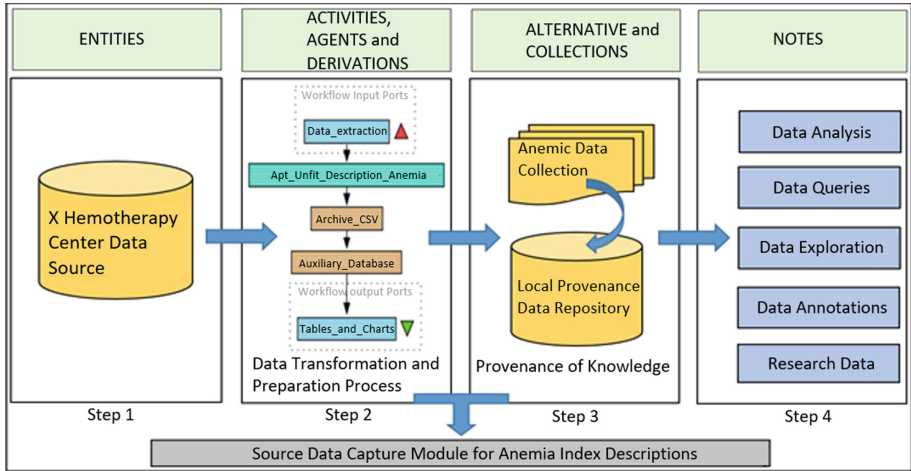
Finally, the workflows were generated with Taverna Workbench Core, and the Jasper-Reports library was used to generate a report that was, subsequently, transformed into a reduced dataset (see Table 1).

#### 4.1 Case Study

PROV-DM is divided into six components that contain both the elements and the possible relationships between them. They are [24]: a) Entities and Activities. Entities can represent any object (real or imaginary) and Activities represent the processes that use and generate Entities; b) Agent and Responsibilities: Agents are Entities that influence, directly or indirectly, the execution of the Activities, receive attributions from other Agents and may have some kind of connection (ownership, rights, etc.) over other Entities; c) Derivations: Describes the relationship between different Entities during the transformation cycle performed by the Activities, allowing to demonstrate the dependency between the used and generated Entities; d) Alternative: Describes the relationship between different views of the same Entity; e) Collections: These are Entities that have members, which are also Entities, and may have their provenance shown collectively; f) Annotations: Provides mechanisms for adding annotations to elements of the model.

Figure 1 presents the proposed method. Four steps are proposed in order to accomplish the task: i) Entities; ii) Activities, Agents, and Derivations; iii) Alternative Collections; and iv) Notes.

These steps comprise the adaptation of some PROV-DM components for provenance management, the creation of a specific workflow for data extraction, and the analysis of 19 years of anemia indices found in candidates who were considered unfit for blood donations.



**Fig. 1.** Proposed method

To better explain the proposed method presented in Fig. 1, the following subsections present each step with its respective components.

### Step 1 (Entities)

Data provided by the X Hemotherapy Center data source are collected and selected so that they can undergo the second step of this transformation process.

### Step 2 (Activities, Agents, and Derivations)

In this step, “ACTIVITIES, AGENTS, and DERIVATIONS” are represented by a workflow created specifically for the cycle of activities required to prepare the data collected before to be manipulated and studied, regarding the statements reported by the experts. Here the workflow aims to extract data that will be seen as reliable to perform analysis for blood donor improper anemia indices (rejections) and, soon after, the extracted data is transformed into a CSV file for the creation of an auxiliary database to generate tables and graphs (data preparation and transformation process).

So, the “ACTIVITIES” are represented by the dataset collected from the X Hemotherapy Center over 19 years, containing dates, times, consumption, processes, transformation, modification, relocation, and use of the original data in relation to anemia rates. The activities are performed by the agents. The “AGENTS” are represented by both blood donor candidates and health specialists who appear in X Hemotherapy Center reports. The “DERIVATIONS” represents the transformation of the X Hemotherapy Center, after the application of the method proposed here, in an entity that will serve as a reference model for the application of this method in other Brazilian hemotherapy centers with the same structure.

