Pontifícia Universidade Católica
do Rio de Janeiro

Ricardo Alfredo Quintano Neira

A multi-criteria process mining optimization tool and its application in a sepsis clinical pathway

Tese de Doutorado

Thesis presented to the Programa de Pós-graduação em Engenharia de Produção of PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Engenharia de Produção.

Advisor : Prof. Silvio Hamacher
Co-advisor: Dr. Erin Stretton

Rio de Janeiro
September 2018

Pontifícia Universidade Católica
do Rio de Janeiro

Ricardo Alfredo Quintano Neira

A multi-criteria process mining optimization tool and its application in a sepsis clinical pathway

Thesis presented to the Programa de Pós-graduação em Engenharia de Produção of PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Engenharia de Produção. Approved by the undersigned Examination Committee.

Prof. Silvio Hamacher
Advisor
Departamento de Engenharia Industrial – PUC-Rio

Dr. André Miguel Japiassú
Instituto de Pesquisa Clínica Evandro Chagas – IPEC-FIOCRUZ

Prof. Julia Lima Fleck
Departamento de Engenharia Industrial – PUC-Rio

Prof. Simone Diniz Junqueira Barbosa
Departamento de Informática – PUC-Rio

Prof. Vincent Augusto
École Nationale Supérieure des Mines de Saint-Étienne – EMSE

Prof. Márcio da Silveira Carvalho
Vice Dean of Graduate Studies
Centro Técnico Científico – PUC-Rio

Rio de Janeiro, September the 27th, 2018

Ricardo Alfredo Quintano Neira

Ricardo Quintano is graduated from PUC Minas in Computer Science and has master degree in Health Informatics from Universidade Federal de São Paulo. He has mainly worked with the development of health care solutions since 2001. He has worked as developer, business analyst, health architecture, and has coordinated several IT projects.

Bibliographic data

Neira, Ricardo Alfredo Quintano

A multi-criteria process mining optimization tool and its application in a sepsis clinical pathway / Ricardo Alfredo Quintano Neira ; advisor: Silvio Hamacher. – 2018.
158 f. : il. color. ; 30 cm

Tese (doutorado)–Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Industrial, 2018.
Inclui bibliografia

1. Engenharia Industrial – Teses. 2. Sepse. 3. Mineração de processos. 4. Análise de processos. 5. Protocolos clínicos. 6. Otimização. I. Hamacher, Silvio. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Industrial. III. Título.

CDD: 658.5

To all sepsis patients. To the sepsis community.

Acknowledgments

My PhD experience at PUC-Rio meant a lot to me. I have grown a lot professionally and personally. Many people had a greatly important role in this process.

I would like to first offer a special thanks to my advisor Silvio Hamacher for all the continuous support he gave me during these years. I have learned a lot from him. He supported me in very challenging situations.

A very special gratitude goes to my co-advisor Erin Stretton, at Philips Research. It was an incredible experience to have her counselling and work with her on the Clinical Pathways project.

It was an honour for me having the process mining and ProM guidance from Bart Hompes, Joos Buijs and Wil van der Aalst from Technische Universiteit Eindhoven (TU/e).

I also want to thank the Philips Research team. In special I would like to acknowledge Ana Leitão, Carsten Schirra, Douglas Teodoro, Gert-Jan de Vries, Gijs Geleijse, Hong Chao Nie, Jennifer Caffarel, Marcelo Santos, Ricardo Heiss, Ricardo Santos, Thomas Wendler and Zoran Hristov.

I would like to express my gratitude to Aline Medeiros, André Japiassú, Daniela Latgé Mannheimer, Fernando Bozza, Giulia Paruolo, Julia Fleck, Leonardo Bastos, Rafael Martinelli, Raphaela Gasparini, Simone Barbosa and Vincent Augusto for all their important help in my research studies.

I greatly acknowledge the support provided by Hospital Samaritano de São Paulo and Philips Blumenau in the execution of my research study regarding the evaluation of a sepsis clinical pathways.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq [grant number 140511/2018-0], and Philips Research. I would like to acknowledge CAPES, CNPq and Philips Research for performing my PhD, especially when I was abroad during my internship at TU/e.

Above all, I am extremely grateful to my family and friends. This work would have never been possible without their support and love.

Abstract

Neira, Ricardo Alfredo Quintano; Hamacher, Silvio (Advisor). **A multi-criteria process mining optimization tool and its application in a sepsis clinical pathway**. Rio de Janeiro, 2018. 158p. Tese de Doutorado – Departamento de Engenharia Industrial, Pontifícia Universidade Católica do Rio de Janeiro.

Sepsis is considered a worldwide health and economic burden. In Brazil, sepsis is the major cause of death in Intensive Care Units, as well as, one of the main causes of late hospital mortality. In this thesis, we first provide a 10 years population-level epidemiology report of sepsis in Brazil, using data from the Brazilian Unified Health System. Secondly, we present a research study that supports health care facilities in the evaluation and optimization of their sepsis clinical pathways (CP) using process mining techniques. A CP consists of a well-defined care plan, which includes a clear order and time for the execution of interventions with expected outcomes. During the execution of this study, it became clear for us the lack of existing process mining techniques for the optimization of CPs. Thus, we proposed, implemented and tested a novel process mining technique that supports users to improve their processes and we applied it for CP improvement. Our developed technique (Multi-CAT) identifies and highlights a set of activities and sub-sequences that provide positive or negative outcomes considering multiple simultaneous criteria. We successfully applied our technique in a real sepsis CP, and we acquired more optimization insights that we got in our previous manual analysis. We conclude that Multi-CAT has high potential to help in the optimization of processes with a good performance. In the most complex test scenario, with 58 activities, 50,000 cases and 14,451 variants, Multi-CAT took 2.04 minutes to execute the analysis. Finally, the tool was validated in two different processes, indicating potential capability to be applicable to other business areas.

Keywords

Sepsis; Process Mining; Process Analytics; Clinical Pathways; Optimization

Resumo

Neira, Ricardo Alfredo Quintano; Hamacher, Silvio. **Ferramenta de mineração de processos multi-critérios para otimização e a sua aplicação em um protocolo clínico de sepse**. Rio de Janeiro, 2018. 158p. Tese de Doutorado – Departamento de Engenharia Industrial, Pontifícia Universidade Católica do Rio de Janeiro.

Sepse é considerada uma carga na saúde e na economia global. No Brasil, a sepse é a principal causa de morte em Unidades de Terapia Intensiva, bem como uma das principais causas de mortalidade hospitalar tardia. Nesta tese, inicialmente apresenta-se um relatório epidemiológico brasileiro de sepse contemplando 10 anos utilizando dados do Sistema Único de Saúde (SUS). Em seguida, mostra-se um estudo que apoia os estabelecimentos de saúde na avaliação e otimização de seus protocolos clínicos de sepse usando técnicas de mineração de processos. Um protocolo clínico consiste em um plano de cuidados bem definido, que inclui uma ordem clara e tempo para a execução de intervenções com resultados esperados. Durante a execução deste estudo, identificou-se a falta de técnicas de mineração de processos para a otimização de protocolos clínicos. Assim, neste trabalho foi proposta, implementada e testada uma nova técnica de mineração de processos que auxilia usuários na otimização de seus processos. Esta técnica foi aplicada para a melhoria de protocolos clínicos. A técnica desenvolvida (Multi-CAT) identifica e destaca um conjunto de atividades e subsequências que promovem resultados positivos ou negativos, considerando múltiplos critérios simultâneos. A técnica foi aplicada com sucesso em um protocolo clínico de sepse, na qual foram adquiridas mais recomendações de otimização do que foi previamente obtido em análise manual. Conclui-se que a técnica desenvolvida apresenta grande potencial para auxiliar na otimização de processos com bom desempenho. No cenário de testes mais complexo, com 58 atividades, 50.000 casos e 14.451 variantes, Multi-CAT utilizou 2,04 minutos para executar a análise. Para finalizar, a ferramenta foi validada em dois processos distintos, indicando potencial para ser aplicada em outras áreas de negócio.

Palavras-chave

Sepse; Mineração de Processos; Análise de Processos; Protocolos Clínicos; Otimização

Table of contents

List of figures

List of tables

List of Abbreviations

AIH — *Autorização de Internação Hospitalar*
BPIC – Business Process Intelligence Challenge
BPMN – Business Process Model and Notation
CIHI – Canadian Institute for Health Information
CMD – *Conjunto Mínimo de Dados*
CNES — *Cadastro Nacional de Estabelecimentos de Saúde*
CP – Clinical Pathways
CSV — Comma Separated Values
DATASUS – *Departamento de Informática do SUS*
DES – Discrete Event Simulation
ED – Emergency Department
EHR – Electronic Health Record
ERP – Enterprise Resource Planning
IBGE – *Instituto Brasileiro de Geografia e Estatística*
ICD-10 – International Statistical Classification of Diseases and Related Health Problems 10^{th} revision
ICU – Intensive Care Unit
ILAS – *Instituto Latino Americano de Sepse*
IPCA – *Índice Nacional de Preços ao Consumidor Amplo*
HDI – Human Development Index
HIS – Hospital Information System
KPI – Key Performance Indicator
LOS – Length of Stay
Multi-CAT – Multi-Criteria Analysis Tool
P4P – Pay-for-Performance
RAM – Random Access Memory
SIHSUS – *Sistema de Informações Hospitalares do SUS*
SIM – *Sistema de Informações sobre Mortalidade*
SUS – *Sistema Único de Saúde*
UV – Unique Value
XES – eXtensible Event Stream

"Stop sepsis, save lives"

Global Sepsis Alliance, *World Sepsis Day.*

1
Introduction

According to the Latin American Sepsis Institute (*Instituto Latino Americano de Sepse*, ILAS), sepsis in Brazil "is the main cause of death in Intensive Care Units (ICU) and one of the main causes of late hospital mortality, surpassing myocardial infarction and cancer" [ILAS, 2016].

Sepsis, defined as "life-threatening organ dysfunction caused by a dysregulated host response to infection" [Singer et al., 2016], is considered a worldwide health and economic burden. The incidence of sepsis is growing and presents high lethality rates[1]. In the United States, researches [Gaieski et al., 2013; Kempker and Martin, 2016; Kumar et al., 2011; Stoller et al., 2016] show an increase in sepsis incidence and the mean lethality rate ranged from 12.1% to 35.2%. In other countries, the lethality rates varied from 3% to 46% [CIHI, 2009; Group et al., 2004; Knoop et al., 2017; Papali et al., 2017; Rodríguez et al., 2011; Zhou et al., 2017]. Health professionals and economic costs associated with sepsis are high. According to Arefian et al. [2017], the mean worldwide hospital cost per case was US$32,421.

In Brazil, from 1992 to 2006 the sepsis lethality rate was 19.9% for children [Mangia et al., 2011], and from 2002 to 2010 the proportion of sepsis-associated deaths relative to the total number of deaths presented in the Brazilian National Mortality Information System (SIM, *Sistema de Informações sobre Mortalidade*), increased from 9.77% to 16.46% [Taniguchi et al., 2014]. According to ILAS [ILAS, 2018b], the average lethality rate was 40% from 2005 to 2016 for Brazilian hospitals participants of the Surviving Sepsis Campaign. Machado et al. [2017] presented a national overall sepsis lethality rate of 55.7% (2014) for patients with hospitalization in the ICU. According to the PROGRESS study [Beale et al., 2009] (using data from December 2002 to December 2005), the overall sepsis hospital mortality rate in Brazil was 67.4%. Figure 1.1 presents a comparison of the overall sepsis hospital mortality rate in different countries (all data was extracted from Beale et al. [2009]). Related to costs, the median total costs were US$ 9,490 for private hospitals and US$ 9,773 for public hospitals (2003-2004) per episode of care in the ICU [Sogayar et al., 2008].

[1] lethality rate = (number of death cases / number of cases with a specific disease) * 100

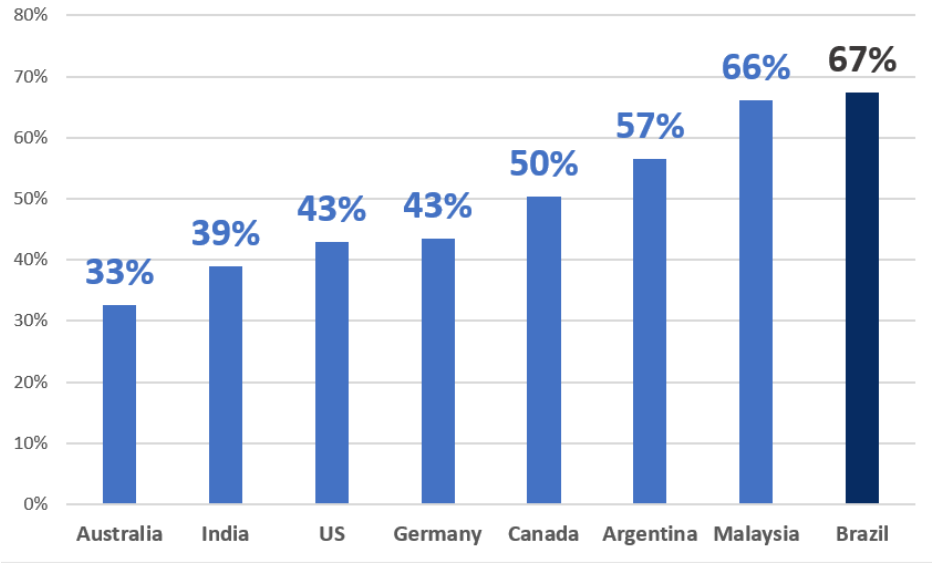


Figure 1.1: Comparison of the overall sepsis hospital mortality rate in different countries. Graph created using data from Beale et al. [2009].

Epidemiology studies of sepsis in developing countries are very scarce [Gobatto et al., 2017; Kempker and Martin, 2016; Papali et al., 2017; Silva et al., 2004; Taniguchi et al., 2014]. To the best of our knowledge, there is no Brazilian population-level epidemiological sepsis report considering all ages and severities of patients. Most of the epidemiological research studies are performed for adult patients with admission to the ICU. In addition, previous studies presented different case-mix configurations and analysis periods, making it impossible to follow trends and to get a correct status and evolution of the disease. Thus, our first objective in this thesis is to present the population-level epidemiology (all severities of sepsis, all types of hospitalizations and all patient ages) of sepsis in Brazil, considering a 10 years period using data from the Brazilian Unified Health System (SUS, *Sistema Único de Saúde*). We show important Brazilian health indicators (e.g. incidence, mortality, lethality, length of stay and associated costs) from 724,458 sepsis cases from 4,271 public and private Brazilian hospitals. We expect that our epidemiology report may help international organizations (e.g. the Society of Critical Care Medicine) and SUS in planning policies to improve the sepsis prevention and treatment.

The employment of Clinical Pathways (CP) is one possible solution to tackle the lethality of sepsis. A CP consists of a well-defined care plan, which includes a clear order and time for the execution of interventions with expected outcomes. CPs are created based on evidence-based medicine directives (guidelines), promoting the improvement and standardization in patients' care, in addition to a reduction of expenses and consumption of

resources [Baker et al., 2017; Fujino et al., 2014; Panella et al., 2003; Zhang et al., 2015]. The correct implementation of sepsis CPs promote positive outcomes like providing early sepsis recognition, right treatment, and reduction of lethality and hospital costs [Palleschi et al., 2014]. Health care providers usually design and implement their sepsis CPs based on the sepsis guidelines available from the Surviving Sepsis Campaign [ILAS, 2018a; SCCM, 2018].

However, the management and evaluation of the execution of CPs is not an easy task and has several challenges including professional time constraints, lack of qualified resources, administrative burdens and absence of tools for evaluating and reporting [Chawla et al., 2016; Khalifa and Alswailem, 2015]. Process mining has been successfully applied in a variety of areas (e.g. logistics, banking, production, healthcare) for the identification of executed processes, deviations and performance metrics [Mans et al., 2015] and we believe it can contribute significantly for the CPs evaluation.

Process mining consists of the analysis of business processes using data extracted from systems [TU/E - Math&CS Department, 2018a; van der Aalst, 2016; Yoo et al., 2016]. It allows for automatically identifying process models (process discovery), recognizing the adherence in the execution of a process (conformance checking), identifying performance metrics (performance checking), besides many other applications. We identified only two studies [Caron et al., 2014; Lismont et al., 2016] that applied existing process mining techniques to help in the evaluation and management of CPs. Hence, our second objective in this thesis is to demonstrate that the use of a set of process mining techniques is helpful in the evaluation of a sepsis CP. Differently from these two studies [Caron et al., 2014; Lismont et al., 2016], in our research, we selected and employed process mining techniques to tackle real needs and challenges identified from a Brazilian hospital. In addition, we validated the outcomes and their utility with a panel of experts. In this research study, we identified the real sepsis treatment process executed by the hospital, the adherence, deviations, performance indicators and bottlenecks in the execution of the CP. We also provided recommendations to optimize their CP, reducing the time to administer antibiotics.

During the execution of our study, it became clear to us the lack of existing process mining techniques for the optimization of CPs. As a result of a literature review, we did not find any technique that provided suggestions to re-design a process work-flow to improve outcomes and that could deal simultaneously with multiple criteria (e.g. minimize lethality and maximize profits). Thus, our last objective in this thesis is to propose, implement and test a new process mining technique that supports users to improve their

processes, considering multiple simultaneous criteria. Our developed technique (Multi-criteria analysis technique - Multi-CAT) identifies and highlights a set of activities and sub-sequences that provide positive or negative outcomes. It also supports the user in the identification of the set of recommendations that mostly contribute to improve the process with few modifications.

Even though our primary objective is to support healthcare facilities to improve CPs, our tool was developed in a generic way to be able to work in different context processes.

1.1 Objectives

In this thesis, our main objective is to propose, implement and test a process mining technique that supports users to improve a CP, considering multiple simultaneous criteria.

Our secondary objectives are to provide a broad picture of the epidemiology of sepsis in Brazil and to contribute with a set of solutions to help in the evaluation of sepsis CPs.

Our specific objectives are:

- to provide a broad population-level epidemiology report of sepsis in Brazil, considering a 10 years period using data from SUS Hospital Information System (SIHSUS) (see Chapter 2);
- to evaluate the execution of a sepsis CP in an adult Emergency Department (ED) of a Brazilian hospital through process mining techniques (see Chapter 3) and,
- to propose, implement and test a process mining technique that identifies and highlights a set of activities and sub-sequences that provide positive or negative outcomes, considering multiple criteria (see Chapters 4, 5 and 6).

We hope that this thesis can help healthcare facilities, sepsis communities and the Brazilian government to promote better sepsis outcomes.

2
Population-level epidemiology of sepsis for Brazilian hospitalizations from 2006 to 2015

As previously stated, the incidence of sepsis is growing and is associated with high lethality rates, consisting of a healthcare and economic burden. For example, in the United States, Kumar et al. [2011] present an incidence increase of 140% from 2000 to 2007, and Stoller et al. [2016] present an annual incidence increase of 26% from 2008 to 2012. The same studies show that the lethality in the US is decreasing. The mean lethality rate ranged from 12,1% to 35,2%. In other countries, the case fatality rates were: Canada 31% (2008-2009) CIHI [2009], China 20.6% (2012-2014) [Zhou et al., 2017], Colombia 3% to 46% (2007-2008) [Rodríguez et al., 2011], France 42% (2001) [Group et al., 2004], Haiti 24.2% (2012) [Papali et al., 2017], and Norway 26.4% (2011-2012) [Knoop et al., 2017].

