Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 210 (2022) 224–229

Procedia Computer Science

www.elsevier.com/locate/procedia

The 3rd International Workshop on Healthcare Open Data, Intelligence and Interoperability (HODII)
October 26-28, 2022, Leuven, Belgium

Association Models for Relating Problems with Semiologic Data in Intensive Medicine

Inês Tavares[a], Júlio Duarte[a]*, Hugo Peixoto[a], Álvaro Silva[b], Maria Manuel[b], César Quintas[b]

[a] *Algoritmi/LASI research center, University of Minho, Portugal*
[b] *Centro Hospitalar Universitário do Porto, Portugal*

Abstract

In Intensive Medicine, the large amount of data that medical professionals are subject to can be overwhelming, leading to the use of techniques and treatments that may not be the most effective in treating patients. Should there be a need to cross planning registries made by doctors and nurses with patients' problems, the situation becomes unmagenable. To support health professionals' decision-making process, and consequently allow health professionals to make informed and timely decisions, by promoting proactive actions, the current study approaches the establishment of a correlation between medical problems and medication and therapies, using association rule mining algorithms, so that physicians can have the correct and timely information regarding patients and consequently, the most appropriate treatments for them in every situation.
The main objective is for doctors and nurses to be able to look through problems and have them associated with the most frequently used and reliable therapies and medication, in order to assist patients with the highest healthcare quality.
The results of this work corroborate that in order to improve the care provided to Intensive Care Units patients, it is essential to implement intelligent systems that can support hospital staff and assist to provide healthcare more efficiently.

Keywords: Data Mining;Intensive Medicine; Decision Support Systems; Association Rules; Association Algorithms.

* Corresponding author.
E-mail address: jduarte@di.uminho.pt

1877-0509
10.1016/j.procs.2022.10.141

1. Introduction

Intensive Medicine is a medical field that specializes in giving treatment to patients who are acutely unwell and require critical medical care. Intensive Care Units (ICU) are separate, self-contained areas that provide the critical care and life supports for these patients. [1]

Collecting data in this continuously evolving environment is very challenging. Firstly, the task of collecting several types of data such as demographic, historical, ICU stay, laboratory and medication is quite work-intensive. Secondly, the integration of all these categories is extremely complex because, in addition to containing heterogenous data, are collected from different data sources, which arises standardization problems, confidentiality, among others. [2]

The role of Data Mining models and algorithms has been crucial in the adoption of measures and optimization of processes in order to improve medical decision making and the healthcare service provided to patients.

This project has been developed in a health institution through a partnership with Centro Hospitalar Universitário do Porto (CHUP). This article aims to demonstrate that by using Data Mining techniques, specifically association rule algorithms, it is possible to generate patterns and establish associations between semiologic data (planning registries) and patient's problems.

The document is structured in the following way: The first section is the introduction where the main ideas of this work are presented, the second is the background where the problem is defined as well as the theory behind the work, the third section is the description of the study methods, where the tools used are described. In the fourth chapter of this paper is the discussion where some views on the results are presented. Finally, the last section presents some conclusions and basic ideas about the work to be done in the future.

2. Background

2.1. Intensive Medicine and Intensive Care Units

Intensive Medicine (IM) is a multidisciplinary field of medical sciences, dedicated to the treatment and monitoring of patients whose clinical condition is alarming and, consequently, requires constant supervision.

The IM specialty is practiced in Intensive Care Units (ICU) and is recognized as being one of the areas of medicine with the highest degree of risk and severity, requiring increased attention with regard to technological innovation and the support given to health professionals in decision making. [3][4]

2.2. POMR and SOAP

POMR (Problem Oriented Medical Records) is a medical records model that represents the events observed by health professionals during patient's ICU stay. The creation of this model was based on the need to provide to medical teams' easy access to information about each patient, since on a daily basis, doctors are confronted with countless patients with distinct health problems from the most varied scopes. That said, it is not humanly possible for a health professional to retain information about all of their patients, allowing them to make decisions quickly, so a model like the POMR is an ally for improving their performance.[5]

The POMR system organizes the clinical information registered by the physician, dividing the record into four sections: database, problem list, initial plan, and progress notes. Each record follows a structure namely SOAP (Subjective, Objective, Assessment and Plan) [6].