The second step starts with the workflow created specifically for the process performed at the X Hemotherapy Center, which is the process of collecting provenance data

related to anemia rates, so that it can be evaluated and analyzed regarding their origin, thus generating new knowledge.

### **Step 3 (Alternative and Collections)**

“ALTERNATIVE” represents the view of the X Hemotherapy Center in the declarations of anemia indices of candidates for blood donation considered inept. By declaring the reports provided, it was possible to quantify in order to generate the results of the analysis of the data declared by the experts. “COLLECTIONS” represents the collection of anemic candidate data to be inserted into the local source data repository. In here, one can also apply provenance of knowledge concepts for possible audits of source data. It is important to notice the “COLLECTIONS” component stores data that constitute documents, in which each document has its own provenance, but the file itself also has its origin information: who kept it, which documents contained it at what time, how it was assembled, etc. Therefore, in addition to the procedures for collecting the provenance data, i.e. the provenance of knowledge applied to this method, it was possible to provide an overview of the provenance of the data described in the reports on the anemia indices of candidates for unfit donations. This made possible a better understanding of the risks and the reasons for the rejection of the blood donations.

### **Step 4 (Notes)**

After storing the provenance data in a local repository, in the fourth step, “NOTES” represents the relationship of the important points on anemia indices, such as: analysis, consultations, exploration, annotations, and reuse of data for further research. In here, it is possible to generate reports that can be cross-referenced with other data, as needed by health specialists. Consequently, this step creates an interaction between refined data and expert reporting.

## **5 Results**

The reports provided by the X Hemotherapy Center are annual from 2000 to 2018, dated January 1 to December 31 of each year to better simplify and reduce the presentation of the data during these 19 years. The body of each report contained a series of attributes, from which only the data that had the potential to generate the expected results was selected.

The selected attributes were: i) number of male and female fit donors as well as the number of male and female unfit donors, all aged from 16 to 60 years old or older, including first-time, repeat or sporadic donors; ii) anemia indices of male and female unfit donors, i.e. low hematocrit and low hemoglobin, considered the reasons for refusal according to statements made by experts in the submitted reports, which in fact built the set of information necessary to generate the process of data provenance and provenance of knowledge; and finally iii) the screening performed by the experts in each year surveyed, i.e. the discovery of diseases through blood donation at the X Hemotherapy Center. These attributes can be better observed in Fig. 2 below.



	ano [PK] integer	indanemfem integer	indanemmasc integer	quantaptofem integer	quantaptomasc integer	quantinapfem integer	quantinapmasc integer	trimed integer
1	2000	647	237	1608	4418	1612	2307	71
2	2001	876	278	1674	4316	1844	2409	0
3	2002	636	203	1974	4313	1852	2328	0
4	2003	269	65	2006	4332	1267	1767	0
5	2004	188	46	1928	3970	1447	1962	0
6	2005	99	16	2209	4285	1265	1484	0
7	2006	102	13	2359	4667	1505	1383	0
8	2007	123	13	2716	4722	1349	1152	0
9	2008	146	28	3021	4899	1263	1091	0
10	2009	167	25	2905	4919	1401	1142	0
11	2010	158	32	3116	4886	1336	1091	46
12	2011	59	8	3672	5362	1318	991	114
13	2012	109	3	3886	5176	1232	861	19
14	2013	102	9	4019	5165	1124	813	74
15	2014	2	0	4506	5516	1194	929	17
16	2015	10	1	4924	5783	1241	1047	12
17	2016	3	2	3943	4527	1011	842	1
18	2017	137	16	3475	4021	928	723	19
19	2018	206	16	3692	4380	916	834	381
*								

**Fig. 2.** Auxiliary database with selected attributes.

The auxiliary database presented in Fig. 2 demonstrates the selected attributes taken from 19 years of reports provided by the X Hemotherapy Center. They are a massive set of information that was possible to be retrieve and group into a reduced dataset in order to be analyzed.