Costs associated with sepsis are high [Arefian et al., 2017; Chalupka and Talmor, 2012] and vary according to each country and study, and factors like age, the severity of sepsis, and type of institutions influence costs. For example, the geometric mean cost for sepsis in the United States was US$19,330 (2007) [Lagu et al., 2012]; the median cost per episode of patients admitted in 21 Brazilian ICUs was US$9,632 (2003–2004) [Sogayar et al., 2008]; the mean cost per hospitalization for sepsis patients admitted in the ICU of 10 Chinese university hospitals was US$11,390 (2004–2005)[Cheng et al., 2007]; and in France the mean cost of sepsis hospitalizations in the ICU was of €22,800 (US$26,647.50 on July 26th 2018) (1997–2000) [Adrie et al., 2005].

Health indicators are essential to define strategies to improve the treatment of diseases, but the epidemiology information of sepsis in developing countries is scarce [Gobatto et al., 2017; Kempker and Martin, 2016; Papali et al., 2017; Silva et al., 2004; Taniguchi et al., 2014]. To the best of our knowledge, there is no Brazilian population-level epidemiological sepsis report considering all ages and severities of patients. Thus, in this chapter our aim is to assess trends in the incidence, lethality, and costs of sepsis for Brazilian Unified Health System (SUS) hospitalizations for the period from January 2006 to December 2015.

2.1 Materials and methods

2.1.1 Data sources

We used data from two databases available for public access from the DATASUS [DATASUS, 2018b] website. DATASUS is the Informatics Department of the Brazilian Unified Health System. The first database is the SIHSUS Hospital Information System that presents authorizations for hospital encounters (AIH – *Autorização de Internação Hospitalar*) performed under the SUS. Each AIH registry contains data from a hospital encounter: demographic information, hospital length of stay (LOS), costs, diagnoses, and patient hospital outcome [Santos, 2009]. No data present in this base have information that could identify patients. The second base is the National Registry of Healthcare Facilities (CNES – *Cadastro Nacional de Estabelecimentos de Saúde*), which is updated monthly and contains information about each facility. This database was used to define the size and type of hospitals (private or public), as well as the relation of ICU beds per total number of beds. The size of hospitals is defined according to the number of beds: small hospitals have a maximum of 50 beds, medium hospitals have from 51 to 150 beds, large hospitals have from 151 to 500 beds, and very large hospitals have more than 500 beds [Ministério da Saúde, 2018].

2.1.2 Selection of hospitalizations

As the SIHSUS database does not provide enough information to identify the presence of infection or organ dysfunction during a hospitalization, in this work we selected sepsis cases using a defined list of diagnosis codes as described in previous studies [Acosta et al., 2013; Angus et al., 2001; CIHI, 2009; Stoller et al., 2016; Taniguchi et al., 2014].

We selected AIH registries of patients with the primary diagnosis (most responsible diagnosis) of sepsis who had been hospitalized between 2006 and 2015. We used the list of sepsis diagnoses (ICD-10-CA, Canadian Revision) provided by the Canadian Institute for Health Information (CIHI) [CIHI, 2009] (Table 2.1 presents the complete list of diagnoses). Specific ICD-10-CA codes (A41.50, A41.51, A41.52, A41.58, A41.80, and A41.88) were not used, because they are not part of the SUS ICD-10 terminology. Registries of patients that had the primary diagnosis as one of those described in Table 2.1 were classified with sepsis and considered in our study.

Table 2.1: The Canadian Institute for Health Information list of International Statistical Classification of Diseases and Related Health Problems 10th revision (ICD-10) codes used to define sepsis.

ICD-10 code	**Description**
A02.1	Salmonella sepsis
A03.9	Shigellosis, unspecified
A21.7	Generalized tularaemia
A22.7	Anthrax sepsis
A23.9	Brucellosis, unspecified
A24.1	Acute and fulminating melioidosis
A26.7	Erysipelothrix sepsis
A28.0	Pasteurellosis
A28.2	Extraintestinal yersiniosis
A32.7	Listerial sepsis
A39.2	Acute meningococcaemia
A39.3	Chronic meningococcaemia
A39.4	Meningococcaemia, unspecified
A40	Streptococcal sepsis
A40.0	Sepsis due to streptococcus, group A
A40.1	Sepsis due to streptococcus, group B
A40.2	Sepsis due to streptococcus, group D
A40.3	Sepsis due to Streptococcus pneumoniae
A40.8	Other streptococcal sepsis
A40.9	Streptococcal sepsis, unspecified
A41	Other sepsis
A41.0	Sepsis due to Staphylococcus aureus
A41.1	Sepsis due to other specified staphylococcus
A41.2	Sepsis due to unspecified staphylococcus
A41.3	Sepsis due to Haemophilus influenzae
A41.4	Sepsis due to anaerobes
A41.5	Sepsis due to other Gram-negative organisms
A41.8	Other specified sepsis
A41.9	Sepsis, unspecified
A42.7	Actinomycotic sepsis
B00.7	Disseminated herpesviral disease
B37.7	Candidal sepsis
P35.2	Congenital herpesviral [herpes simplex] infection
P36	Bacterial sepsis of newborn
P36.0	Sepsis of newborn due to streptococcus, group B
P36.1	Sepsis of newborn due to other and unspecified streptococci
P36.2	Sepsis of newborn due to Staphylococcus aureus
P36.3	Sepsis of newborn due to other and unspecified staphylococci
P36.4	Sepsis of newborn due to Escherichia coli
P36.5	Sepsis of newborn due to anaerobes
P36.8	Other bacterial sepsis of newborn
P36.9	Bacterial sepsis of newborn, unspecified
P37.2	Neonatal (disseminated) listeriosis
P37.5	Neonatal candidiasis

2.1.3 Data analysis

The SIHSUS data were processed to remove duplicated registries, to adjust the age of patients, and to deflate all costs to December 2015 using the Broad National Prices Index for the Consumer (IPCA - *Índice Nacional de Preços ao Consumidor Amplo*) [IBGE, 2018b]. All costs were converted to US dollars using the rate from December 2015 (R$1 was US$0.253) [UOL Economia, 2018]. The costs presented in this work refer to what the government reimburses hospitals for sepsis hospitalizations (including the payment of the hospital staff, hospital and intensive care unit accommodations, procedures and exams). For intensive care unit (ICU) hospitalization costs, we considered only the hospitalization costs of registries that had the length of stay (LOS) in the ICU equal or greater than one day. For calculating the mean costs per case and the mean LOS we excluded the 5% extremes values to remove outliers. In this work, we considered hospital encounters that had discharge disposition as "discharge to home" or "death" since we wanted to extract the outcomes of the effectiveness of the treatment (patient lives or dies).

To calculate the sepsis incidence and mortality per 100,000 persons, we used the population projection by gender and age provided by the Brazilian Institute of Geography and Statistics (*Instituto Brasileiro de Geografia e Estatística* - IBGE) [IBGE, 2018c].

For race/ethnicity indicators, we considered data from 2008 to 2015 since there was no information available from 2006 and 2007. Regarding the hospital type, which can be private or public, for the public type, we considered federal, state, and municipal hospitals.

To understand the influence of age in the mortality rate, we created multiple logistic regression models using the patient age, gender and race as independent variables. For these analyses, we did not consider registries in which the race or gender was not informed.

The two-tailed Chi-Square test was applied for independent samples using nominal variables. To compare independent samples with continuous data (which were not normally distributed) we used the two-tailed Mann-Whitney U test. We considered the level of significance to be $\alpha = 0.05$; that is, a result was considered statistically significant whenever $p < 0.05$. To calculate the correlation between LOS and costs, we applied Pearson's test. All statistical analyses were conducted using R software [R core team, 2018].

2.2 Results

From the original AIH database with 115,392,208 records, 96,570,859 (83.69%) were non-duplicated registries with discharge disposition as "discharge to home" or "death," and 724,458 (0.63%) records were of hospitalizations with the primary diagnosis of sepsis. The flow diagram for the selection of sepsis cases can be found in Figure 2.1. These sepsis cases were treated in 4,271 different Brazilian hospitals.

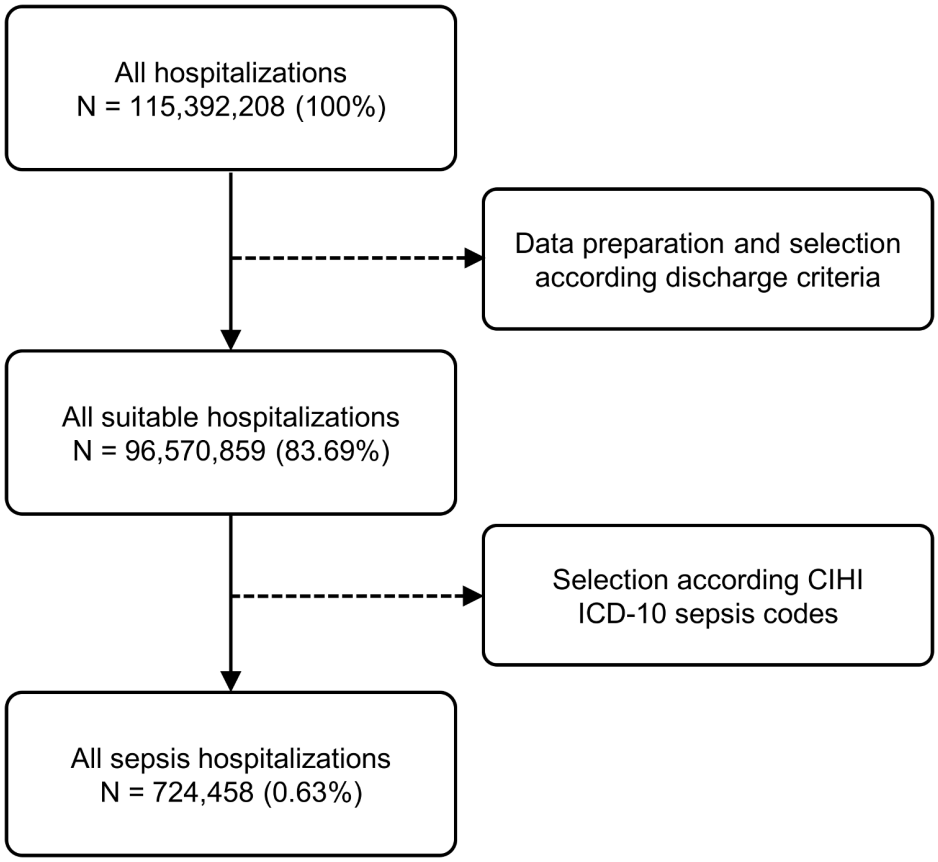


Figure 2.1: Flow diagram for a selection of sepsis cases. CIHI = Canadian Institute for Health Information; ICD-10 = International Statistical Classification of Diseases and Related Health Problems 10^{th} revision.

Figure 2.2 shows the incidence, lethality and mortality from 2006 to 2015. During this period, the incidence of sepsis increased 50.5% from 31.5/100,000 to 47.4/100,000 persons per year. The number of sepsis cases over the total number of cases (considering all diseases that had discharge disposition as "discharge to home" or "death") was 0.75% for the period.

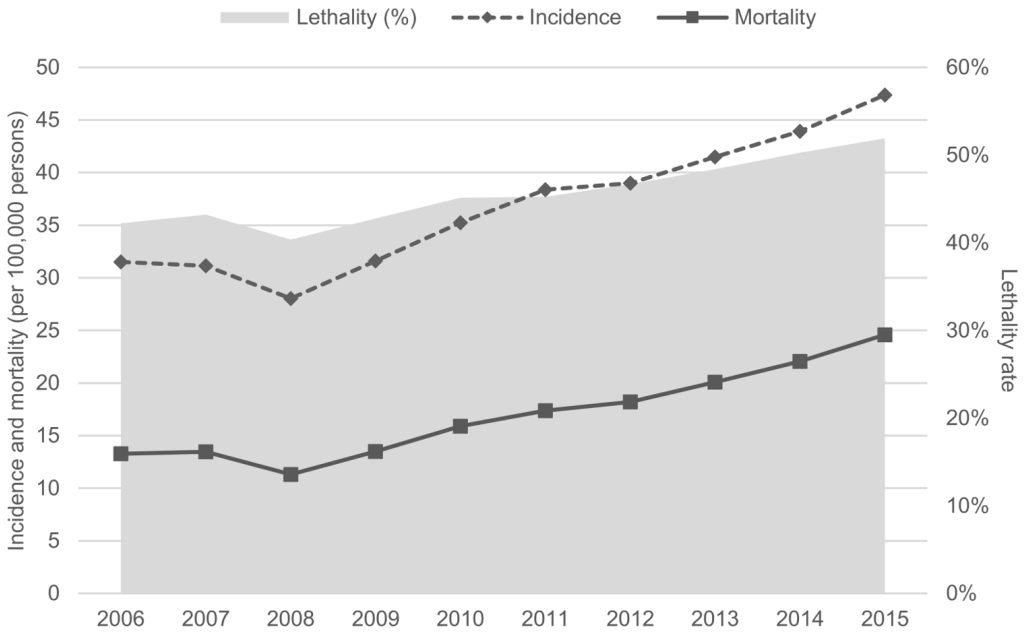


Figure 2.2: Incidence, mortality (per 100,000 persons) and lethality of sepsis from 2006 to 2015.

Table 2.2 presents the number of sepsis cases by patient characteristics from 2006 to 2015 grouped every 2 years. During this period, the average proportion of cases for female patients was 47.7%. Regarding age, the group of adults (from 18 to 64 years old) had the highest proportion of sepsis, which represented 32.5% of cases, followed by the children/teenagers group (from 0 to 17 years old) with 29.8% of cases. The number of sepsis cases for children/teenagers decreased 5.7%, and for the adults, the elderly (from 65 to 84 years old), and the very old (greater than 85 years old), it increased 68.8%, 135.0%, and 205.8%, respectively. The mean patient age was 45.2 years old increasing 34.8% (from 38.2 to 51.5 years old) during the study time frame. Regarding race, white and brown (mixed-race) subjects were the majority of individuals representing 36.8% and 26.2% of cases, respectively; black, yellow (Asian), and indigenous races represented 3.9% of cases. 33.1% of cases had no information with respect to race.

Table 2.3 presents the incidence of sepsis per 100,000 persons. From 2006 to 2015, the incidence for female patients increased 53.2% from 29.5/100.000 to 45.2/100.000 persons per year and for male patients, it increased 47.6% from 33.6/100.000 to 49.6/100.000. Regarding age, the group of very old people had the highest incidence of sepsis, which represented 517.6/100,000 persons, followed by the elderly group at 169.0/100,000 persons. The sepsis incidence for children/teenagers increased 0.5%, and for the adults, the elderly, and the very old, it increased 47.2%, 72.2%, and 86.7%, respectively.

Table 2.2: Number of sepsis cases by patient characteristics from 2006 to 2015.

Indicator / Attribute	**2006-2007 N**	**2008-2009 N**	**2010-2011 N**	**2012-2013 N**	**2014-2015 N**	**Average N (%)**	**Annual Growth Percentage**[a]
Total							
Cases of sepsis	118,014	114,834	144,636	161,100	185,874	72,446[b]	64.1%
% from all hospitalizations	0.60%	0.59%	0.74%	0.85%	0.99%	0.75%	76.3%
Cases per gender							
Male	62,098	60,413	75,742	84,263	96,553	37,907[b] (52.3%)	60.8%
Female	55,915	54,421	68,894	76,837	89,321	34,539[b] (47.7%)	67.9%
Cases per age							
0-17	43,920	40,012	44,860	44,699	42,708	21,620[b] (29.8%)	-5.7%
18-64	37,964	37,763	46,998	52,523	60,503	23,575[b] (32.5%)	68.8%
65-84	28,865	29,440	41,215	49,056	62,247	21,082[b] (29.1%)	135.0%
85+	7,265	7,619	11,563	14,822	20,416	6,169[b] (8.5%)	205.8%
Mean patient age in years	39.1	40.9	44.4	47.0	50.8	45.2	34.8%
Race/Ethnicity[c]							
White	NA[d]	43,810	54,067	58,044	67,345	27,908[b] (36.8%)	69.2%
Black	NA[d]	3,605	4,733	4,961	5,936	2,404[b] (3.2%)	97.4%
Brown (Mixed-race)	NA[d]	25,475	33,953	42,899	56,578	19,863[b] (26.2%)	163.1%
Yellow (Asian)	NA[d]	481	591	653	1,296	378[b] (0.5%)	355.6%
Indigenous	NA[d]	439	301	246	278	158[b] (0.2%)	-45.4%
No Information	NA[d]	41,024	50,991	54,297	54,441	25,094[b] (33.1%)	41.0%
Death cases	50,388	47,799	65,359	76,638	95,009	33,519[b]	101.9%

[a]Growth percentage from 2006 to 2015.
[b]Annual average.
[c]Following the nomenclature and order presented in SIHSUS database.
[d]NA = not available.

Table 2.3: Incidence and mortality per 100,000 persons, and lethality by patient characteristics from 2006 to 2015.

Indicator / Attribute	**2006-2007** N	**2008-2009** N	**2010-2011** N	**2012-2013** N	**2014-2015** N	**Average**	**Annual Growth Percentage**[a]
Incidence	31.3	29.8	36.8	40.2	45.6	36.9	50.5%
Incidence per gender							
Male	33.3	31.7	39.0	42.6	48.0	39.1	47.6%
Female	29.4	28.0	34.7	38.0	43.3	34.8	53.2%
Incidence per age groups							
0-17	35.7	32.9	37.4	37.8	36.8	36.1	0.5%
18-64	16.5	15.8	19.1	20.8	23.3	19.2	47.2%
65-84	132.6	127.5	167.1	185.0	217.0	169.0	72.2%
85+	384.1	363.7	493.5	561.8	692.3	517.6	86.7%
Mortality	13.4	12.4	16.6	19.1	23.3	17.1	85.0%
Lethality	42.7%	41.6%	45.2%	47.6%	51.1%	46.3%	23.0%
Lethality for ICU hospitalizations	60.7%	61.7%	63.0%	65.2%	68.4%	64.5%	14.4%

[a]Growth percentage from 2006 to 2015.

Table 2.4 shows the number of sepsis cases per type and size of hospital. The average proportion of cases for private hospitals was 49.9%. During the study period, the number of sepsis hospitalizations for private hospitals increased 33.0% and for public hospitals, increased 103.5%. With respect to the size of hospitals, large hospitals treated the highest proportion of sepsis cases (47.8%) followed by medium (33.0%), very large (10.1%) and small hospitals (9.1%). From 2006 to 2015, the number of sepsis cases for small, medium, large and very large hospitals increased 15.7%, 41.8%, 94.5% and 74.4%, respectively.

From 2006 to 2015, the mortality due to sepsis increased 85.0%, going from 13.3/100,000 to 24.6/100,000 persons per year. The overall lethality rate of sepsis was 46.3%, and for hospitalizations with admission to the ICU, it was 64.5%. The number of sepsis deaths over the total number of deaths (all admissions) was 8.2% for the period. The average lethality rate for female patients (46.8%) was higher than male patients (45.8%) (Chi-square test results: $\chi^2 = 77.9$; df = 1; $p < 0.001$; two-tailed). With respect to the age of patients, very old patients had the highest lethality rate of 75.9% followed by the elderly, the adults, and the children/teenagers groups with 67.7%, 49.3%, and 13.6%, respectively ($\chi^2 = 154{,}830$; df = 3; $p < 0.001$; two-tailed). The lethality rate for children/teenagers decreased 40.1% and for adults, elderly and very old, it increased 11.5%, 6.1%, and 2.8%, respectively. Regarding the race/ethnicity, indigenous and brown (mixed-race) patients had the smallest lethality rate of 30.1% and 42.1%, respectively, while black, yellow (Asian), and white patients had the rate of 52.0%, 51.6%, and 49.9%, respectively ($\chi^2 = 2{,}643.9$; df = 4; $p < 0.001$; two-tailed). The sepsis lethality rate in public hospitals (55.5%) was higher than private hospitals (37.0%) ($\chi^2 = 25{,}036$; df = 1; $p < 0.001$; two-tailed). Regarding the size of hospitals, small hospitals had the smallest lethality rate with an average of 22.9%, medium hospitals had an average of 39.7%, large hospitals had 51.7%, and very large hospitals had 63.2% ($\chi^2 = 30{,}991$; df = 3; $p < 0.001$; two-tailed).