2.3. Association Rules

Association rules are a paradigm for the representation of knowledge, due to their simple and intuitive structure, which makes them easily understandable and similar to the typical logic of human reasoning.[7] This technique is mainly used to discover interesting relationships between variables in large datasets.

An association rule is composed of two parts: an antecedent (if), which is something found in the data; and a consequent (then), which is something found in combination with the antecedent.

Despite this, an important aspect to take into account is that not all patterns identified or relationships discovered by the association rules are of interest to be studied. In order to clarify this point, two metrics should be considered:

• Support: Indicates how often the "if/then" relationship appears in a given database.

• Confidence: Indicates the number of times that the relationships identified by the support metric were found to be true.

That said, in this context, the association rules allow to understand that certain combinations of conditions may indicate that there is an increased risk of other complications, being important to associate certain therapeutic treatments with infections or specific health problems.

2.4. Algorithms for Generating Association Rules

The methods used to derive association rules arose from the need to analyze extremely long transaction lists, in order to identify frequent and recurring patterns in the data. From these models, knowledge that is easily interpretable and useful for business managers, who can translate it into concrete action plans, can be extracted. [7]

There are several algorithms available to discover association rules, whose role is to identify frequent patterns in the database. [8] In the context of this study were utilized the ECLAT Algorithm and the FP-Growth Algorithm.

3. Description of the Study

3.1. Methods and Tools

The methods used throughout the study are Design Science Research (DSR) and the Cross Industrial Standard Process for Data Mining (CRISP-DM), as research and Data Mining methods, respectively.

3.2. Business Understanding

This project aims to provide to Intensive Medicine, a system that allows health professionals to make more informed decisions, through the organization and processing of the data received by the various monitors and laboratory analyses, as well as the SOAP records recorded about the clinical status of each patient and present the various possible diagnoses so that decision-making is effectively more efficient.

3.3. Data Understanding

The data used came from the CHUP UCI database, during the period of March 2020 and October 2021, organized into two datasets. All the datasets used ended up having many variables that were not considered relevant for the context of this study, as it was necessary to filter them.

The first dataset provided is related to SOAP records, which contains the columns listed in the table below, and encompasses the subjective, objective, evaluation and plan data corresponding to a given patient. During the period of March 2020 and October 2021, data from 47 patients were recorded.

Table 1. Variables of SOAP dataset

Variable	Definition
SEQ_NUMBER	Admission sequential number
PROC_NUMBER	Patient process number
DAY_DATE	SOAP record day
EPISODE	Patient admission number
RPI	Notes regarding therapeutic treatment

The second dataset provides all information related to the administration of medication. All medications administered to patients for each day of hospitalization were recorded, in a total of 36 patients in the period between March 2020 and October 2021.

Table 2. Variables of Medication Dataset

Variable	Definition
FMP_DATA	Medication administration day
SEQ_NUMBER	Admission sequential number
MED_DESIGNATION	Name of the medication administered
ART_DESIGNATION	Detailed designation of the medication administered
FMP_QUANTITY	Dose of medication administered
FMP_PRICE	Price of the dose of medication administered

3.4. Data Preparation

In order to be able to match the different dataset, and thus ensure that the information was used in its entirety, a large amount of the data was lost, due to null values, incorrect readings performed by electronic equipment or data heterogeneity during the extraction from the database. Variables that had negative values or were outside the normal parameters of these same variables were found, resulting in additional lost data.

Since the planning data is recorded manually by health professionals, it is, consequently, target of spelling errors or the use of different terms to refer to the same subject. That said, it was necessary to resort to the NLP language to process them. The process involved manipulating the Portuguese language, starting with cleaning and correcting characters and expressions considered irrelevant, as well as removing white spaces and fix formatting. The final goal was to count the frequency of words and establish correlations between expressions, in order to generate patterns and trace trends.