The attributes selected are the following (in Brazilian Portuguese acronyms): a) ano (donation reference year); b) indanemfem (female anemia index); c) indanemmasc (male anemia index); d) quantaptomasc (amount of male fit donors); e) quantaptofem (amount of female fit donors); f) quantinapfem (amount of female unfit donors); g) quantinapmasc (amount of male unfit donors); and finally h) trimed (screening by the attending physicians).

Table 1, shown below, is populated with data extracted from the outcome of the analysis undertaken on the auxiliary database. This database contains statements about the anemia indices of unfit donors. All the attributes shown in Table 1 helped perform the development of the specific workflow to create the provenance data method for the X Hemotherapy Center.

Table 1 shows the number of eligible and unsuitable blood donation candidates, both male and female, as well as the percentage of anemic unsuitable blood donation candidates. Importantly, candidates unfit for blood donations were rejected for anemia rates, i.e. were also filtered from at least 80 reasons for refusal before becoming eligible for blood donations.

These reasons for refusal are diverse, ranging from something simple as a fever, weight loss, flu manifestations, among others, to more complex reasons such as diabetes, heart disease, cancer, HIV, etc. In this work, we gathered the unfit donors who had several reasons for refusal. From this subset, we extracted those unfit by anemia, by presenting that percentage for each year. Table 1 also shows the amount of screenings performed by

doctors each year, which in other words means the number of blood donation candidates who discovered disease during the donation process.

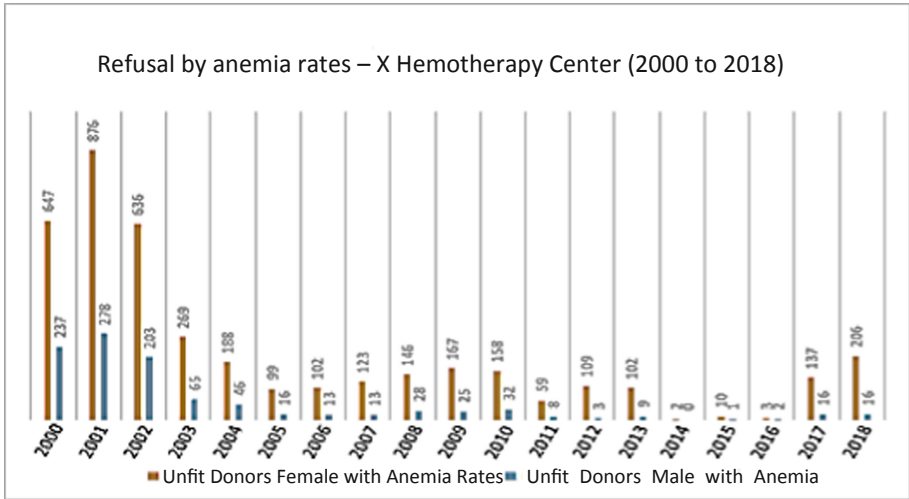
The 19 years of data analyzed revealed 197,551 blood donation candidates. Out of this total, 114,813 were male and after blood tests 89,657 were found fit and 25,156 were found unfit for blood donations. Anemia rates were present in 1011 candidates, totaling 4.02% of candidates unfit for blood donations. Of the remaining 82,738 female blood donation candidates, 57,633 became fit and 25,105 became unfit for blood donations after blood tests. Besides, there were 4,039 anemia inducing female candidates, totaling 16.09% of donors unfit for blood donations.

Table 1 also shows that from 2001 to 2009 there was no screening, but anemia rates continued to fluctuate between male and female donors, with a female predominance. In 2018, the highest number of tests was observed, evidence for some unidentified specific reason in relation to blood donations that year. The year 2001 is the year in which the anemia rates between men (11.54%) and women (47.51%) represent the highest rates observed. If compared to 2018, when there was the highest screening rate, the anemia rate was well below the 2001 average.