Table 2.4: Number of sepsis cases and lethality per hospital type and size from 2006 to 2015.

Indicator / Attribute	**2006–2007 N**	**2008–2009 N**	**2010–2011 N**	**2012–2013 N**	**2014–2015 N**	**Average N (%)**	**Annual Growth Percentage**[a]
Cases per type of hospital							
Private	64,787	61,373	72,187	78,655	84,196	36,120[b] (49.9%)	33.0%
Public	53,062	53,416	72,440	82,441	101,675	36,303[b] (50.1%)	103.5%
Death cases per type of hospital							
Private	23,132	19,280	24,475	29,806	36,868	13,356[b] (39.9%)	68.8%
Public	27,188	28,511	40,880	46,831	58,141	20,155[b] (60.1%)	131.3%
Lethality per type of hospital							
Private	35.7%	31.4%	33.9%	37.9%	43.8%	37.0%	26.9%
Public	51.2%	53.4%	56.4%	56.8%	57.2%	55.5%	13.6%
Cases per size of hospital							
Small	12,892	10,989	12,892	14,096	14,899	6,577[b] (9.1%)	15.7%
Medium	41,704	39,309	48,091	51,892	57,964	23,896[b] (33.0%)	41.8%
Large	51,628	52,870	70,011	78,935	92,945	34,639[b] (47.8%)	94.5%
Very large	11,430	11,571	13,633	16,173	20,063	7,287[b] (10.1%)	74.4%
Death cases per size of hospital							
Small	2,625	2,340	2,848	3,407	3,859	1,508[b] (4.5%)	59.5%
Medium	15,109	13,258	19,152	21,545	25,888	9,495[b] (28.3%)	83.1%
Large	25,691	24,334	34,325	41,777	52,800	17,893[b] (53.4%)	125.6%
Very large	6,822	7,855	9,030	9,908	12,462	4,608[b] (13.8%)	79.5%
Lethality per size of hospital							
Small	20.4%	21.3%	22.1%	24.2%	25.9%	22.9%	38.2%
Medium	36.2%	33.7%	39.8%	41.5%	44.7%	39.7%	29.3%
Large	49.8%	46.0%	49.0%	52.9%	56.8%	51.7%	16.1%
Very large	59.7%	67.9%	66.2%	61.3%	62.1%	63.2%	3.0%

[a]Growth percentage from 2006 to 2015.
[b]Annual average.

Figure 2.3 presents the sepsis incidence (per 100,000 persons) and lethality according to each age group. Analyzing the figure, we can note that children younger than 1-year-old had a high incidence of sepsis (475.9 cases per 100,000 persons) with a lower lethality rate (13.1%). Children in the group of 5–9 years old had the lowest lethality rate (12.3%). In contrast, people more than 90 years old had the highest incidence of sepsis (601.4 cases per 100,000 persons) with the highest lethality rate (77.5%). Performing multiple logistic regressions (death as a dependent variable, and age, gender, and race as independent variables), we observed that age has a strong association with mortality. When the age variable is removed, the prediction error increases from 29.6% to 46.8%. Performing a logistic regression with death as a dependent variable and age as an independent variable, the prediction error was 29.6%. All ROC curves are available in Appendix A. Analyzing the results of the last logistic regression, we conclude that the older the patient is, the higher the probability is to die when diagnosed with sepsis (for 1-year change in age, the odd to die increases 1.036 times).

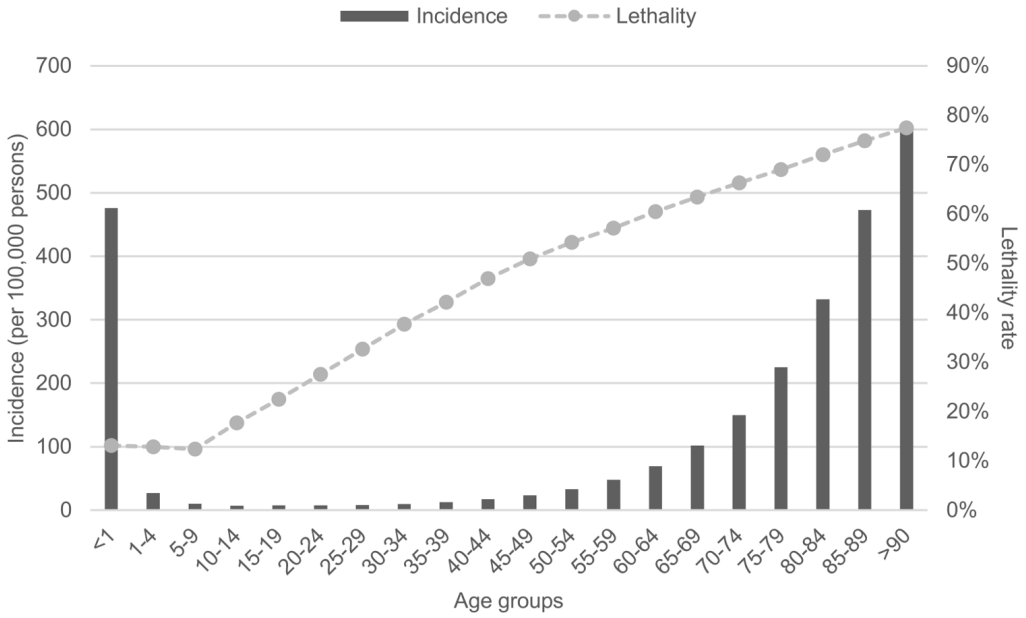


Figure 2.3: Sepsis incidence (per 100,000 persons) and lethality rate according to the age groups.

Regarding race lethality, brown (mixed-race) and indigenous races had the lowest lethality rates (42.1% and 30.1%, respectively). Comparing the age of brown and indigenous races with white, black, and yellow (Asian) races, the age is significantly lower for brown (mean 40.1 years; median 46 years) (Mann-Whitney U test results: U = 15,572,966,908; n1 = 159,680; n2 = 247,170; $p < 0.001$; two-tailed) and for indigenous patients (mean 17.0 years; median 1

year) (U = 67,934,676; n1 = 1,281; n2 = 247,170; $p < 0.001$; two-tailed) than for the other races (mean 52.4 years; median 61 years). Female patients had in average higher lethality rates (46.8%) than male patients (45.8%). Comparing the age of each gender, male patients were in average younger (mean 43.7 years; median 52 years) than female patients (mean 46.9 years; median 56 years) (U = 60,897,122,525.5; n1 = 379,069; n2 = 345,388; $p < 0.001$; two-tailed). Thus, as brown and indigenous patients, in general, were younger than the other race groups, as well as, male patients were younger than female patients, we can consider that age is an important attribute to explain these differences in lethality rates.

Table 2.5 presents hospitalizations, costs, and LOS grouped every 2 years. The mean cost per hospitalization was US$624.0 (median US$353.9) with a growth percentage of 23.5% between 2006 and 2015, and the mean cost per ICU hospitalization was US$1,708.3 (median US$1,293.1) with a growth percentage of 42.2%. The mean hospitalization LOS was 9.0 days (median 7.0 days), while the mean LOS in the ICU was 8.0 days (median 6.0 days). From 2006 to 2015, the LOS decreased 2.2% going from 9.3 to 9.1 days and the ICU LOS increased 5.2% going from 7.7 to 8.1 days.

The mean hospitalization LOS for public hospitals was higher than private hospitals, with a difference of 35.5% ranging from 7.6 days (median 6.0 days) to 10.3 days (median 8.0 days) (U = 51,666,193,009; n1 = 344,310; n2 = 345,562; $p < 0.001$; two-tailed).

The average daily hospitalization cost was US$94.4 (median US$66.1) for private hospitals, and US$90.3 (median US$58.2) for public ones (U = 54,091,530,825; n1 = 344,893; n2 = 342,978; $p < 0.001$; two-tailed). The mean case cost was US$586.2 (median US$343.6) for private hospitals, and US$662.9 (median US$372.4) for public hospitals (U =57,418,114,611.5; n1 =347,283; n2 =345,789; $p < 0.001$; two-tailed). Related to the average cost per case, public hospitals had higher costs than private hospitals mainly because the LOS in public hospitals was higher. The Person's correlation between LOS and costs was 0.71, indicating a strong positive relationship.

Analyzing lethality rates, LOS, and mean cost per case, we can observe that, in general, private hospitals have a more effective treatment for sepsis compared with the public hospitals. This means that private hospitals treat patients faster than public hospitals, with lower total costs and lower lethality rates. We observed a small difference in the average age of patients that were treated in private (44.8 years; median 53 years) and public hospitals (45.7 years; median 55 years) (U = 64,847,829,992.5; n1 = 361,198; n2 = 363,034; $p < 0.001$; two-tailed). Performing a logistic regression for each type of hospital

Table 2.5: Hospitalizations, costs and LOS of sepsis from 2006 to 2015.

Indicator / Attribute	**2006-2007 N (SD)**	**2008-2009 N (SD)**	**2010-2011 N (SD)**	**2012-2013 N (SD)**	**2014-2015 N (SD)**	**Average**	**Annual Growth Percentage**[a]
All hospitalizations	118,014	114,834	144,636	161,100	185,874	72,446[b]	64.1%
ICU hospitalizations	33,466	30,543	40,477	47,894	58,542	21,092[b]	83.9%
% ICU hospitalizations	28.4%	26.6%	28.0%	29.7%	31.5%	29.1%	12.0%
Hospitalizations costs[c]							
Total cost (US$ million)	78.2	104.4	133.1	143.5	156.1	61.5[b]	113.3%
Mean cost per case (US$)[d]	512.6 (420.6)	658.7 (683.7)	669.6 (701.8)	648.1 (687.2)	619.2 (662.2)	624.0	23.5%
Costs for hospitalizations requiring ICU[c]							
Total cost (US$ million)	48.2	71.0	93.0	103.0	114.1	42.9[b]	159.0%
Mean cost per case (US$)[d]	1,220.2 (838.3)	1,928.9 (1,457.5)	1,927.5 (1,416.9)	1,808.8 (1,325.6)	1,657.1 (1,195.6)	1,708.30	42.2%
Mean LOS							
Hospitalization (days)[d]	9.3 (7.5)	8.8 (7.4)	8.6 (7.3)	9.0 (7.7)	9.1 (7.6)	9.0	-2.2%
ICU (days)[d]	7.8 (6.7)	8.1 (7.0)	7.8 (6.7)	8.0 (6.8)	8.2 (6.8)	8.0	5.2%

[a]Growth percentage from 2006 to 2015.
[b]Annual average.
[c]Deflated costs (December 2015, IPCA).
[d]5% extremes excluded.

with death as the dependent variable and age as the independent variable, we observed that the age has similar partial slope coefficients for both cases (private with 0.0337 and public with 0.0385). Thus, the age variable does not explain the inefficiency in public hospitals. Nevertheless, we cannot exclude possible differences in the type of patients admitted and the severity of illness between public and private hospitals, as we did not have access to this sort of information. During the study period, the percentage of ICU beds (total ICU beds/total beds) in private hospitals (6.0%) was close to the percentage in public hospitals (5.5%) (U = 2,498,347,306.5; n1 = 83,138; n2 = 60,843; $p < 0.001$; two-tailed), indicating that both types of hospitals probably have similar resources to monitor and treat severe patients, taking into account the premise that all ICU beds contain all necessary monitor equipment available and properly working.

To give a general overview of the efficiency of hospitals for treating sepsis, Figure 2.4 presents an efficiency matrix. We created it inspired in the efficiency matrix presented by Salluh et al. [Salluh et al., 2017]. Each dot is a group of hospitals of the same size and type. Each dot was added in the matrix according to its average LOS (Y-axis) and average lethality rate (X-axis). The letter near each dot represents the size of the hospital group (S = small, M = medium, L = large, V = very large). The shape of the dot represents the type of hospital (diamond shapes are private, square shapes are public). Hospital groups that are located in the bottom left part of the matrix are more efficient than the groups that are located in the top right. We can observe that smaller hospitals are more efficient than larger hospitals and private hospitals are more efficient than public ones. However, we cannot exclude possible case-mix differences between each group of hospitals, as we did not have access to this type of data. Figures 2.5 and 2.6 present the box plots for LOS and lethality rate per hospital size and type. The number of hospitals and cases per hospital group of the efficiency matrix can be found in Appendix B.

2.3 Discussion

We observed that the number and incidence of SUS sepsis hospitalizations increased from 2006 to 2015. This phenomenon can be related to population aging. According to the World Health Organization [WHO, 2018], Brazilian life expectancy has grown from 73 years to 75 years (from 2006 to 2015), which has increased the older people groups [IBGE, 2018a]. During the study period, the mean age of septic patients in the current research study

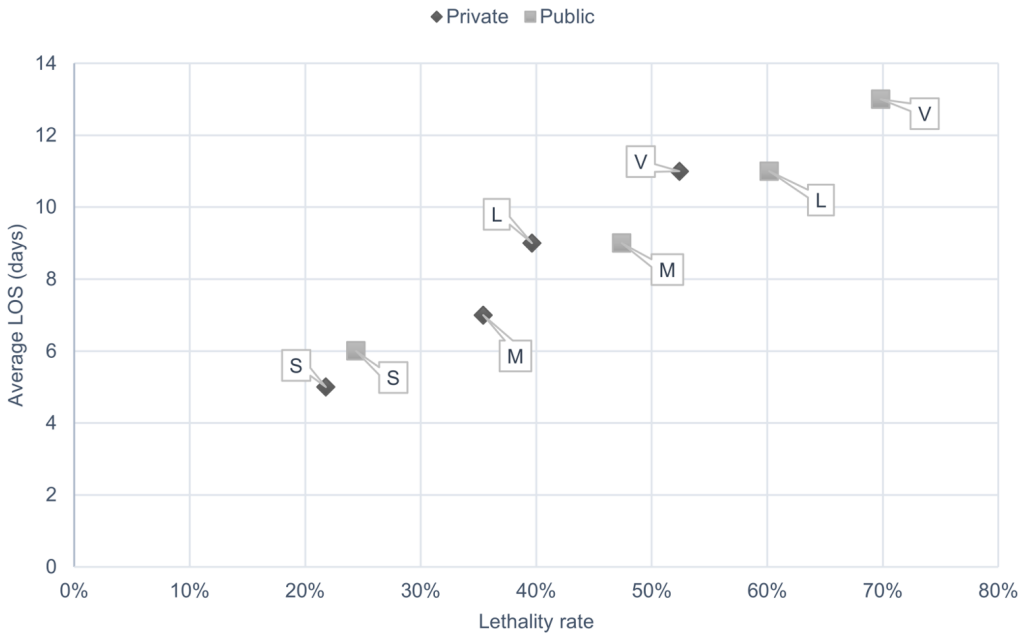


Figure 2.4: Treatment efficiency matrix for sepsis per hospital size and type. Letters refer to the size of hospitals: M = medium size; L = large size; S= small size; V = very large size. LOS = length of stay.

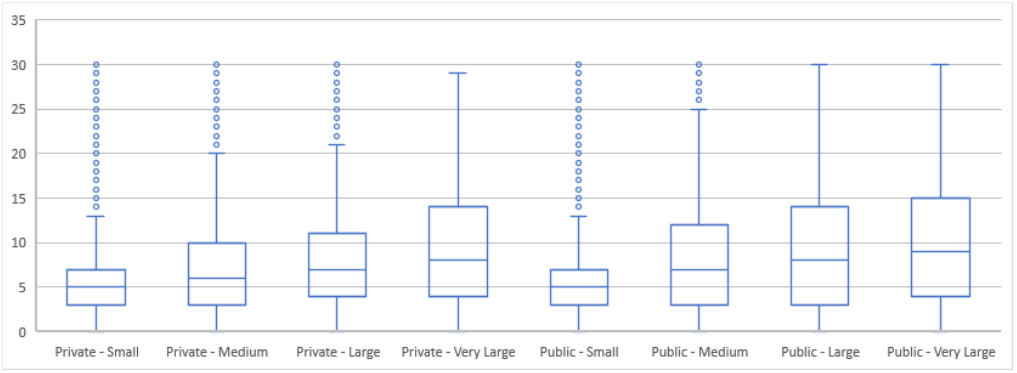


Figure 2.5: LOS box plots per hospital size and type. Hospitalizations with LOS greater than 30 days are not considered in the chart.

increased from 38.2 to 51.5 years old. Another reason can be associated to initiatives like the Surviving Sepsis Campaign [ILAS, 2018a; SCCM, 2018], which could improve the awareness of health professionals to provide the right diagnosis of sepsis [Gaieski et al., 2013]. Finally, as presented in previous studies, the increase in the number of patients with immunocompromised status [Gobatto et al., 2017], such as cancer and HIV positive patients, and patients using immunosuppressive drugs, which leads to a higher risk to present sepsis with concurrent infections [Japiassú et al., 2010] could also have contributed to the growth in the incidence.

Analyzing the multiple logistic regressions results, age presented a high

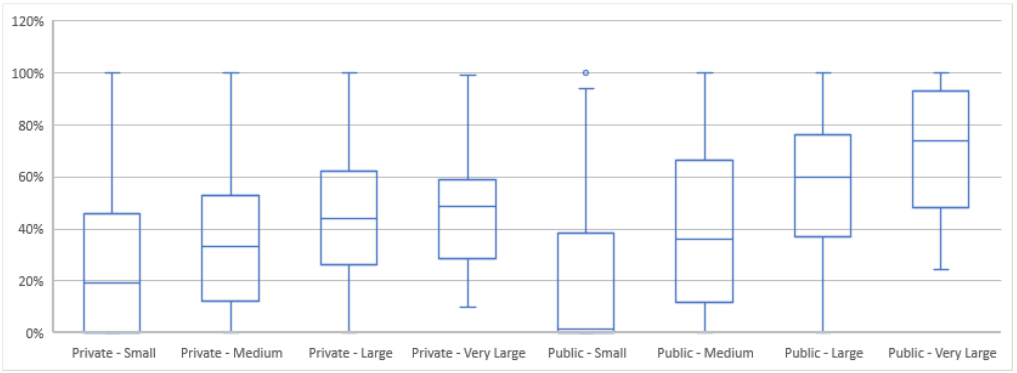


Figure 2.6: Lethality rate box plots per hospital size and type.

association with patients' death (for 1-year change in age, the odds to die increase 1.036 times). From 2006 to 2015, the incidence of sepsis cases for the elderly and the very old age groups increased 72.2% and 86.7%, respectively, and these two groups had the highest average lethality rate of 67.7% and 75.9%, respectively. Nevertheless, during the study period, we observed that the lethality rate for adults, elderly and very old increased 11.5%, 6.1%, and 2.8%, respectively, indicating that age is not the only factor for worsening mortality.

The fact that the lethality rate in public hospitals (55.5%) was higher than private ones (37.0%), can be a result of delayed sepsis recognition and treatment, or fewer resources in the public hospitals than in the private ones, as previously published [Conde et al., 2013]. The different characteristics in care provided by each type of hospital, the structure of the ICU health team, the delay in transferring the patient to the ICU, and the access to the best standard of care may contribute to these different rates in mortality [Silva et al., 2004].