Subsequently, the medications were grouped by patient and by date and the document related to the problems was added. Finally, the new files were loaded together and concatenated.

Prior to modeling, it was necessary to format the data so that they could be read and interpreted by the association algorithms.

3.5. Modeling

The modeling process consisted of building models, adapted to each algorithm, using the Python language. First, the Top 3 Problems with the highest number of medication entries and planning records were filtered, and then each problem was executed individually. The Top 3 Problems are COVID Pneumonia, Kidney Failure Type One and Cardiovascular Dysfunction. Due to the considerable heterogeneity of planning data and the fact that the registration is a manual process, it was only possible to generate the support parameter. In the table below are featured some examples of the data retrieved.

Table 3. Association of Problems and Planning Records with Support

Problem	Planning Record	Support
COVID Pneumonia	“ajusta ventilação”	0.1061947
Kidney Failure Type I	“sem ercorrencias”	0.1142857
Cardiovascular Dysfunction	“prone”	0.1333333

Contrary to planning records, medication proved to be a much more attractive variable for generating association rules, as it was possible to calculate parameters such as support, trust, lift and leverage. In the table below are some examples of the obtained results.

Table 4. Association of Problems and Medication with Support, Confidence, Lift and Leverage

Problem	Antecedents	Consequents	Antecedent Support	Consequent Support	Support	Confidence	Lift	Leverage
COVID Pneumonia	frozenset({'fresubin original) sonda 500/1000mL'})	frozenset({'decnp - dieta entérica completa normalizada polimérica#decnp (nutrison std'})	0.1061947	0,39604	0,39604	0,39604	1	2,525
Kidney Failure Type I	frozenset({'isousource std'})	frozenset({'fresubin original) sonda 500/1000mL'})	0.1142857	0,333333	0,333333	0,333333	1	3
Cardiovascular Dysfunction	frozenset({'paracetamol', 'isousource std'})	frozenset({'pantoprazol'})	0.1333333	0,103448	0,534483	0,103448	1	1,870968

4. Evaluation and Discussion

The ECLAT and FP-Growth algorithms proved to be very efficient processes and suitable for large datasets.

Regarding data related to therapeutic treatments and treatment planning observations, an extraordinary heterogeneity of results was observed. This is due to the fact that it is a daily record, that can be repeated several times on the same day, and that it is a manual writing record. As expected, even with the use of NLP techniques, it is quite limiting to try to establish trends and recording patterns, not only because the speech differs from person to person, but also because the comments about the patient's clinical status are extremely specific and adapted to the conditions and other health problems they may have. As a consequence, it was impossible to generate association rules for therapeutic treatment records, even with a minimum confidence of 10%.

In the context of Data Mining, it is considered that the achievement of results could be significantly improved if the recording of this type of data were carried out in a more automatic and standardized way. A section could be added with pre-defined options that identify the therapeutic procedures and treatments to be carried out, so that the generation of patterns and identification of trends could be done.

As far as the business is concerned, there should be a uniformity in the language used to make comments and recommend therapies, so that results could have greater substance.

Regarding medicines, it was observed that the information provided is much more organized and formatted, less redundant and capable of generating very interesting association rules. As a result of the high number of rules obtained for each of the problems identified in the top three, it was inserted in this document one example rule for each problem. As mentioned earlier, due to the high amount of data, the support associated with each rule is relatively low, between 50% and 12%, depending on the issue at hand. On the other hand, there is a consistently high degree of confidence, which reveals the existence of a high probability of coexistence between the antecedent and the consequent, but which has to be analyzed together with the support, since rules with low support can also generate rules with high confidence. The confidence of a rule varies between 0 and 1, giving the estimate between the probability of observing the consequent, given the antecedent. That said, it is necessary to establish a relationship with the parameters of lift and leverage, which indicate the degree of relevance of the rule.