**Table 1.** Number of donors (fit, unfit, anemic unfit, percentage of anemic unfit and screening).

Years	Male fit	Male unfit	Anemia unfit male	% Anemics unfit male	Female fit	Female unfit	Anemia unfit female	% Anemics unfit female	Screenings
2000	4418	2307	237	10,27%	1608	1612	647	40,14%	71
2001	4316	2409	278	11,54%	1674	1844	876	47,51%	0
2002	4313	2328	203	8,72%	1974	1852	636	34,34%	0
2003	4332	1767	65	3,68%	2006	1267	269	21,23%	0
2004	3970	1962	46	2,34%	1928	1447	188	12,99%	0
2005	4285	1484	16	1,08%	2209	1265	99	7,83%	0
2006	4667	1383	13	0,94%	2359	1505	102	6,78%	0
2007	4722	1152	13	1,13%	2716	1349	123	9,12%	0
2008	4899	1091	28	2,57%	3021	1263	146	11,56%	0
2009	4919	1142	25	2,19%	2905	1401	167	11,92%	0
2010	4886	1091	32	2,93%	3116	1336	158	11,83%	46
2011	5362	991	8	0,81%	3672	1318	59	4,48%	114
2012	5176	861	3	0,35%	3886	1232	109	8,85%	19
2013	5165	813	9	1,11%	4019	1124	102	9,07%	74
2014	5516	929	0	0,00%	4506	1194	2	0,17%	17
2015	5783	1047	1	0,10%	4924	1241	10	0,81%	12
2016	4527	842	2	0,24%	3943	1011	3	0,30%	1
2017	4021	723	16	2,21%	3475	928	137	14,76%	19
2018	4380	834	16	1,92%	3692	916	206	22,49%	381
<b>Total</b>	<b>89.657</b>	<b>25.156</b>	<b>1.011</b>	<b>4,02%</b>	<b>57.633</b>	<b>25.105</b>	<b>4.039</b>	<b>16,09%</b>	<b>754</b>

These comparisons help draw estimates and thresholds for new studies in the area of hemotherapy from a data provenance perspective, which could ultimately contribute to the prevention of anemia rates in the X Hemotherapy Center. In order to provide a more detailed comparison, Fig. 3 shows a graph of the profile of anemic unsuitable donor candidates during the 19 years analyzed.



**Fig. 3.** Refusals due to anemia (male and female unfit)

It can be observed that men's anemia rates are lower than women's. This can be demonstrated as a result of the interval between blood donations from women and men. Actually, the X Hemotherapy Center has reported that women tend to be at risk of developing anemia earlier than men. This happens due to more frequent blood donations done by women in a shorter period of time, or even repeated donations.

Other factors that are associated with higher rates of anemia in women are pregnancy and the monthly blood loss from menstruation. For men, the possibilities are that continued blood loss caused by some type of bleeding, associated with blood donation, or regular blood donation may be related to the risk of developing anemia [1]. According to the literature, the recommendations to reduce anemia rates in all male and female population are related to the iron supplementation after blood donation and the increasing of the waiting time between donations [13, 14].

## 6 Conclusions

This paper proposed a method for data collection in a Brazilian hemotherapy center, based on the data provenance approach. This method proved to be important for generating a reduced dataset, in order to allow knowledge extraction and mining of a volume of data generated over a 19-year period.

The application of the concept of data provenance along with the provenance of knowledge and the scientific-workflow techniques for the development of the proposed method led to the conclusion that these elements together may, in fact, contribute to the advancement of research in Brazilian hemotherapy centers. They helped guide data collection in relation to the anemia index statements found in the 19 years of data presented in the reports provided by the X Hemotherapy Center. Moreover, they contributed to generate knowledge through the analysis performed in the anemia rates found. It could be observed that most of the population studied was female, who develops more frequently anemia rates in conjunction with other factors after blood donations, which may be a research factor for the discovery of other diseases. The method proposed here can be applied to another Brazilian hemotherapy center that has the same features and security policies presented here.