The overall sepsis lethality rate was 46.3%, and for hospitalizations with admission to the ICU, it was 64.5%. The mortality rates presented in studies for countries with a similar Human Development Index (HDI) (2014) [United Nations Development Programme, 2017] to Brazil (HDI 0.755) were: China (HDI 0.727) with 33.5–48.7% for sepsis (multicenter investigations in ICUs) [Liao et al., 2016]; Colombia (HDI 0.720) with 21.9% for sepsis, and 45.6% for septic shock (multicenter investigation of cases of patients admitted in the emergency department, general wards and ICUs) [Rodríguez et al., 2011]; and Mexico (HDI 0.756) with 30.4% (multicenter investigation in ICUs) [Carrillo-Esper et al., 2009]. The lethality rates presented in the current study are higher than the lethality rates presented above, even when comparing to countries with a lower HDI, such as Colombia and China.

Concerning hospitals costs, the mean cost per hospitalization was US$624.0 (median US$353.9) and per ICU hospitalization was US$1,708.3 (median US$1,293.1). Applying the ratio of the purchasing power parity (PPP) conversion factor to the market exchange rate [The world bank, 2018], the cost was US$1,040.0 (median US$589.8) per hospitalization and US$2,847.2 (median US$2,155.2) per ICU hospitalization. As Sogayar et al. [2008] discussed, a simple comparison between studies is not easy, since reimbursement rates, costs, price factors, and healthcare systems may vary. Nevertheless, the values presented in the current work are low if we compare them with the costs of other sepsis studies. Chalupka and Talmor [2012] presented that the costs per sepsis case had a wide variability ranging from US$4,888 (Argentina) to US$103,529 (United States). Comparing the median ICU case costs (US$2,155.2) of the current research study (2006-2015), that represent the government reimbursement to hospitals for treating sepsis, with the median case costs of 21 ICUs of private and public Brazilian hospitals (US$9,632) presented by Sogayar et al. [2008] (2003-2004), one may suppose that the reimbursement values are lower than the effective costs. According to Victora et al. [2011], in Brazil, private institutions debate that the reimbursement provided by SUS barely allows them to cover all costs. The SUS's low reimbursement to hospitals for treating sepsis may be one of the reasons for the high lethality rates.

The current SUS reimbursement model, in which the hospital receives a fix value as base for the main executed procedure, has limitations, e.g, it does not consider the patient's severity [GTRH, 2010]. The careful evaluation and possible modification of the SUS reimbursement model could possibly lead to better outcomes. For example, the Pay-for-Performance (P4P) model [Greene and Nash, 2009], in which the hospital receives a bonus in case of good quality measures, could encourage health care facilities to look for strategies to improve their treatments.

Even though this work provides 10 years of sepsis indicators, it has limitations. As we had no access to clinical variables (type of infection, severity scores, vital signs, laboratory exam results, co-morbidities), we could not improve the regression models to have a more precise estimative, we could not verify the accuracy in the selection of cases, and we could not define the severity of the sepsis (definition of case-mix). Jolley et al. [2015a] performed a research study using data from three Canadian hospitals to identify the sensitivity and specificity in selecting septic patients using the CIHI ICD-10 codes. The sensitivity was 46.4% and the specificity was 98.7% for selecting ICU adult septic patients. For selecting non-ICU septic patients, the sensitivity

was 6.7% and the specificity was 100%. These results indicate that in our study we underestimated the total number of sepsis cases.

It was a challenge for us to define the right list of ICD-10 codes to use in this study since there is no official list. There are several lists of ICD-10 codes used in different studies [Jolley et al., 2015b]. In addition, as Tsertsvadze et al. [2016] presented, it is difficult to determine the true sepsis incidence of a population because there is an absence of valid standard methods for defining sepsis. Thus, we consider it is important to create a unique list of sepsis diagnosis codes and a standard approach for selecting sepsis cases. This would provide the generation of homogenized indicators, which would allow appropriate comparisons and would encourage health professionals to use the right diagnosis codes.

More than 75% of the Brazilian population depends on and exclusively uses the SUS health services [Watts, 2016]. The rest of the population has access to private health services, but they can also use the SUS services since they are available to any person. Thus, since we could not obtain the exact percentage of users that exclusively accessed private health services, we decided to use the Brazilian population projection [IBGE, 2018c] without any adjustment to generate the sepsis incidence and mortality per 100,000 persons.

2.4 Conclusions

The incidence of SUS sepsis hospitalizations increased 50.5% in Brazil during the period from 2006 to 2015. The overall lethality rate of sepsis was 46.3%, and for hospitalizations with admission to the ICU, it was 64.5%. During the study period, the lethality rate for children/teenagers improved, but for all other age groups it became worse with an increase of 11.4%. The lethality rate in public hospitals (55.5%) was higher than in private hospitals (37.0%), which possibly reflects the differences in the number of resources for processes and/or structure. The SUS's low reimbursement to hospitals for treating sepsis may be one of the reasons for the high lethality rates.

In next chapter we will present the evaluation of the execution of a sepsis CP using a set of process mining techniques.

3
Evaluation of the execution of a sepsis clinical pathway in the emergency department through process mining techniques

In Chapter 2 we presented a broad sepsis epidemiological report of Brazilian SUS hospitalizations. The overall sepsis lethality rate was 46.3%, and for hospitalizations with admission to the ICU, it was 64.5%. These results clearly show that the Brazilian sepsis situation needs close attention.

The employment of clinical pathways (CP) is one possible solution to tackle the high lethality rates as it promotes early sepsis recognition, correct treatment process and the improvement of patient's outcomes [Palleschi et al., 2014]. Yet, the management and evaluation of the execution of CP's is not an easy task, as there are many challenges like professional time constrains, lack of qualified resources, administrative burdens and lack of tools for evaluating and reporting [Chawla et al., 2016; Khalifa and Alswailem, 2015].

Thus, in this chapter we present the evaluation of the execution of a sepsis CP from an adult Emergency Department (ED) of a Brazilian hospital through process mining techniques. We demonstrate that the use of a set of process mining techniques can help the hospital staff in evaluating their sepsis CP.

Our research study was performed using data from Hospital Samaritano de São Paulo [Hospital Samaritano, 2018b]. The hospital staff extracts monthly Key Performance Indicators (KPI) to check the adherence in the execution of the CP (for example the adherence with respect to the administration of volume expansion and antibiotics). The extraction process of these KPIs is arduous, and the hospital has a dedicated team to consolidate all required information. In addition, these KPIs do not allow them to identify all deviations and performance indicators in the process. Considering this scenario, we believe that the use of process mining techniques may help the hospital staff to better evaluate and manage their sepsis CP.

3.1
Background

In this section we present some basic process mining concepts that are important for understanding the techniques applied and developed in this

thesis. All concepts were extracted/adapted from van der Aalst [2016].

Process mining consists of the analysis of business processes using data extracted from systems. It allows for automatically identifying process models (process discovery), recognizing the adherence in the execution of a process (conformance checking), identifying performance metrics (performance checking), besides many other applications. Process mining was applied in several research studies in health care areas [Rojas et al., 2016; Yang and Su, 2014]. For example, Augusto et al. [2016] presented an automatic way to simulate and evaluate CPs. They applied their method in a severe heart failure case study. Xu et al. [2016] created a method that summarizes patients daily interventions and improves the discovery of executed CPs. They implemented their method for intracerebral hemorrhage. Fernandez-Llatas et al. [2015] introduced a tool and a method to visualize an executed process using indoor location systems data. They tested the tool in the surgical area.

Activity (or transition) is a possible action to be executed in a process. For example, in the sepsis CP use case we present some activities like "Registry of Triage", "Registry of Clinical Notes".

Arc is a connection between 2 *activities*. *Arcs* are used in process models to define rules regarding the execution order of *activities*.

An *event* consists in the execution of a specific *activity* by a *case*. In its most basic form, it is composed of a unique identifier representing a *case* (e.g. hospitalization ID, patient ID), the performed *activity* for that *case* (e.g. "Administration of Antibiotic" or "Patient Discharge"), and the date and time that the *activity* was performed. An event can store other types of attributes (variables), like for example, the patient gender, age, type of discharge, vital signs value, resource. An example of *event* is: the *activity* "Registry of Triage" was performed at "20/07/2017 02:00PM" for the patient Joe Smith (case ID 487974).

A *case* is one executed instance of the process and it is composed of one or more *events*. A *case* in the health care context can be the process followed by a patient to perform a radiographic examination.

A *complete case* is a *case* from the *event log* that has a start and an end of a given process. When executing process mining analyses it is essential to consider only *complete cases* as the utilization of incomplete cases may generate incorrect outputs (e.g. wrong mean case duration, wrong analysis of deviations).

An *event log* is a repository of *events* from an executed process. The standard format to store an *event log* is the eXtensible Event Stream (XES) [IEEE, 2016]. *Event logs* serve as input for most process mining techniques.

A *process variant* (or executed path, sub-graph) represents a unique ordering of *activities* executed in a process. For example, in the context of sepsis, "Registry of Triage" >[1] "Registry of Clinical Notes" > "Prescription of Antibiotic" > "Administration of Antibiotic" represents one *variant*, and "Registry of Triage" > "Prescription of Antibiotic" > "Administration of Antibiotic" > "Registry of Clinical Notes" represents a second *variant*. Each *variant* is executed by one or more different *cases*.

Sub-sequence is a part of a *process variant*. For example, in the *process variant* "A > B > C > D", "A > B" and "B > C > D" are *sub-sequences*. The relation between *activities* in a given *event log* can happen in a direct or indirect way. Activity A is *directly followed* by activity B if activity B happens immediately after activity A. The representation of the relation *directly followed* is "A > B". Activity A is *eventually followed* by activity B if activity B happens after activity A, but not necessarily immediately after it. The representation of the relation *eventually followed* is "A » B". For example, in the *process variant* "A > B > C > D", activity A is *directly followed* by activity B and it is *eventually followed* by activities C and D.

3.2 Materials and methods

3.2.1 Sepsis clinical pathway – the normative process

Figure 3.1 presents the Emergency Department (ED) CP process model. The process starts when a patient arrives in the ED and gets a queue number (activity 1). The patient waits until a receptionist calls them using the queue number. The receptionist registers the patient's admission in the Hospital Information System (HIS) (activity 2). The patient waits at the reception until a nurse calls them to start the triage. During the triage, the nurse measures the patient's vital signs (temperature, blood pressure, heart rate and respiratory rate) (activity 3), asks specific questions to the patient, prioritizes the patient according to their severity level, and registers all the triage information in the Electronic Health Record (EHR) (activity 4). If the nurse suspects that the patient has sepsis, than they must start the formal sepsis pathway (activity 5A). The nurse must fill out a specific sepsis CP paper-based form, must register a clinical note in the EHR that the sepsis CP was started and must communicate immediately with the first available physician that the patient is a sepsis suspect. The physician evaluates the patient (anamnesis and physical

[1] A > B means that activity A was directly followed by activity B

examination) and records all information in the clinical notes (activity 6). If the physician corroborates a sepsis diagnostic hypothesis, then they must prescribe antibiotics and volume expansion (activities 7 and 9) and must request the blood culture and lactate exams (activities 8 and 10). If the sepsis CP was not formally started previously, then the physician must start it (activity 5B after activity 6). Finally, the patient goes to the medication room and a nurse technician collects the blood culture and lactate exams (activities 11 and 13) and administers the antibiotics (activity 14) and the volume expansion (activity 12). The designed CP presents all sepsis treatment steps until the administration of antibiotics, representing the first bundle of the Surviving Sepsis Campaign [SSC, 2015] (grey area of Figure 3.1). The administration of antibiotics must happen after the collection of the blood culture and, as prescribed by the hospital, this activity must be executed within one hour of the sepsis presentation (when the formal sepsis CP starts). The faster the patient receives the antibiotics, the greater is their survival probability [Kumar et al., 2006].

3.2.2 Data extraction

We extracted data from the HIS of Hospital Samaritano located in São Paulo city, Brazil. It is a large private hospital (with more than 300 beds) and it stands out for excellence in healthcare [Hospital Samaritano, 2018a; Revista Exame, 2018].

We extracted 2 years of sepsis hospitalizations, constituting 4,516 cases. We selected cases of patients diagnosed with sepsis and/or who died of sepsis, identified through the list of International Statistical Classification of Diseases and Related Health Problems 10^{th} revision (ICD-10) sepsis codes provided by the Canadian Institute for Health Information [CIHI, 2009]; or cases in which the sepsis medication template for prescribing was used. The list of diagnosis codes is presented in Chapter 2 (Table 2.1).

We extracted administrative (patient admission, patient discharge, declaration of death) and clinical data (vital signs, clinical notes, prescription items, medication administration, exam collection and results). In the extraction, we anonymized all patient and hospitalization data to guarantee that no-one could identify a patient or associate the extracted data with the HIS database. In this stage, following the "Safe Harbor" de-identification method from the HIPAA Privacy Rule [HIPAA, 2012], to anonymize the data:

– We encoded any identification code like patient/professional codes, hospitalization IDs, chart numbers, and prescription and administration codes;

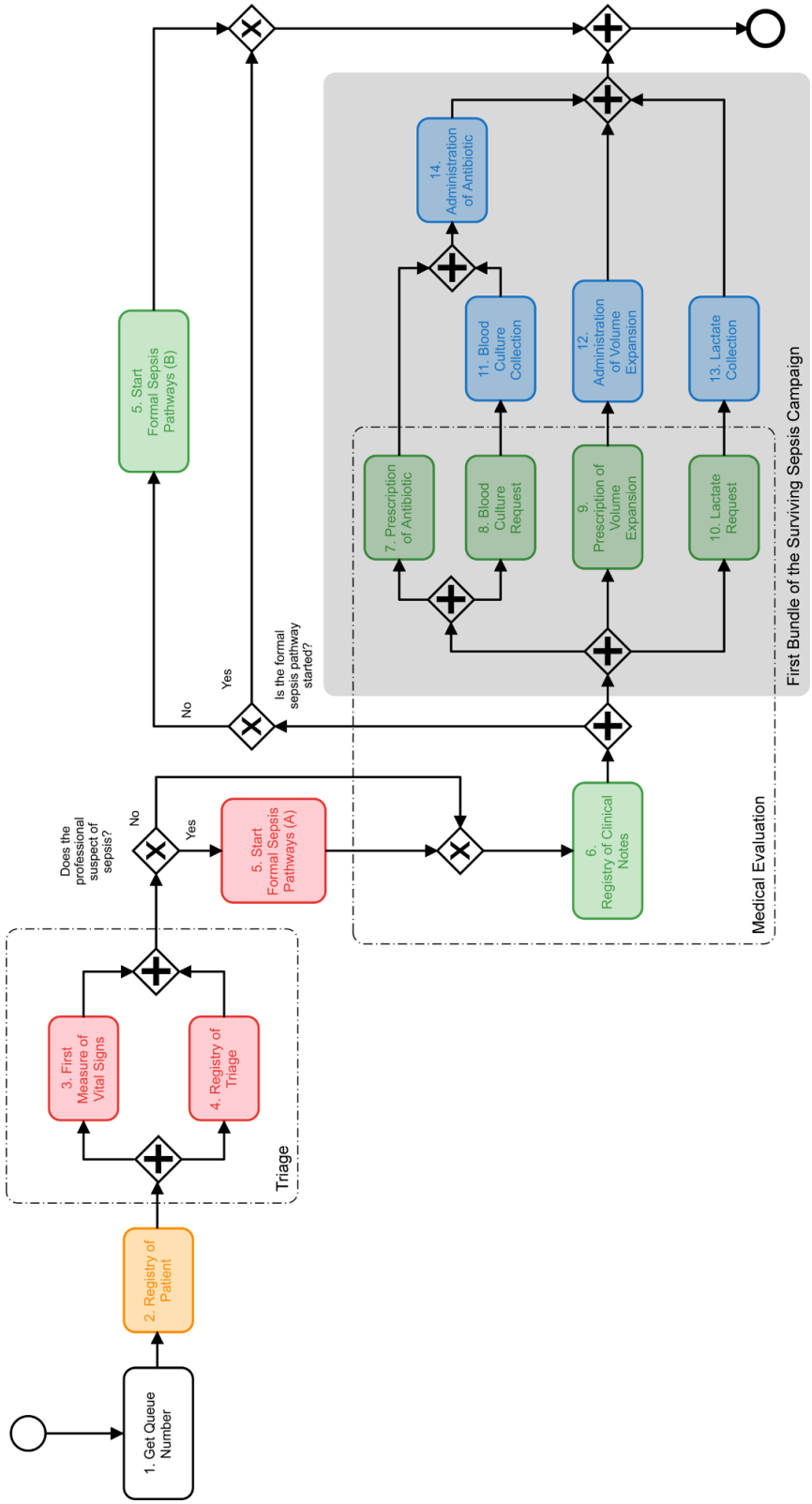


Figure 3.1: Sepsis clinical pathway of the emergency department (normative process) using the Business Process Model and Notation (BPMN). The orange activity is performed by a receptionist; pink activities are executed by nurses; green activities are performed by physicians; blue activities are executed by nurse technicians.

- Instead of extracting the patient's day of birth, we calculated their age based on the admission date. All cases of patients older than 90 years were classified as 90 years old since they have higher probability of being identified;
- All cases of patients with weight greater than 130kg were classified as 130kg since they have higher probability of being identified;
- All dates were shifted to a given time interval;
- We encrypted names of patients and professionals, specific numbers (e.g. chart, bed, hospitalization numbers) from text fields (like clinical and discharge notes).

More details of the extraction phase can be found in Appendix C.

3.2.3 Preparation of the event log

To execute process mining algorithms, the HIS system data was converted to an event log. The created event log contains all complete sepsis cases that started and ended in the ED or cases that started in the ED with transfer to the wards. All patients are adults (>= 18 years old), with the presence of two or more Systemic Inflammatory Response Syndrome signals and with sepsis suspicion from a physician. We used these selection criteria to reduce the heterogeneity in the case mix.

The resulting event log has 1,710 cases, 20,605 events, 14 activities (those described in Figure 3.1) and 292 variants. The granularity of time is minutes. Table 3.1 presents the number of events per activity in the event log.

3.2.4 Data analysis

We performed conformance checking analyses to verify the adherence and to find deviations in the process. Each deviation was classified in one category: activity executed in a different order, activity not performed, activity performed by a different role, and violation of target time. With the identified deviations and observed behaviors from the event log, we updated the CP process model (normative model) to create the AS-IS model (what the ED staff actually executed to treat septic patients). To identify bottlenecks in the process, we executed performance checking. In this last analysis, we removed outliers cases using the interquartile rule (we removed all cases that had the case duration greater than the interquartile range * 1.5, that means the case duration should be less or equal to 67 minutes). We conducted all process

Table 3.1: Number of events per activity in the event log.

Activity	**Events**	
	N	**% of cases**
1. Get Queue Number	1,710	100%
2. Registry of Patient	1,710	100%
3. First Measure of Vital Signs	1,710	100%
4. Registry of Triage	1,694	99%
5. Start Formal Sepsis Pathway	1,710	100%
6. Registry of Clinical Notes	1,710	100%
7. Prescription of Antibiotic	1,672	98%
8. Blood Culture Request	1,695	99%
9. Prescription of Volume Expansion	187	11%
10. Lactate Request	1,669	98%
11. Blood Culture Collection	1,657	97%
12. Administration of Volume Expansion	187	11%
13. Lactate Collection	1,633	96%
14. Administration of Antibiotic	1,661	97%

mining analyses using ProM [TU/E - Math&CS Department, 2018b] and Disco [Fluxicon, 2018].