According to the literature, we know that for the lift, the values can vary between $[0,+\infty[$, and that values close to 1 show that the antecedent and consequent are independent, making the rule uninteresting. If the value is greater than 1 and the greater the distance, it indicates that the evidence of the antecedent provides more information about the consequent, that is, there is a high level of co-occurrence.

For leverage, the values can vary between [-0.25,0.25], to measure how much more count is obtained from the co-occurrence of the antecedent and consequent, of the expected, that is, of independence.

After analyzing the data in the table above, it is concluded that:

- COVID Pneumonia: The association rules for this problem have approximately 40% support, confidence value of 1, 2.5 lift, and 0.23 leverage. Based on the values of each metric and the relationships between each one of them, it is verified that the association rules of this problem allow establish a strong relationship between the use of the drugs shown in the table and the resolution of the health problem Pneumonia by COVID-19.
- Kidney Failure Type I: The association rules taken for this problem have very similar results to the health problem explained above, with the conclusion drawn about the medication used for its treatment being the same.
- Cardiovascular Dysfunction: The association rules obtained for Cardiovascular Dysfunction require a closer analysis, due to the dispersion of values that occur in certain relationships. Support values range between 10% and 38%, confidence remains at the value 1 and the leverage parameter ranges from 0.04 to 0.24. It is therefore necessary to distinguish that a rule that has a low level of support and that, despite having high confidence levels, has a low leverage value, may not have as much relevance as a rule that has 38% support and degrees equally high reliability, lift and leverage.

5. Conclusions and Future Work

Based on the results obtained, it has been collected information that could make the choice of medicines more efficient in the treatment of patients, since associations with more precarious values were discovered with regard to the frequency with which they are used and the importance of its association with other medications.

In the future, it is crucial to standardize and automate the therapeutic planning recording process, so that patterns that allow health professionals to make a faster and more efficient decision on therapies can be found, as well as to extinguish practices that do not have relevance and make the entire patient information system more organized.

Regarding drug administration, the team noted that the registration of diets and drugs is carried out in the same place, not allowing a Data Mining analysis to establish associations and find trends exclusively for the medications used. There should be a greater effort and rigidity on the part of health professionals in the registration of procedures and medications administered to patients, so that projects like this can add useful and correct information that not only adds value to the care provided in the Units of Intensive Care, but also enabling doctors and nurses to do their jobs more effectively.

Acknowledgments

The work has been supported by FCT – Fundação para a Ciência e Tecnologia within the Project Scope: DSAIPA/DS/0084/2018.

References

1. Department of Health G of WA Intensive care units (ICUs). https://www.healthywa.wa.gov.au/Articles/F_I/Intensive-care-units-ICUs
2. Ramon J, Fierens D, Güiza F, Meyfroidt G, Blockeel H, Bruynooghe M, Van Den Berghe G (2007) Mining data from intensive care patients. Adv Eng Informatics 21:243–256 . https://doi.org/https://doi.org/10.1016/j.aei.2006.12.002
3. Pinsky MR, Mancebo J, Brochard L, Hedenstierna G (2009) Applied physiology in intensive care medicine. Appl Physiol Intensive Care Med (Second Ed 1–469 . https://doi.org/10.1007/978-3-642-01769-8
4. Filipe Portela (2013). Pervasive Intelligent Decision Support in Critical Health Care, PhD Thesis, University of Minho)
5. Juarez JM, Campos M, Gomariz A, Morales A (2012) Computing problem oriented medical records. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 6924 LNAI:117–130 . https://doi.org/10.1007/978-3-642-27697-2_9
6. Skurka MA (2012) Health Information Management: Principles and Organization for Health Information Services. Wiley
7. Vercellis C (2009) Business Intelligence: Data Mining and Optimization for Decision Making
8. Sharda R, Delen D, Turban E, Liang T-P (2013) Business intelligence and analytics