The main contributions of the proposed method are: a) the improvement of the analysis that discovers anemia indices that make blood donation unfeasible; b) the provision of a beneficial view of donor groups in their hemotherapy centers; c) the raising of the quality of blood donations by creating preventive mechanisms for blood donor evasion; d) the permission, when necessary, to query the local source data repository, in order to create data-quality metrics and to perform an audit processes for this data. These contributions demonstrate that the use of data provenance together with the provenance of knowledge are differential requirements of what can be found in the literature and, in fact, contribute to the relevance of the results found in this research.

After several searches in the literature and considering the few works found, it became clear to us the relevance of the studied subject, since we could not find similar proposals. We believe that methods for data collection that are able to highlight, synthesize and explain elements in the area of Hemotherapy may be a guide for the development of new research in this area.

It also became clear that Computer Science approaches, such as the Provenance of Data, combined with an Information Science viewpoint could be very useful to the context of hemotherapy information systems. Computer Science provides the technological support for the development of data provenance. On the other hand, Information Science provides the methods and techniques for informational treatment, making use of technological applications to apply the provenance of knowledge.

## 7 Future Perspectives

Some paths can be envisaged in the follow up of this research: a) Data collection done directly on the hemotherapy center's database with the help of an already-connected auxiliary database (a local provenance data repository composed of anemia declarations). This would be done without modifying the original database structure, i.e. by generating an automatic CSV file (or others) in a cloud computing structure capable of handling large amounts of data, providing better data quality for future research; b) Automating the data description process, i.e. generating predictions for more complex analysis by using data provenance as a preventive factor for anemia indices that result in blood-donation refusals; c) Improving the data provenance method by performing data-cross-referencing processes (reasons for refusal of blood donations in all hemotherapy centers in Brazil);

and d) Integrating more hemotherapy centers in this research, or even performing the study in other regions of Brazil, by adapting the proposed method whenever is necessary.

Another future perspective means using data provenance for preventing anemia problems, by indicating the reasons why they occur more frequently in certain regions of Brazil. We can also envisage the evaluation of how data provenance models could assist in generating more complex and complete analysis, in order to discover the most prominent anemia rates, by Brazilian region.

For all of the aforementioned scenarios to become true, we believe that Brazil should improve its data storage infrastructure and computational tools applied to the Brazilian hemotherapy centers. Furthermore, it would be advisable to think about the creation of research institutes all over the country, in order to study and to prevent anemia rates and other reasons for refusing blood donations. That indeed would be a challenge.

## References

1. Almeida, F.N.: Descrição da Proveniência de Dados para Extração de Conhecimento em Sistemas de Informação de Hemoterapia, p. 114 (2012). f. Tese (Doutorado) - Curso de Bioinformática, Bioinformática, Universidade de São Paulo - USP, São Paulo (2012)
2. Altintas, I., Berkley, C., Jaeger, E., Jones, M.: Kepler: an extensible system for design and execution of scientific workflows. In: Proceedings of 16th International Conference on Scientific and Statistical Database Management, Santorini Island, Greece, 23 June 2004, pp. 423–424. IEEE (2004)
3. Borko, H.: Information science: what is it? *Am. Doc.* **19**(1), 3–5 (1968)
4. Bose, R., Frew, J.: Lineage retrieval for scientific data processing: a survey. *ACM Comput. Surv.* **37**(1), 1–28 (2005)
5. Buneman, P., Khanna, S., Wang-Chiew, T.: Why and where: a characterization of data provenance. In: Van den Bussche, J., Vianu, V. (eds.) *ICDT 2001*. LNCS, vol. 1973, pp. 316–330. Springer, Heidelberg (2001). [https://doi.org/10.1007/3-540-44503-X\\_20](https://doi.org/10.1007/3-540-44503-X_20)
6. Buneman, P., Tan, W.C.: Provenance in databases: tutorial outline. In: Proceedings of ACM SIGMOD International Conference on Management of Data, Beijing, China, 11–14 Jun 2007. ACM (2007)
7. Capurro, R., Hjørland, B.: O conceito de informação. *Perspectivas em Ciência da Informação*, Belo Horizonte **12**(1), 148–207 (2007)
8. Cuevas-Vicenttin, V., Dey, S., Wang, M.L.Y., Song, T., Ludäscher, B.: Modeling and querying scientific workflow provenance in the D-OPM. In: Proceedings of 2012 SC Companion High Performance Computing, Networking, Storage and Analysis, Washington, EUA, 10–16 November 2012, pp. 119–128. IEEE (2012)
9. Curbera, F., Doganata, Y., Martens, A., Mukhi, N.K., Slominski, A.: Business provenance – a technology to increase traceability of end-to-end operations. In: Meersman, R., Tari, Z. (eds.) *OTM 2008*. LNCS, vol. 5331, pp. 100–119. Springer, Heidelberg (2008). [https://doi.org/10.1007/978-3-540-88871-0\\_10](https://doi.org/10.1007/978-3-540-88871-0_10)
10. Davidson, S.B., Freire, J.: Provenance and scientific workflows: challenges and opportunities. In: ACM SIGMOD International Conference on Management of Data, pp. 1345–1350 (2008)
11. Deelman, E., Gannon, D., Shields, M., Taylor, I.: Workflows and e-science: an overview of workflow system features and capabilities. *Future Gen. Comput. Syst.* **25**(5), 528–540 (2009)
12. Genetics in Medicine: ACMG. <https://www.nature.com/articles/gim2016196>. Accessed 22 Sept 2019