For deviations related to the order of activities, i.e., activities performed in a different order compared to the CP, we verified if they could reduce the time for the administration of antibiotics (target time). We defined two groups for comparison of the mean target time: Group A, cases in which the order of activities happened as prescribed in the CP (e.g. "registry of clinical notes" followed by "prescription of antibiotic"); and Group B, cases in which the order of activities happened in a different order (e.g. "prescription of antibiotic" followed by "registry of clinical notes"). We used the Mann-Whitney U test to compare pairs of groups (with continuous data, which were not normally distributed; $\alpha = 0.05$). We conducted all statistical analyses using R software [R core team, 2018].

3.2.5 Validation of results

We performed the validation of this research with 3 physicians, 2 nurses, and 1 quality analyst who all actively work in the sepsis CP. Two of the physicians manage the sepsis CP. The validation happened in July of 2017 with a group interview and structured questionnaires. The questionnaires are presented in Appendix D.

We validated the AS-IS process, 4 deviations (one of each category), bottlenecks and the deviations that could potentially optimize the execution of the CP.

3.3 Results

3.3.1 Conformance analysis and deviations

Confronting the event log with the sepsis CP (normative) model, we identified that the trace fitness was 0.85, indicating that the hospital has performed the process close to the one defined in its CP. The trace fitness is a number that varies from 0 to 1 that shows how well the model can reproduce the event log. A number close to 1 means that the model represents well the event log reality. Table 3.2 presents a detailed list of conformance analysis measurements.

Table 3.2: Process mining measurements for conformance analysis.

Measurement	Value
Trace Fitness	0.85
Move-Model Fitness	0.77
Move-Log Fitness	0.98
Precision	0.86
Backwards Precision	0.72
Balanced Precision	0.79

We identified 43 different types of deviations in the execution of the CP, constituting 5,184 instances of deviations. Table 3.3 presents the list of the 10 most frequent types of deviations. The non-prescription and non-administration of volume expansion were the most frequent type of deviations (89% of cases), followed by the start of the formal CP during triage (45%).

3.3.2 The real executed process (AS-IS)

To obtain the real executed process model, we adjusted the CP model incorporating the most frequent types of deviations and observed behaviors from the event log. The adjusted model had a better fitness (0.97). Figure 3.2 presents the AS-IS model. All the differences between the CP and the AS-IS model are:

- The first measure of vital signs (activity 3) usually happens before the registration of triage (activity 4);
- The formal start of the sepsis pathway (activity 5A) usually starts during triage;

Table 3.3: List of the 10 most frequent deviations identified with conformance checking analysis.

Deviation	Cases		Category
	N	%	
Cases without prescription of volume expansion	1,523	89%	activity not performed
Cases without administration of volume expansion	1,523	89%	activity not performed
Start of formal sepsis pathway during triage	768	45%	activity executed in a different order
Registry of clinical notes after prescription of treatment	390	23%	activity executed in a different order
Start of formal sepsis pathway before triage	160	9%	activity executed in a different order
Activity '8. Blood Culture Request' performed by 'Nurse Technician'	130	8%	activity performed by a different role
Cases without lactate collection	77	5%	activity not performed
Activity '8. Blood Culture Request' performed by 'Laboratory Assistant'	75	4%	activity performed by a different role
Blood culture collection without its request	55	3%	activity executed in a different order
Cases without blood culture collection	53	3%	activity not performed
Lactate collection without its request	50	3%	activity executed in a different order

– The prescription and administration of volume expansion are not always executed.

3.3.3 Performance analysis and bottlenecks

Table 3.4 presents the average waiting time between activities in the process. This table is an output from the ProM plug-in "Replay a Log on Petri Net for Performance Conformance Analysis". Evaluating Table 3.4, we identified two places in the process that we consider potential bottlenecks, since they represented significant waiting times until the administration of antibiotics:

– Patients waiting in the reception before triage: mean of 18 minutes – from "1. Get Queue Number" to "3. First Measure of Vital Signs" activity;
– Prescription of medications and request of exams: mean of 5 minutes – from "6. Registry of Clinical Notes" to "7. Prescription of Antibiotic", "8. Blood Culture Request" and "10. Lactate Request" activities. We did not consider in this analysis the "9. Prescription of Volume Expansion" activity since its frequency was low (11% of cases).

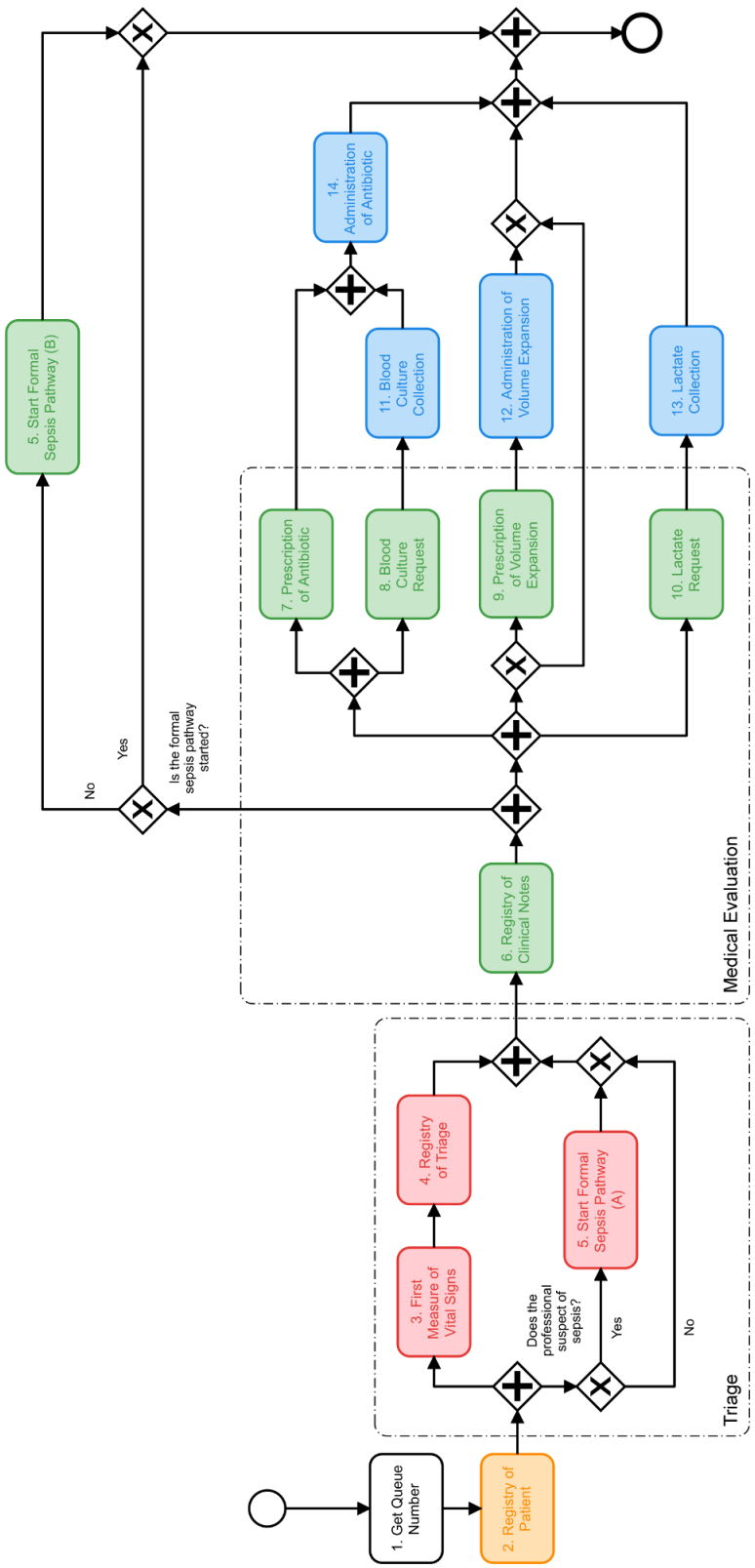


Figure 3.2: AS-IS sepsis treatment process model of the emergency department using the Business Process Model and Notation (BPMN). The orange activity is performed by a receptionist; pink activities are executed by nurses; green activities are performed by physicians; blue activities are executed by nurse technicians.

Table 3.4: Average waiting time (in minutes) between activities in the execution of the sepsis clinical pathway. White cells indicate that there is no connection between activities.

From\To	1. Get Queue Number	2. Registry of Patient	3. First Measure of Vital Signs	4. Registry of Triage	5. Start Formal Sepsis Pathway	6. Registry of Clinical Notes	7. Prescription of Antibiotic	8. Blood Culture Request	9. Prescription of Volume Expansion	10. Lactate Request	11. Blood Culture Collection	12. Administration of Volume Expansion	13. Lactate Collection	14. Administration of Antibiotic
1. Get Queue Number	0.00	8.40	17.70	19.90	22.80	26.60	30.70	30.60	25.50	30.60	33.30	49.80	33.30	34.10
2. Registry of Patient		0.00	9.38	11.50	14.50	18.30	22.30	22.30	19.00	22.30	25.00	43.30	25.00	25.80
3. First Measure of Vital Signs			0.00	2.13	5.89	8.95	13.10	13.10	12.00	13.00	15.80	36.40	15.70	16.60
4. Registry of Triage				0.00	8.35	6.85	11.10	11.00	10.10	11.00	13.70	34.50	13.70	14.50
5. Start Formal Sepsis Pathway			1.32	1.07	0.00	6.13	9.33	9.32	8.57	9.28	12.00	32.50	12.00	12.80
6. Registry of Clinical Notes					0.41	0.00	5.08	5.08	4.16	5.03	7.77	27.40	7.73	8.57
7. Prescription of Antibiotic					6.68		0.00	11.10		3.15	2.66	24.00	2.63	3.50
8. Blood Culture Request					6.70		0.07	0.00	0.00	0.01	2.66	24.10	2.68	3.57
9. Prescription of Volume Expansion					9.32		0.34	0.27	0.00	0.33	2.19	24.30	2.19	3.03
10. Lactate Request					6.64		0.07	0.09	0.00	0.00	2.71	24.30	2.66	3.58
11. Blood Culture Collection					4.02		2.21			2.18	0.00	21.60	0.00	1.02
12. Administration of Volume Expansion					0.28							0.00		
13. Lactate Collection					3.94		2.40	14.60			0.11	21.80	0.00	1.02
14. Administration of Antibiotic					3.23					1.18		21.40	2.00	0.00

3.3.4
Analysis of deviations that can optimize the process

The prescription of the treatment before registering clinical notes (cases that had the prescription of medicines and request of exams eventually followed by the registry of clinical notes) was the deviation we identified that could optimize the execution of the CP, reducing the time for the administration of antibiotics by 3.5 minutes (median 3 minutes) (Mann-Whitney U test results: U = 261,472.5; n1 = 366; n2 = 1,199; $p < 0.001$; two tailed).

3.3.5
Validation of results with hospital staff

With respect to the deviations, the hospital staff considered three of them (registry of clinical notes after prescription of treatment, cases without prescription of volume expansion, and antibiotic not administrated until 1 hour since identification) as real deviations and one of them as not valid (blood culture was requested by nurse technician) since, in practice, it should never happen.

Regarding the AS-IS process, the hospital staff recognized it as the one executed in the ED and they would not change anything in the presented model.

Concerning the bottlenecks, the health team considered both of them as real and they did not provide any new bottleneck from their own experience.

About the deviation that could reduce the time for the administration of antibiotics, the hospital staff agreed that prescribing the treatment before registering the clinical notes clearly reduces the administration target time since the delivery of medication process starts early. However, in general, they believe that the CP should not be updated since this deviation is associated to severe patients. According to them, physicians know when and how to prioritize the treatment of a patient, and updating the CP would remove their autonomy.

All professionals considered important to have access to the outcomes presented in this research, since they can help the hospital staff to identify problems in the execution of the CP. For them, this is a key feature to improve their CP.

3.4
Discussion

This research study showed that the use of process mining techniques can support hospitals in the evaluation of CPs. This work could provide the identification of the adherence, deviations, the AS-IS process, performance

indicators and a deviation that can improve outcomes in the execution of the CP. The careful extraction of the HIS data, the rigorous preparation of the event log and active communication with health professionals were key elements for the success of this work.

With respect to the adherence in the execution of the CP, the lack of prescription and administration of the volume expansion (89% of cases) was the main factor to decrease the trace fitness. If we set the volume expansion as optional in the clinical pathway model, the fitness increases to 0.97.

The two bottlenecks identified in the process can possibly seem to represent a small amount of time (mean of 18 and 5 minutes). If we compare them with the target time to administrate antibiotics (1 hour), they represent a significant amount of time (30% and 8%). It is important to remember that the hospital stands out for excellence in healthcare, thus, probably in other hospitals with different profiles, these waiting times can be even higher. Regarding the bottleneck of patients waiting in the reception, the hospital staff suggested a simplification in the registration of triage information. In addition, they will update the process to perform the triage before the registration of the patient. This action will help the hospital to identify earlier the sepsis suspicious patients. With respect to the bottleneck of the prescription of the treatment, the hospital staff suggested a simplification in the registration of the prescription.

We believe that the update of the sepsis CP considering our optimization recommendation, which entails prescribing first and then registering the clinical notes, would benefit the process and patients. This would maximize the probability that the expected outcomes would be achieved in any operational condition, regardless of patient severity. In this way, the time to give antibiotics would probably be reduced for all sepsis patients.

The application of the CP evaluation method we presented in this chapter could benefit Brazilian hospitals to manage and improve their sepsis CPs. A pre-requisite for its execution is that the registration of hospitalization information (e.g. triage, medical evaluation, prescription) is done using integrated systems, allowing the creation of the event log. Sadly, the adoption of HIS and EHR in Brazil is still small [Computer World, 2012; iMedicina, 2017].

3.4.1 Limitations

Regarding the data quality, for providing the event times in the event log, we used the time that the action was entered in the system instead of the time that the action was actually performed (execution time). The execution

time was rarely available in the HIS data. Nonetheless, in our visits to the hospital we observed that the staff registers the information in a very close time of executing the action.

The administration of the antibiotics and the collection of blood culture for some hospitalizations had the same registration time and, according to the CP, the administration of antibiotics must happen after the blood collection. This behavior happens as nurse technicians execute all treatment and later register all executed actions in the HIS. To solve this issue, we added one minute to all events of the "14. Administration of antibiotic" activity.

We created the volume expansion prescription and administration activities calculating and analyzing the total amount of solution given to each patient. These activities were added in the event log if the total given solution volume, in a period of 4 hours, was greater than the expected for a volume expansion (30ml * patient weight, considering a threshold of 80%). We followed this approach since there was no information in the prescription to indicate the intention of the physician when prescribing a solution.

The hospital staff considered the use of the time that the action was entered in the system (instead of the execution time), adding one minute to the administration of antibiotics and the creation of the volume expansion activities as valid assumptions.

3.4.2 Challenges

During the execution of this research study we faced the following challenges:

- Regarding the data extraction from the HIS, the complete process took us more time than we expected. The major reasons were: A. The HIS was not initially planned to manage the sepsis clinical pathway; B. Limited time availability from the HIS development team to help us in mapping all required database tables and attributes; C. It was challenging for us to understand the data structure of the HIS; D. We had to guarantee that all extracted data was correctly de-identified; E. We had restricted time to extract all data to not compromise the HIS usage by the hospital staff;
- The interpretation of the deviations identified by the ProM plugin "Replay a Log on Petri Net for Conformance Analysis" was not easy. All results are associated to a Petri Net model. For example, ProM presented that from place p1 (before the execution of activity "2. Registry of Patient") two cases moved on log to the activity "3. First Measure

of Vital Signs". Thus, we converted all ProM deviations to meaningful deviations. Regarding our example, we converted the deviation to "two cases anticipated triage";

- The tools to automatically discover processes (e.g. Disco or ProM plugins like Inductive miner, Heuristic miner) generated process models that are very different from the CP model (normative model). For example, Inductive miner removed the dependency arcs from prescription activities to treatment activities, making all of them in parallel. Thus, we opted to update manually the CP model incorporating the most frequent types of deviations and observed behaviors from the event log;
- We did not find any process mining tool to provide recommendations to improve the CP. All our analyses to identify deviations that provided positive outcomes were performed in a manual way;
- We had to adapt the research validation as the hospital staff had limited time availability to support this initiative.

3.5 Conclusions

In this chapter, we have successfully applied process mining techniques to evaluate the execution of a sepsis CP. The hospital performed the process very close to the defined in its sepsis CP, demonstrating a very high health care quality, as the treatment started in average within 13 minutes (the time recommended by the clinical pathway is 60 minutes). We consider the results of this research study very promising since they can help the hospital in the management of the CP and can reduce their burden in the extraction of KPIs as was confirmed through a structured interview with the hospital staff.

In the next chapter we will present a novel process mining technique that helps in the optimization of processes.

4
Multi-criteria analysis technique

In Chapter 3, we identified actions that could improve the sepsis treatment process in the ED. For this task, we identified CP deviations (regarding the order of activities) and for each of them, we verified their contribution to improve the process. This work was done manually and demanded a significant execution time. Taking into consideration the hospital reality, in which it is arduous to generate monthly sepsis KPIs, adding such manual type of demand probably would not be practical. Thus, it became necessary to identify an approach that promotes process optimization with a minimal work from the user.

We looked for process mining techniques that could provide insights to optimize a process considering a set of criteria defined by the user. For example, taking into account our sepsis use case, we want to answer in a simple way "what are the set of actions executed in the past that can reduce the time to administrate the antibiotics?". Another example can be related to a selling process "what are the set of activities and sub-sequences that maximized the profit, and minimized cancellation rate and case duration?".

We executed a literature review (January 29, 2018) using Scopus [Elsevier, 2018] and Web of Science [Clarivate Analytics, 2018] citation indexing services. Our query was: *("improvement" OR "enhancement" OR "optimization" OR "optimisation") AND ("process mining" OR "process analytics")*. We considered manuscripts that describe process mining techniques with the aim to provide insights to optimize process work-flows. We excluded papers that present optimization for process mining techniques (e.g. process discovery improvement, reduction of event log noise, model simplification, repairing alignments), other types of techniques (process discovery, performance, petri net model extension), methods for executing process mining projects, and case studies applying process mining. We evaluated the title and/or abstract from 426 non-duplicated publications, and from this subset, we selected 7 papers for complete screening. We evaluated the references of these 7 publications and added a new one in our results. A total of 8 papers met our defined criteria. Figure 4.1 presents a diagram for the selection of publications.

We classified the publications identified from the literature review in

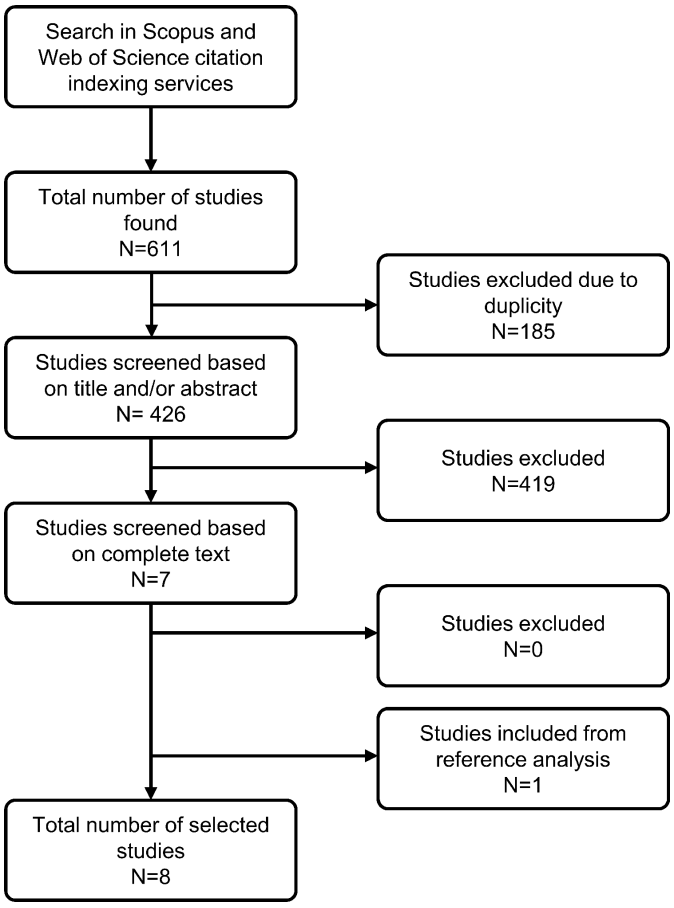


Figure 4.1: Selection diagram of publications in the literature review.

three categories according to their aim. Two studies provide insights for re-designing a process (update the activities and arcs of a normative process). Lakshmanan et al. [2013] implemented an approach to identify frequent patterns associated to outcomes. The technique helps users identify similar behaviors from a group of cases, presenting them in a visual way, highlighting frequent patterns in a discovered model. In addition, the approach allows the user to apply trace clustering to remove outlier variants. Dees et al. [2017] describe a method that uses a process model and an event log as input to reconstruct the model incorporating deviations that contribute to improve a KPI value, respecting constraints/rules defined in the model. The method detects deviations confronting the event log with the model, correlates the observed behavior from the deviations to the KPI values, and finally updates the model to incorporate the detected positive behaviors.

Two manuscripts from the same research project analyze the influence of case features to outcomes. Lehto et al. [2016, 2017] propose a generic method using influence analysis that is based on process mining, root cause analysis and classification rule mining. The method processes and evaluates an event

log, and presents as output a list of case antecedents (that can be an attribute, sequence of activities, the number of times a specific activity was executed) and their influence in the process outcomes. The approach helps users to define the order for implementing improvement actions as it provides the impact of change of each antecedent.

Four publications provide insights for optimizing the resource (workers) allocation for executing a set of activities. Low et al. [2014, 2016] and van der Aalst et al. [2015] propose an approach to update an event log changing the resource allocation, and in this way, recreating an optimal execution scenario. The new event log can help users to identify process improvements. For the resource allocation, they applied optimization techniques like integer linear programming, tabu search, hill climbing and a hybrid genetic approach developed by them. Ikeda et al. [2014] propose a formal concept analysis using a pair of features from different process perspectives. The method estimates a weakness value in the combination of two features (e.g. a professional executing a specific activity) from a given event log. The weakness value helps users to identify points in the process that require attention. In their research study, they combined resource and control-flow perspectives, and in this way, the weakness value supports the user to identify points in the process that require attention in the resource allocation.

Table 4.1 presents all selected manuscripts with a summary of their characteristics. Lakshmanan et al. [2013] and Dees et al. [2017] describe approaches that provide recommendations for re-designing a process based in outcomes, which is close to what we executed manually in our sepsis use case. Even though both approaches have the purpose to support process optimization, they have limitations, like for example, none of them considers multiple simultaneous criteria or identifies behaviors that have high contribution for the improvement of outcomes with a small process update. Thus, our last objective in this thesis is to propose, implement and test a process mining technique that identifies a set of activities and sub-sequences executed in the past that can contribute to improve a process, considering a set of criteria defined by the user. In section 4.3 we detail the benefits of our approach.

The technique we propose, using an event log and a set of criteria defined by the user, presents as output a set of activities and sub-sequences that can potentially improve the process. The technique also provides a list of executed variants ordered according their contribution to better process outcomes (KPIs). As a simple example, we could provide as input the event log from our previous research study and define as criterion that we want to minimize the time to give antibiotics. The technique processes the event

Table 4.1: Summary of characteristics of the research studies selected in the literature review.

Characteristic	**Dee et al.**[a]	**Ikeda et al.**[b]	**Lakshmanan et al.**[c]	**Lehto et al.**[d]	**Low and Aalst et al.**[e]
Allows the re-evaluation of constraints from the normative process	No	No	Yes[f]	No[f]	No
Analyzes the influence of case features to outcomes	Yes	No	Yes	Yes	No
Considers multiple simultaneous criteria	No	No	No	No	Yes
Considers (or highlights) behaviors that contributed to diminishing outcomes	No	No	Yes	Yes	No
Considers (or highlights) behaviors that contributed to improve outcomes	Yes	No	Yes	Yes	No
Identify behaviors with high contribution for improvement with small process changes	No	Yes	No	Yes	No
Is implemented as a unique tool or plugin	No	No	No	Yes	No
Resource optimization	No	Yes	No	No	Yes
Process re-design	Yes	No	Yes	No	No

[a]Dees et al. [2017]
[b]Ikeda et al. [2014]
[c]Lakshmanan et al. [2013]
[d]Lehto et al. [2016, 2017]
[e]Low et al. [2014, 2016] and van der Aalst et al. [2015]
[f]Information is not explicit in the manuscript.

log and, as a result, presents that the execution of the "prescription of the treatment" eventually followed by the "registration of clinical notes" was the sub-sequence that generated better outcomes. With this result, the hospital can evaluate their sepsis clinical pathway and decide if they want to incorporate the identified behavior in their process.

In this chapter we will present the concepts of the proposed technique. In the next chapter we will describe its implementation and usage. In the following section we will describe a simple loan request process that will be used to exemplify the proposed technique. Then we will explain the technique itself.

4.1 The loan request process

In this section we will present a simple fictitious example of a loan request process that will be used to exemplify the Multi-criteria analysis technique. This process was inspired from the one used in the Business Process Intelligence Challenge 2017 [BPIC, 2017]. Figure 4.2 presents the loan request process model.

The process starts when a customer requests a loan to the bank. The customer can ask for any amount of money for a specific target (e.g. to buy a house or a car). The customer can start the process accessing a website (activity A) or in person by going to the bank (activity B). Once the request is formalized, the customer needs to provide a set of required documents (e.g. birth certificate, proof of income) (activity D). If the customer takes more than one week to perform the previous activity, the bank calls them to remember to send their documents (activity C). Once the institution receives all documentation, the bank validates the loan request (activity E), approving (activity F) or rejecting it (activity G). During activities C, D and E, the customer can cancel the loan request (activity H). The process finishes after the execution of activities F, G or H.

The bank acts in several countries and, in our example, the quality team wants to get insights into which actions can improve the process for the subsidiaries located in the Netherlands. The bank wants to get recommendations to improve their process, reducing the case duration and increasing the profit for approved cases.

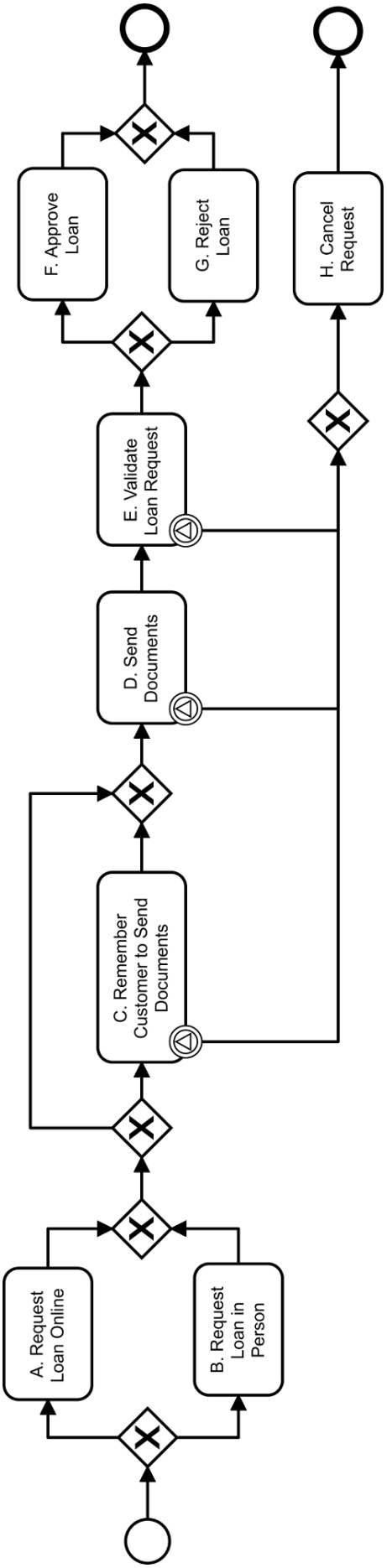


Figure 4.2: Simple loan request process using the Business Process Model and Notation (BPMN) (Note: this model was created as an elucidative example. It does not represent a real process).

4.2 The proposed technique

In this section, we detail the concept of our process mining technique. The aim of this technique is to support the user in optimizing/improving their process, using retrospective data and considering a set of multiple simultaneous criteria.

Figure 4.3 presents an overview of the proposed process mining technique. The approach is composed of 9 steps. In short, the event log is prepared in Steps 1 and 2. Step 3 groups cases by variant. In Step 4, the set of criteria to be used in the analysis is defined. Step 5 treats the event log data, removing outliers and infrequent cases, and normalizing the attribute values. Step 6 generates a unique value or indicator to each variant based on their outcomes. In Step 7, the variants are clustered according their unique value allowing the recognition of variants that promoted similar outcomes. Step 8 simplifies the variants to help in the interpretation and analysis of results. And finally, Step 9 compares the clusters with positive and negative outcomes and provides the results.

Below we will detail each step of the technique. During the explanation of each step, we will exemplify its application using the loan request process presented previously. A detailed description of the utilization of our technique to the loan request process is presented in Appendix E.

4.2.1 Step 1 - Data collection

Retrospective data must be extracted from an information system, preprocessed (e.g. remove duplicated registries, treatment of missing values) and converted to an event log. During this step, it is important to know the database structure and have a clear understanding of the process to be analyzed, as it supports the selection of the tables/variables to extract data and assists the creation of the activities of the event log [Mans et al., 2015].

During the execution of this step, with the intention to simplify the process, it is strongly recommend to evaluate carefully the process and remove all unnecessary (or not important) activities. The process simplification helps in the analysis of process mining results. This simplification procedure requires a deep business knowledge of the process.

Regarding our example, the bank extracted data from 30 different tables from their system and created the event log using the same activities presented in Figure 4.2.

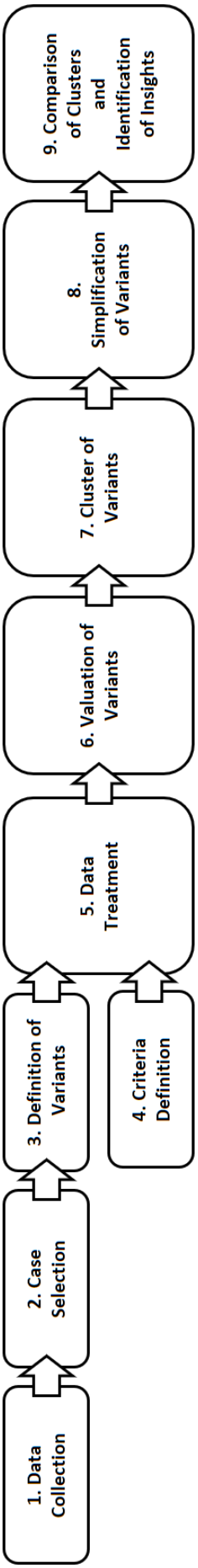


Figure 4.3: Overview of the concept of the multi-criteria analysis technique.

4.2.2
Step 2 - Case selection

The user must select which cases they want to analyze. This step is important to limit the scope of the analysis, like defining particular case characteristics (e.g. patients with the same disease severity), or cases that followed specific variants (e.g. cases that executed the activity F).

With respect to our example, the bank selected only complete and approved cases (that performed the activities A or B, and F) executed in the Netherlands for the period from 2016 to 2018.

4.2.3
Step 3 - Definition of variants

All cases from the event log are grouped by variant. That means that all cases that followed the same sequence of activities will be grouped.

With respect to the loan example, the process was executed in 20 different ways, and as consequence, 20 different groups of cases were created. For example, one group was composed of 1,500 cases that followed the path "A > C > D > E > F" (Variant 1), a second group was composed of 149 cases that followed the path "B > E > F" (Variant 2), a third group was composed of 700 cases that followed the path "B > C > E > F" (Variant 3) and, a fourth group was created with 10 cases that followed the path "A > C > C > C > D > E > F" (Variant 4).

4.2.4
Step 4 - Criteria definition

The user defines a set of criteria that they want to evaluate. A criterion is the combination of:

- an attribute from the process (e.g. case duration, age of the customer, costs, profit);
- a priority (e.g. little importance, extremely important);
- the target/direction (minimize or maximize);
- a type of measure (e.g. mean, variance).

In our example, the bank defined two simultaneous criteria: the first one is extremely important (*priority*) and consists of maximizing (*target*) the mean (*type of measure*) profit (*attribute*); the second one has its priority above average (that means, its priority is less important than the first criterion) and it consists of minimizing (*target*) the mean (*type of measure*) case duration (*attribute*).

4.2.5 Step 5 - Data treatment

All outliers of the selected attributes from Step 4 must be removed from all selected cases. An outlier is defined by Aggarwal [2015] as "a data point that is very different from most of the remaining data". Outliers in general are harmful for data analysis since they influence the mean, variance, minimal and maximum values [Seo, 2006]. In such manner, all values from each attribute selected during the criteria definition (Step 4) are evaluated for identifying outliers. All cases that contain outlier values are removed from the analysis.

Later, the user needs to inform the minimal number of cases they want to consider per variant. Normally, it is not recommended to consider a small number of cases per variant since they do not represent a frequent behavior. Variants with a number of cases (after removing cases with outliers) less than the informed by the user are removed from the analysis.

Finally, all values must be normalized to adjust different scales to a common scale. This is essential in the simultaneous comparison of different attributes (e.g. age and costs) [Commission et al., 2008]. In this way, all values from each attribute (selected in Step 4) are converted to a unique common scale (for example, a scale from 0 to 1).

Step 5 can only be executed after the execution of Steps 3 and 4, since the outliers will be analyzed and removed only for the attributes selected by the user in Step 4, and the normalization can only be applied for the remaining cases, i.e., cases selected in Step 2 without outliers.

Regarding the loan example, the number of cases from Variant 1 reduced to 1,480 since 20 cases had outliers values for case duration or profit. Variant 4 (with 10 cases) was removed from the analysis as the bank defined a minimal number of 50 cases per variant. All remaining attribute values were converted to a common scale from 0 to 1 (for example, a profit value of €3,645 was converted to 0.4586 and a case duration value of 12 days was converted to 0.1176).

4.2.6 Step 6 - Valuation of variants

A Unique Value (UV), or indicator, is calculated for each group of cases per variant (defined in Step 3). The UV can be the mean, the variance, or a combination of different metrics (composite indicator), as defined in the set of criteria created in Step 4. The UV is essential to rank the variants.

In our example, the following formula was used for calculating the UV:

$$\frac{4 \times mean(normalization(total\_time))}{8 \times mean(normalization(profit))} \quad (4\text{-}1)$$

The formula represents a minimization problem, and for this reason, the term from the "profit" criterion (that has a maximizing target) was added in the denominator part of the formula. The smaller the UV, the better outcomes the variant provides. The UV was 2.7249 for Variant 1, 0.0587 for Variant 2 and 0.1166 for Variant 3. These results show us that Variant 2 had the better outcomes, followed by Variants 3 and 1. A detailed description regarding the creation of the UV formula is present in section 5.1.6.

4.2.7 Step 7 - Cluster of variants

Different variants have different order in the execution of activities, but at the same time, some of them can provide similar outcomes. Taking the loan example, the best variant has 149 cases with a UV of 0.0587, and the second best variant has 700 cases with a UV of 0.1166. Both variants have a similar UV and the analysis of both variants together (with 849 cases) would provide more realistic insights than just evaluating the first variant with 149 cases. In Step 7, the variants are clustered using the UV created in Step 6, allowing the identification of variants with similar outcomes. This step allows for discovering the group of variants that provided positive outcomes in the process, and the group of variants that provided negative outcomes. The UV must be calculated for each cluster to allow ranking the clusters. This is necessary to identify the best and worst clusters.

In the loan process example, 2 clusters were created. The first one has all cases with the best outcomes, with a UV ranging from 0.0587 to 0.2274 (Cluster 1, UV = 0.1626) and the second one has all cases with worse outcomes, with an UV varying from 1.7057 to 3.8269 (Cluster 2, UV = 2.0927). Variants 2 and 3 were added in the first cluster.

4.2.8 Step 8 - Simplification of variants

In process mining analyses, the process simplification is a powerful technique for helping in the identification of insights. One way to simplify the process is to reduce the number of activities from the event log consolidating repeated sub-sequences. The aim of this step is to consolidate repeated sub-sequences. This consolidation can be applied in two cases:

1. When a given sub-sequence is present in all variants, it can be replaced by a new single activity. For example, suppose that the sub-sequence "A

> B > C" is identified in all variants, then the sub-sequence is replaced by one unique activity [A > B > C]. We call this created activity as *sub-sequence block*. Figure 4.4 presents an example of the creation of *sub-sequence blocks*. We can observe in the example a reduction of 2 activities after replacing the sub-sequences by blocks.

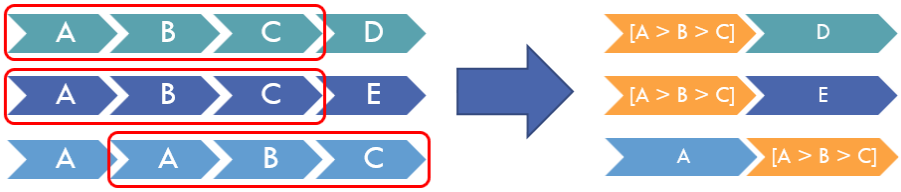


Figure 4.4: Example of simplification of variants: creation of *sub-sequence blocks*.

2. When a set of the same activities are part of different sub-sequences and the only difference is regarding their order (e.g. "A > B > C" or "B > C > A"), them all these sub-sequences can be replaced by one unique activity. For example, suppose that the sub-sequences "A > B > C", "A > C > B" or "B > C > A" are identified in all variants, then these sub-sequences are replaced by one unique activity [[A > B > C]]. We call this created activity as *permutation sub-sequence block*. Figure 4.5 presents an example of the creation of *permutation sub-sequence blocks*. We can observe in the example a reduction of 2 activities after replacing the sub-sequences by blocks.

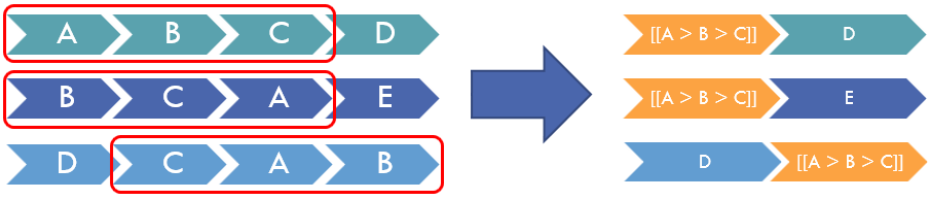


Figure 4.5: Example of simplification of variants: creation of *permutation sub-sequence blocks*.