13. Mendrone, A.J.R., et al.: Anemia screening in potential female blood donors: comparison of two different quantitative methods. *Transfusion* **49**, 662–668 (2009)
14. Meyers, D.G.: The iron hypothesis: does iron play a role in atherosclerosis? *Transfusion* **40**(8), 1023–1029 (2000)
15. Moreau, L., et al.: The open provenance model core specification (v1.1). *Future Gen. Comput. Syst.* **27**(6), 743–756 (2011)
16. Moreau, L., Freire, J., Futrelle, J., McGrath, R.E., Myers, J., Paulson, P.: The open provenance model: an overview. In: Freire, J., Koop, D., Moreau, L. (eds.) *IPAW 2008. LNCS*, vol. 5272, pp. 323–326. Springer, Heidelberg (2008). [https://doi.org/10.1007/978-3-540-89965-5\\_31](https://doi.org/10.1007/978-3-540-89965-5_31)
17. Moreau, L., Groth, P.: Provenance: An Introduction to PROV. *Synthesis Lectures on the Semantic Web: Theory and Technology*, vol. 3, no. 4, pp. 1–129. Morgan & Claypool Publishers, California (2013)
18. Oinn, T., Li, P., Kell, D., Goble, C.: Taverna/<sup>my</sup>Grid: aligning a workflow system with the life sciences community. In: Taylor, I.J., Deelman, E., Gannon, D.B., Shields, M. (eds.) *Workflows for e-Science*, pp. 300–319. Springer, London (2007). [https://doi.org/10.1007/978-1-84628-757-2\\_19](https://doi.org/10.1007/978-1-84628-757-2_19)
19. Saracevic, T.: Ciência da Informação: origem, evolução e relações. *Perspectivas em Ciência da Informação* **1**(1), 41–62 (1996)
20. Silva, P.P., Mcguinness, D.L., Mccool, R.: Knowledge provenance infrastructure. *Proc. IEEE Data Eng. Bull.* **25**, 179–227 (2003)
21. Simmhan, Y.L., Plale, B., Gannon, D.: A survey of data provenance techniques. Technical report TR-618, Computer Science Department, Indiana University (2005)
22. Stevens, R., Zhao, J., Goble, C.: Using provenance to manage knowledge of in silico experiments. *Brief. Bioinform.* **8**, 183–194 (2007)
23. Stolfus, R.J.: Defining iron deficiency anemian public health terms: a time for reflection. *J. Nutr.* **131**, 565S–567S (2001)
24. W3C: PROV-DM. <http://www.w3.org/TR/prov-dm/>. Accessed 21 Sept 2019
25. WHO. <https://www.who.int/topics/anaemia/en/>. Accessed 21 Sept 2019