Concerning the loan example, the sub-sequence "E > F" was substituted by the *sub-sequence block* [E > F] since this sub-sequence is present in all remaining variants of the event log.

4.2.9 Step 9 - Comparison of clusters and identification of insights

The final step of our method is to analyze the clusters (created in Step 7) and present to the user the set of activities and sub-sequences that can optimize/improve their process according to the criteria defined in Step 4. Here, we denominate the Positive cluster as the one that has the set of variants that provided positive outcomes, and we denominate the Negative cluster as the one with worse outcomes than the Positive cluster. The Positive and Negative clusters can be composed of one or more clusters identified in Step 7.

In this step, first, the variants from the Positive and the Negative clusters are compared. This comparison puts in evidence the activities and sub-sequences that contributed to positive outcomes, as well as, that contributed to negative outcomes. Activities and sub-sequences that are only presented in one of the two clusters will be identified and highlighted.

Secondly, a matrix with the minimum distances for every pair of variants is created. The distance consists of counting the minimum number of operations (changes of activities or order of activities in a variant) needed to transform one variant into another. The minimum distance between variants was inspired in the minimum edit distances commonly used to compare two strings [Navarro, 2001]. The analysis of the minimum distance together with the variants UV helps the user to identify the activities and sub-sequences that mostly contributes to improve (or diminish) the process outcomes. For example, suppose two variants, the first one presents the sub-sequence "A > B > C > D" with a UV of 0.0021 and the second one has the sub-sequence "A > C > D" with a UV of 4.543. The distance of these two variants is 1 since the first variant executed an extra activity (B) that was not executed by the second variant. Analyzing this scenario, we can observe that with one modification in the process (adding activity B) it is possible to increase the outcomes significantly. To sum up, the matrix with the minimum distances supports the user to identify pairs of variants with a small distance and that at the same time has a significant difference in their UV, and, analyzing these pairs of variants, the user can identify activities and sub-sequences that mostly contribute to improve their process.

Regarding our example, Cluster 1 represents the Positive cluster and Cluster 2 represents the Negative cluster. Performing Step 9 the bank gets results like: 1. the execution of activity "B. Request Loan in Person", and the execution of the sub-sequences "B > C"[1] and "B > [E > F]"[2] contributed to

[1] "B > C" means: activity B directly followed by activity C.
[2] [E > F] is a *sub-sequence block*.

better outcomes; 2. the execution of activity "A. Request Loan Online", and the execution of the sub-sequences "A > C" and "C > C" contributed to worse outcomes. The minimum distance matrix would indicate to the bank that the worst variant (A > E > F, with UV 3.8269) has only one difference (distance of 1) to the best variant (B > E > F, with UV 0.0587). That means that the substitution of activity A (Request Loan Online) to B (Request Loan in Person) probably would improve highly the outcomes. With these results in hand, we could provide to the bank some optimization recommendations like: 1. encourage the client to request loan in person (execution of activity B); 2. improve the content of their website, providing more information for their clients.

4.3 Contributions of the approach

Our approach promotes significant contributions since it:

- identifies and highlights a set of activities and sub-sequences that, in the past, provided positive or negative outcomes. The approach helps the user to include positive behaviors and to remove negative behaviors from their normative process. The technique evaluates the benefits of each variant (according to their outputs and a set of defined criteria) considering the complete event log. Some cases with negative outcomes can perform the same variant from cases with positive outcomes. The solution from Dees et al. [2017] considers only deviations that promoted positive outcomes to adjust the normative model. Lakshmanan et al. [2013] highlights frequent patterns of activities and sub-sequences, but according to our understanding, their technique performs the analysis considering either patterns with positive outcomes or patterns with negative outcomes (not the influence of all patterns simultaneously);
- is designed to work with multiple simultaneous criteria (KPIs) considering different importance levels defined by the user. For example, the user can define a criteria to minimize costs and another to maximize profits. None of the analyzed approaches allows working with multi-criteria in a generalized and flexible way, in which the user define as input a set of criteria;
- generates outputs considering a group of different variants instead of a unique variant. The good execution of a process can happen in many different ways and it is not beneficial to adjust a process considering solely the best executed variant. Possibly, the best variant was performed in

few occasions. Using a group of variants increases the number of analyzed cases and, as a consequence, the set of generated recommendations considers more behaviors;

- provides results that allow the user to re-evaluate constraints (rules) predefined in the normative process. Sometimes constraints of a process can contribute to negative results. Dees et al. [2017] adjust a model respecting its rules, and as a consequence, removes the opportunity for the user to evaluate the constraints. Nevertheless, our approach can perform analyses considering constraints if the user provides a filtered event log including only cases that are compliant to these rules;
- allows the user to work with a sub-group of variants of a given process (sub-graph). The user needs to provide a new event log filtering specific sub-sequences using one of the available process mining tools;
- helps the user to identify the set of activities and sub-sequences that mostly contributes to improve (or diminish) the process outcomes (using the minimum distance matrix);
- can, as a side effect, provide to the user some deviations in the execution of the process. For example, in the loan use case, the sub-sequence "activity C directly followed by activity C" contributed to worse outcomes. The self arc for activity C is not present in the normative process, and thus, the sub-sequence "C > C" is a deviation.

To conclude, the implementation of the presented technique has potential to help organizations to customize and improve their processes taking into consideration their own needs, reality and past actions. In the next chapter we will present technical details regarding the implementation of this approach.

5
Multi-CAT: Multi-criteria analysis tool

In this chapter, we will describe the implementation and usage of the process mining technique we presented in the previous chapter. We implemented the tool as a ProM [TU/E - Math&CS Department, 2018b] plugin. ProM is an extensible framework maintained by the Eindhoven University of Technology that supports the highest number of process mining plugins [Dramski, 2017]. The framework was developed using the Java technology [Oracle Corporation, 2018], and as a consequence, can run in any operating system. ProM is one of the most popular process mining tool [Claes and Poels, 2012] and is open source. We call the developed plugin as "Multi-Criteria Analysis Tool" (Multi-CAT).

First, we will describe how we implemented each step from the multi-criteria analysis technique concept. Later we will present the Multi-CAT usage process.

5.1
Tool implementation

In this section we will describe how we implemented in Multi-CAT each step from the concept we presented in Chapter 4.

5.1.1
Step 1 - Data collection

The user is responsible for executing this step. There is a plethora of existing processes and there can be many different data-sources (systems) to support a given process. For example, different companies can execute the same "product quotation" process, but they can use different Enterprise Resource Planning (ERP) systems.

For each process and context, the user needs to identify and extract the associated data to be evaluated and convert it to an event log.

After defining the tables and attributes, the user may consider developing a tool to automate the extraction process and the creation of the event log. The user can use the plugin "Convert CSV to XES" from ProM to convert the event log to a XES format.

5.1.2 Step 2 - Case selection

This step is also performed by the user. The user can adopt one of the many existing tools to filter an event log. ProM has "Filter log by attributes", "Filter log on trace attribute values" and "Filter events based on attribute values" plugins that can help the user in the case selection. Disco [Fluxicon, 2018] has some robust filters, like the "Follower", that allows the user to select cases that followed specific sub-sequences.

5.1.3 Step 3 - Definition of variants

In this step, Multi-CAT groups all cases from the event log by variant.

Before grouping cases per variant, Multi-CAT sorts the event log by date and time of the execution of activities to guarantee that the cases will have the activities correctly ordered. If the date and time of two or more activities of a same case are equal, than the tool will order the activities in alphabetical order.

5.1.4 Step 4 - Criteria definition

In this step, the user must inform all criteria for the analysis they want to perform. Multi-CAT is designed to support one or more criteria. Each criterion is composed of an attribute from the event log to be analyzed (e.g. case duration, costs, profits), the importance of the criterion to the user (little importance, under average, above average, extremely important), the direction/target of the criterion (minimize or maximize), and the type of measure (mean, variance, standard deviation).

Concerning the importance of each criterion, we did not find in the literature a formal method for defining weights that gives flexibility in the characterization of the importance distance between different criteria and at the same time that is simple for user usage. Stillwell et al. [1981] presents different methods to convert criteria rank order into weights: 1. the rank sum; 2. the rank exponent; 3. the rank reciprocal; 4. the rank-order centroid. All these four approaches are simple to be used since the user only needs to define a criteria rank, but these approaches do not allow the user to define the importance distance of 2 criteria (e.g. criterion 1 is much more important than criterion 2). The Analytic Hierarchy Process (AHP) [Saaty, 1987] is a method for organizing and analyzing complex decisions considering multi-criteria. The technique decomposes a decision problem into a hierarchy set of sub-problems.

The first step of AHP, that defines the priority weights for a set of criteria, allows the definition of importance distance between different criteria but it is not a simple process as the user needs to compare all criteria in pairs (e.g. considering three criteria, the user should compare criterion 1 with 2, criterion 1 with 3 and criterion 2 with 3). Thus, we implemented in Multi-CAT a scale composed of four importance weights (items) as presented in table 5.1 [On Target, 2018]. The scale uses a set of exponential values and, as a consequence, each score doubles the previous score value. As the scale is composed of only four items, we believe it is simple to be used and can reduce human indecision.

Table 5.1: Scale of criterion importance implemented in Multi-CAT. Source: [On Target, 2018]

Description	Weight Value
Little importance	1
Under average	2
Above average	4
Extremely important	8

5.1.5 Step 5 - Data treatment

In this step, Multi-CAT removes outliers, excludes variant groups that contain a number of cases less than the specified by the user, and normalizes the data attributes. We implemented two methods to remove outliers [Seo, 2006]:

- 5% extremes (standard deviation method): attribute values are classified as outliers if they are out of the interval from the inferior bounder *mean(attribute) - 2*standard_deviation(attribute)* and the superior bounder *mean(attribute) + 2*standard_deviation(attribute)*;
- Interquartile rule (Turkey's method): attribute values are classified as outliers if they are out of the interval from the inferior bounder *first quartile - (IQR[1] *1.5).* and the superior bounder *third quartile + (IQR[1] *1.5).*

We implemented two methods to normalize data [Commission et al., 2008]:

- Min-Max: *(x - min)/(max - min)*, where x is the current case attribute value, min is the minimal value from all the attribute values (considering all variants), and max is the maximal value;

[1]IQR (InterQuartile Range) = third quartile - first quartile

– Ranking: *Rank(x)*. Attribute values are sorted and their position (rank) is the normalized value. If two or more attributes have the same value, then the average of their position is calculated and attributed to all of them.

We selected both normalization methods as they generate positive values and this is a prerequisite to create the UV (see more information in subsection 5.1.6). Multi-CAT allows the user to choose not to apply any outliers removal method and/or normalization method.

5.1.6 Step 6 - Valuation of variants

For each variant, Multi-CAT calculates a unique value (UV) that is essential to compare the performance between different variants. To calculate the UV per variant, Multi-CAT automatically creates a UV formula that groups all criteria defined by the user.

In this step first, the user must select the aggregation method to connect all defined criteria. The aggregation method can be additive or multiplicative. The selection of the aggregation type is associated in how the user wants to compensate criteria with poor outcomes with criteria with good outcomes [Profit et al., 2010]. The additive aggregation type provides full compensability and may generate a UV with error in direction and dimension. In the other hand, the multiplicative aggregation provides partial compensability, reducing the probability of one criterion to neutralize another. In general, the additive aggregation type is the most used since its simplicity.

Then, the tool creates the terms of the formula that represents each criterion. Each term is composed by a measure value multiplied by the importance defined by the user. Here the importance acts as a weight in the term. The measure value can be the mean, variance or standard deviation.

By default, Multi-CAT creates an equation term that represents a "minimization" problem. In this way, all minimization terms will compose the numerator part of the formula, and all maximization terms will constitute the denominator part. All minimization and maximization terms will be connected considering the type of aggregation selected by the user.

If the aggregation type is "Additive", Multi-CAT will sum all criteria as presented below, in which min represents all minimization terms and max represents all maximization terms:

$$\frac{\sum(weight_{min} * measure_{min}(attribute_{min}))}{\sum(weight_{max} * measure_{max}(attribute_{max}))} \quad (5\text{-}1)$$

If the aggregation type is "Multiplicative", Multi-CAT will multiply all criteria as presented below:

$$\frac{\prod(weight_{min} * measure_{min}(attribute_{min}))}{\prod(weight_{max} * measure_{max}(attribute_{max}))} \quad (5\text{-}2)$$

As an example of UV formula creation, suppose the user defined the four criteria presented in table 5.2 and selected the aggregation type "Additive".

Table 5.2: Example of four criteria to explain the creation of the UV formula.

ID	Priority	Attribute	Target	Type of Measure
1	extremely important	profit	maximize	mean
2	above average	requested_money	maximize	mean
3	under average	case_duration	minimize	mean
4	little importance	client_age	minimize	variance

First, Multi-CAT will create the following four terms:

- Term for criteria 1: *8 * mean(normalization(profit))*
- Term for criteria 2: *4 * mean(normalization(requested_money))*
- Term for criteria 3: *2 * mean(normalization(case_duration))*
- Term for criteria 4: *1 * variance(normalization(client_age))*

Then, Multi-CAT will connect all minimization and maximization terms using the "Additive" type of aggregation. Finally, Multi-CAT will set the minimizing terms as the numerator part of the formula and all maximizing terms as the denominator part of the formula. The resulting formula for the previous example is:

$$\frac{(2 \times mean(normalization(case\_duration))) + (1 \times variance(normalization(client\_age)))}{(8 \times mean(normalization(profit))) + (4 \times mean(normalization(requested\_money)))} \quad (5\text{-}3)$$

For each variant, Multi-CAT will calculate the UV using the created formula. The lower the UV, the better the variant contributes to improve the outcomes based on the selected criteria.

The UV formula assumes that the resulting value of each term of the formula (after normalizing and applying the type of selected measure) is a positive real number $\Re_{>0} = \{\chi \in \Re | \chi > 0\}$.

5.1.7 Step 7 - Cluster of variants

Multi-CAT clusters the variants based on the UV. We selected the following centroid-based clustering methods (or partitioning methods) from the Java Machine Learning Library (Java-ML) [Java-ML, 2012], as they work with numerical data [Gan et al., 2007] and our clustering problem has only one dimension (UV) with no need for more sophisticated clustering methods (like high density or spectral clustering):

- K-means [MacQueen et al., 1967]: is one of the most popular clustering algorithms. K-means clusters data points by minimizing the sum-of-squared-error criterion. In this way, it iterates until the sum-of-squared-error criterion does not reduces significantly. During each iteration, the method re-calculates the mean values of each cluster (centroids) and re-allocate the data points to the closest centroids. As the method uses the mean value, it is sensitive to outliers. The method is efficient in clustering large data sets;
- K-medoids [Kaufman and Rousseeuw, 1987]: is a similar method to K-means but instead of using the mean, the method selects the most centrally located data point in a cluster (medoids) as a centroid. This method is more robust than K-Means in the presence of outliers, but it has a higher computational cost than K-Means;
- Farthest First [Gonzalez, 1985; Sanjoy Dasgupta, 2002]: this method places a new centroid in the furthest place from the previous defined cluster centroids. After defining all centroids, the method will allocate the data points to the closest centroid. This technique provides a well-spaced clusters and is suitable for large data sets. The method may not generate uniform clusters [Kumar et al., 2013].

All selected clustering methods need as input parameter the number of clusters (K value). To simplify the tool usage, Multi-CAT automatically defines the appropriate number of clusters using the Elbow Method. The method verifies if adding a new cluster will improve significantly the clustering results based in the variance analysis (from K to K-1) of the sum of squared errors (distance between data points in a cluster) [Han et al., 2011]. We implemented the Elbow Method using the Euclidean Distance. Multi-CAT performs the elbow analysis from 2 to 7 clusters.

Multi-CAT also allows the user to manually select the number of clusters (from 2 to 7), giving them the flexibility to perform the analysis according to their specific needs.

To allow reproducibility for user analyses, we updated all cluster methods to always use the same seed in the first definition of centroids. This guarantees the creation of the same clusters (number of clusters and variants associated to them) given a context (event log and user parameters).

After clustering the variants by UV, Multi-CAT calculates the UV for each cluster, in the same way as presented in Step 6, but instead of considering only the cases of a variant, it considers all cases from a cluster. In this way, Multi-CAT will apply the same UV formula created in Step 6 for all cases from each cluster.

5.1.8 Step 8 - Simplification of variants

In this step, Multi-CAT will simplify the variants consolidating repeated sub-sequences that are present in all variants of the event log. Below we present the algorithms we implemented for the creation of blocks.

Sub-sequence blocks

This algorithm will replace ordered sub-sequences that are present in all variants by *sub-sequence blocks*. For example, if the sub-sequence "A > B > C" is present in all variants, then Multi-CAT will replace this sub-sequence for a new activity identified as [A > B > C]. The algorithm will prioritize the replacement of larger and more frequent sub-sequences. Steps of the algorithm:

1. Look for the smallest (regarding the number of activities) variant from the event log;
2. Create a list with all possible sub-sequences (combination of activities) for the smallest variant. For example, for the variant "A > B > C" the sub-sequences "A > B > C", "A > B" and "B > C" will be created. Working on the smallest variant increases the tool performance as it reduces the number of activities to create the sub-sequences;
3. Remove from the list all sub-sequences that are not present in all variants;
4. For each sub-sequence from the list, calculate its frequency and size;
5. Sort the sub-sequence list by size and frequency (descending order). This will allow to first replace larger and more frequent sub-sequences to blocks;
6. Iterate over the sub-sequence list and, for each variant, replace each sub-sequence by a *sub-sequence block*. If the sub-sequence is "A > B > C"

then the *sub-sequence block* will be identified as [A > B > C]. This step item will be executed until there are no more available sub-sequences to be replaced.

Permutation sub-sequence blocks

This algorithm will replace sub-sequences (considering the permutation of the activities) that are present in all variants by *permutation sub-sequence blocks*. For example, if the sub-sequences "A > B > C" or "B > C > A" are present in all variants, then Multi-CAT will replace these sub-sequences for a new activity identified as [[A > B > C]]. The algorithm, like the previous one, will prioritize the replacement of larger and more frequent sub-sequences. Steps of the algorithm:

1. Look for the smallest (regarding the number of activities) variant from the event log;

2. Create a list with all possible sub-sequences (combination of activities) for the smallest variant. For example, for the variant "A > B > C" the sub-sequences "A > B > C", "A > B" and "B > C" will be created. In this step item Multi-CAT does not consider *"sub-sequence blocks"* created by the previous algorithm. In the case a variant contains *"sub-sequence blocks"*, then all sub-sequences between the blocks will be used to create the list;

3. For each sub-sequence from the list (created in the previous step item), create all possible permutations sub-sequences (permuting the activities). For example, for the sub-sequence "A > B", the permutations "A > B" and "B > A" will be created;

4. Remove from the list all sub-sequences that are not present in all variants (taking into consideration the permuted sub-sequences);

5. For each sub-sequence from the list, calculate its frequency and size (taking into consideration the permuted sub-sequences);

6. Sort the sub-sequence list by size and frequency (descending order). This will allow to first replace larger and more frequent sub-sequences to blocks;

7. Iterate over the sub-sequence list (taking into consideration the permuted sub-sequences) and, for each variant, replace each sub-sequence by a *permutation sub-sequence block*. For example, if the sub-sequence is "A > B > C" or "C > B > A" then the *permutation sub-sequence block* will

be identified as [[A > B > C]]. This step item will be executed until there are no more available sub-sequences to be replaced.

The user has the option to select how they want to simplify the event log: 1. Creation of *sub-sequence blocks*; 2. Creation of *sub-sequence blocks* and *permutation sub-sequence blocks*; 3. No creation of blocks. In case the user selects the second option ("Creation of *sub-sequence blocks* and *permutation sub-sequence blocks*"), Multi-CAT first will create the *sub-sequence blocks* and later will create the *permutation sub-sequence blocks* since the last algorithm has a high computational cost.

5.1.9 Step 9 - Comparison of clusters and identification of insights

We implemented the comparison of two different clusters (comparison of activities and sub-sequences). The Positive cluster is the one with the best UV and the Negative cluster is the one with the worst UV.

To make easier to understand the comparison of clusters, we will use an example of two clusters detailed in Table 5.3.

Table 5.3: Example of 2 clusters.

Cluster	Variant	Number of Cases
Positive	A >B >C >D	20
Positive	A >C	20
Positive	A >B >D	10
Negative	A >C >B	20
Negative	A >C >D	5
Negative	A >B >D >E	10

Comparison of activities

For comparing activities from both clusters, Multi-CAT identifies the list of activities that are present in the Negative cluster and that are not present in the Positive cluster, and vice versa. Regarding the example presented above, Multi-CAT identifies that activity E is present in the Negative cluster and it is not present in the Positive cluster.

Comparison of sub-sequences for directly followed activities

For comparing the executed sub-sequences of directly followed activities from both clusters, Multi-CAT will identify the directly followed sub-sequences that are present in the Negative cluster and that are not present in the Positive cluster, and vice versa.

With respect to the example presented above, Multi-CAT identifies that the sub-sequence "B > C" contributed positively to the process as it is only present in the Positive cluster, and that the sub-sequences "C > B" and "D > E" contributed negatively to the process as they are only present in the Negative cluster.

The algorithm implemented for the comparison of sub-sequences for directly followed activities is detailed in Appendix F.

Minimum distance for every pair of variants

We used the Levenshtein distance algorithm to implement the matrix with the minimum distances for every pair of variants. The Levenshtein distance algorithm calculates the number of differences between two words (or strings) [Navarro, 2001]. The algorithm calculates the minimum number of operations necessary to transform one word into the other (in its simplified definition, each operation has cost 1). The operations can be: insert, delete or replace. For example, the Levenshtein distance for the words *source* and *force* is 2, since there is one deletion operation (removal of the letter u) and one replacement operation (replacement of letter s by f). In Multi-CAT, we used the Levenshtein distance method available in the Apache Commons Text [Apache Commons, 2018].

5.2 Extra functionalities

To help the user in their analyses, we implemented three extra functionalities in Multi-CAT:

1. Last user input parameters are recovered: every time the user perform an analysis in Multi-CAT, all input parameters are saved. In case the user performs new analyses using the same event log, Multi-CAT recovers the last parameters used;

2. Label of activities: the user can define how they want to label the process activities. The first option is *alias* in which an acronym is created using the original name of the activity (e.g. the activity name "Collection of Blood Culture" is converted to "CoBC"). The second option is *number* in which each activity of the process is converted to a number (e.g. the activity name "Collection of Blood Culture" is converted to "9"). The last option is *number and alias* in which both previous options are combined (e.g. the activity name "Collection of Blood Culture" is converted to "9. CoBC");

3. Global Unique Value: Multi-CAT calculates the UV for the complete analysis considering the remaining cases after the outliers and infrequent variants removal. This Global UV could be used in the comparison of the performance of different organizations (e.g. to compare the results of a group of hospitals regarding the execution of the sepsis CP), or to verify the performance of an organization in different periods (for example, to check the evolution of the process every semester).

5.3 Tool usage

In this section we present how Multi-CAT works. Figure 5.1 presents an overview of the plugin usage. We divided the tool utilization process in 3 main activities that are detailed below.

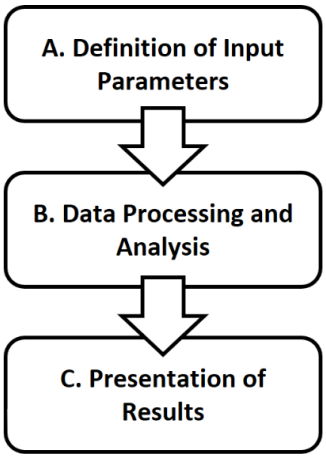


Figure 5.1: Overview of Multi-CAT usage.

5.3.1 Activity A. Definition of input parameters

In this step the user informs all necessary parameters for the analysis. This activity implements step 4, and part of steps 5 and 6 from Section 5.1. For performing this activity, the user must:

1. Load the event log in ProM. The user can provide a CSV or a XES file;

2. Select the event log and run the plugin (see Figure 5.2). ProM will start Multi-CAT presenting its input screen (see Figure 5.3);

3. Define the list of criteria, the method to remove outliers, the method to normalize the data, the type of aggregation (how to combine the

set of criteria), the minimal number of cases per variant, the variant clustering method, how to present the name of activities (alias, number, number and alias) and, how to simplify the variants (creation of blocks). For each criterion, the user must select the attribute (loaded from the event log), the weight (little importance, under average, above average, extremely important), the direction (minimize or maximize), the type of measure (mean, variance, standard deviation) and if the method to remove outliers should be applied for the selected attribute. Appendix G shows all fields options from the input screen.

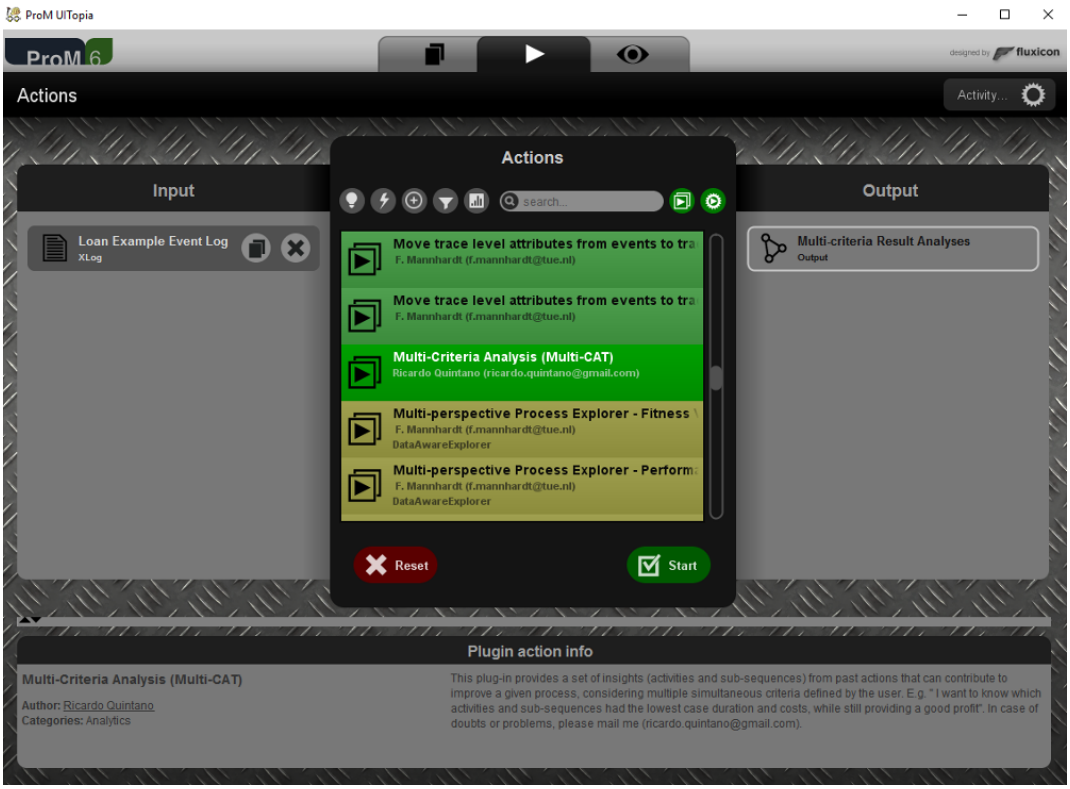


Figure 5.2: Selection of Multi-CAT plugin in ProM.

5.3.2 Activity B. Data processing and analysis

Once the user provided all input parameters, Multi-CAT will process the event log executing steps 3, 5, 6, 7, 8 and 9 described in Section 5.1.

If during the execution of this activity any problem is identified (e.g. duplicated criteria, invalid attribute value) then Multi-CAT will provide a list of errors to the user. Otherwise, Multi-CAT will present the results (described in the next activity).

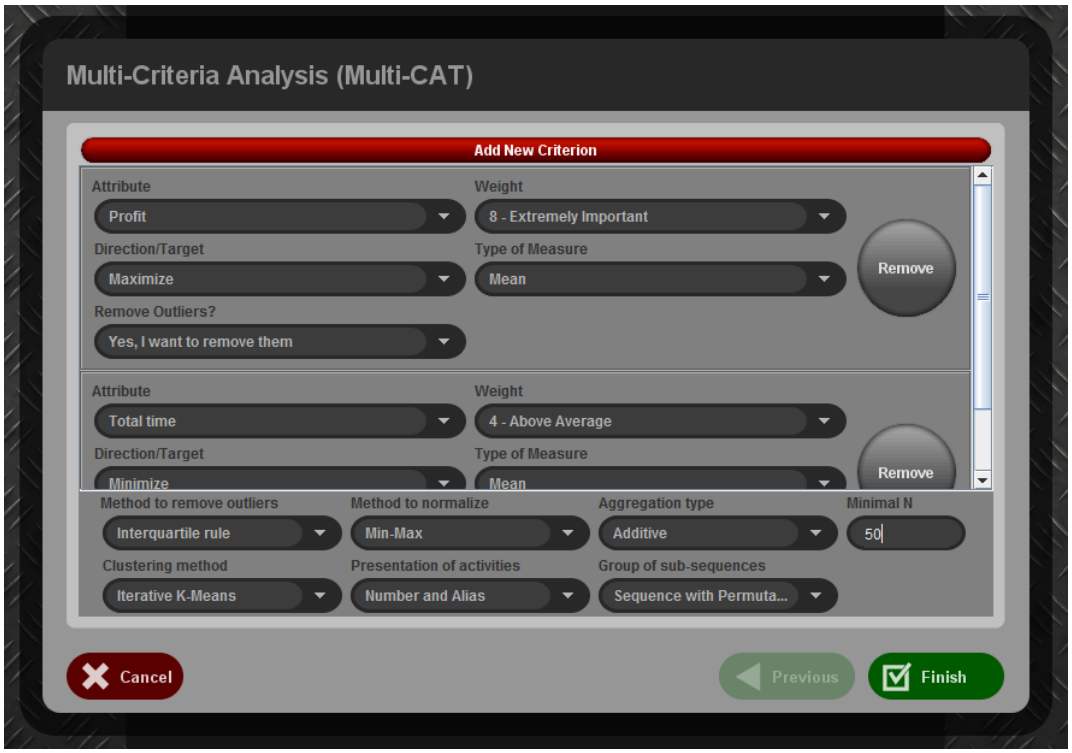


Figure 5.3: Multi-CAT input screen.

5.3.3 Activity C. Presentation of results

The results of the analysis are presented to the user. Figure 5.4 presents the overview of the result report screen from Multi-CAT. We divided the figure in three parts (figures 5.5, 5.6, 5.7) for better visualization. The report is composed of the following items:

- The created UV formula (see item 1 in figure 5.5);
- The Global UV (see item 2 in figure 5.5);
- All criteria defined in Activity A (see item 3 in figure 5.5). The figure presents criteria C1 and C2 with their configuration details (attribute, weight, direction/target and type of measure);
- All parameters specified in Activity A (see item 4 in figure 5.5). The figure shows the method to remove outliers, the method to normalize values, the type of aggregation used in the UV formula, the minimal number of elements (cases) per variant, the clustering method, and how the sub-sequences were simplified (group of sub-sequences);
- A table with all variants (we denominate it as *variants table*) ordered from the best to the worst UV (observe item 5 in figure 5.6). Multi-CAT shows for each variant its sub-sequence, the number of cases (N), the

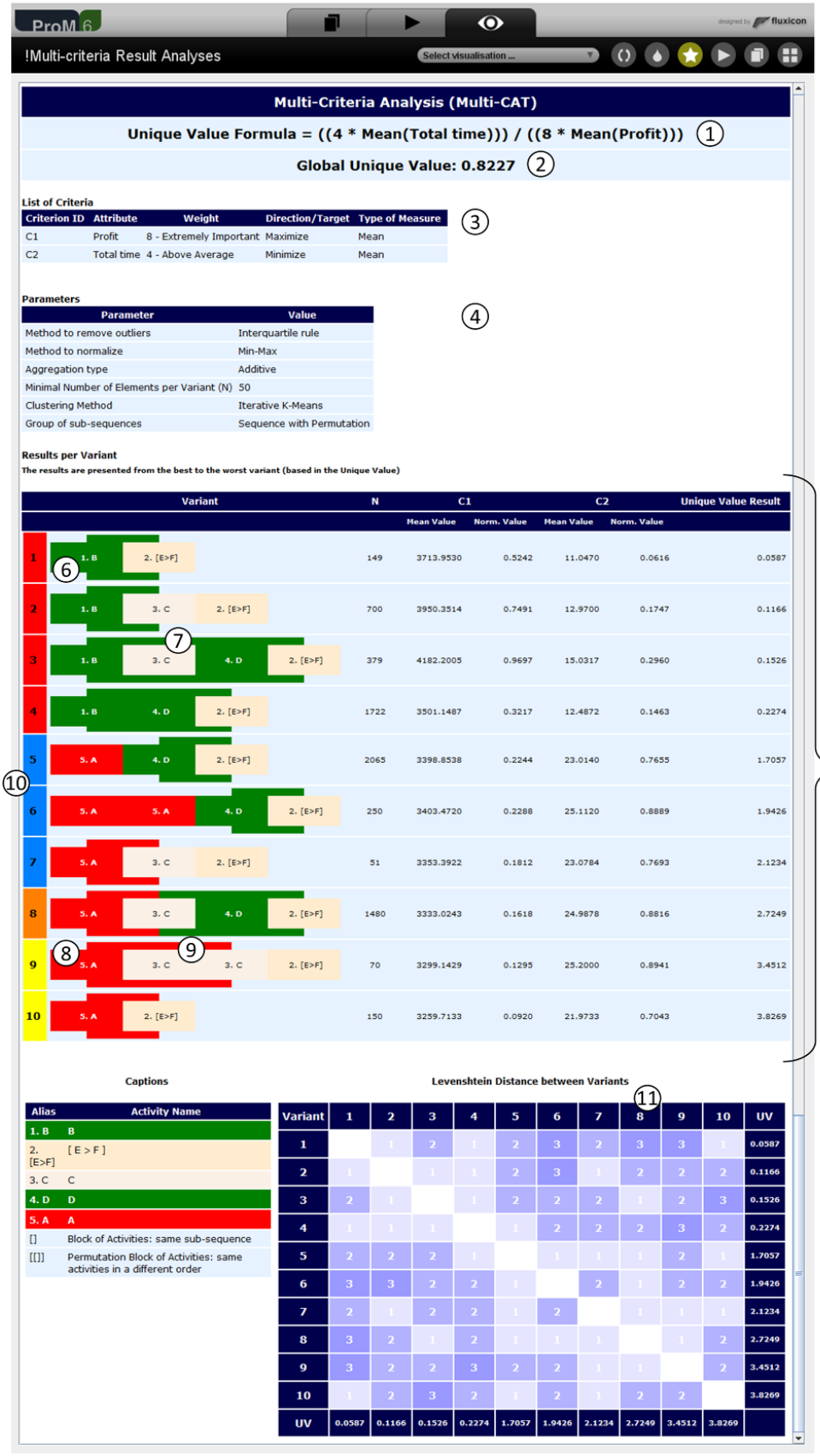


Figure 5.4: Overview of the result report screen from Multi-CAT.

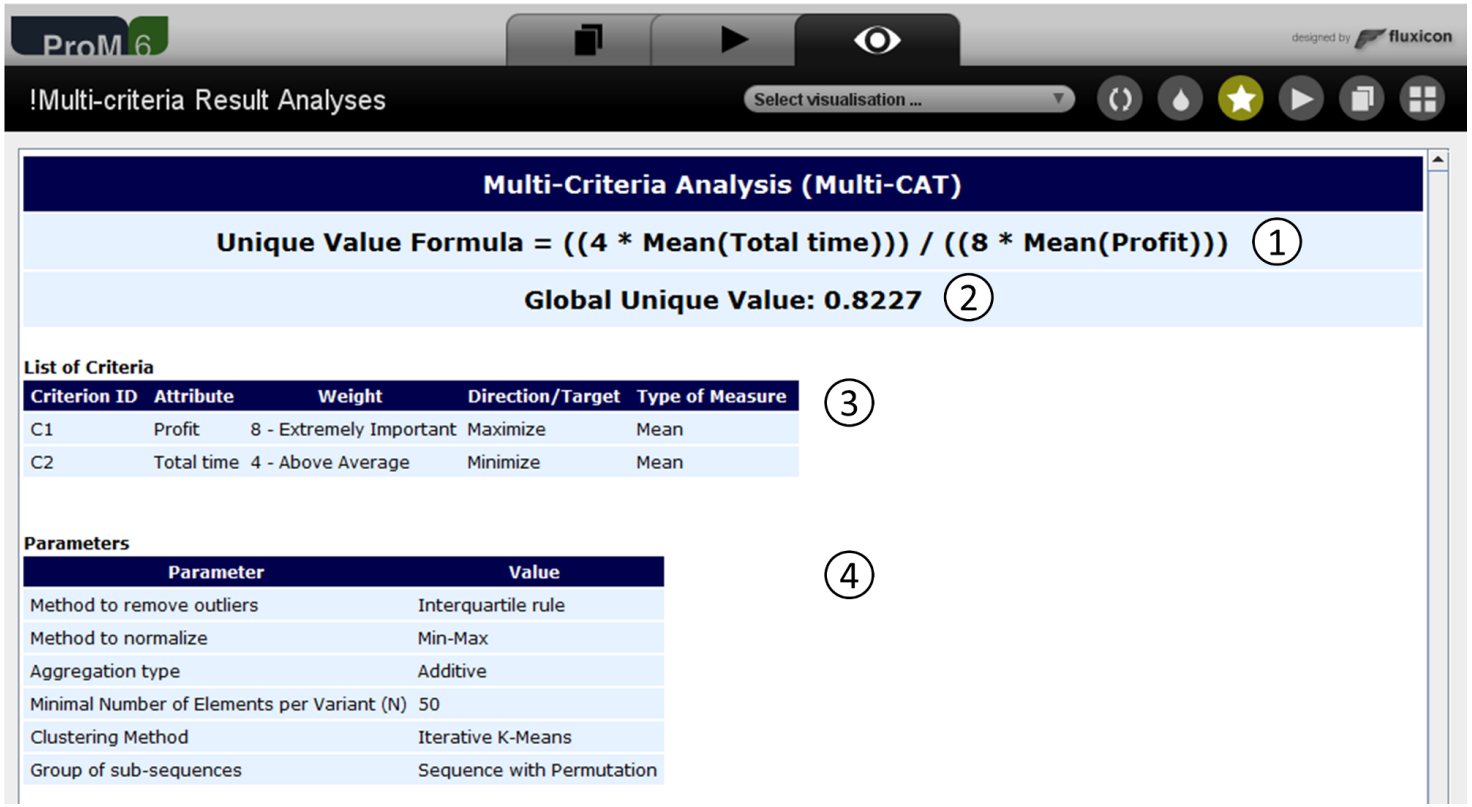


Figure 5.5: Result report screen from Multi-CAT: UV formula, Global UV and parameters of the analysis.

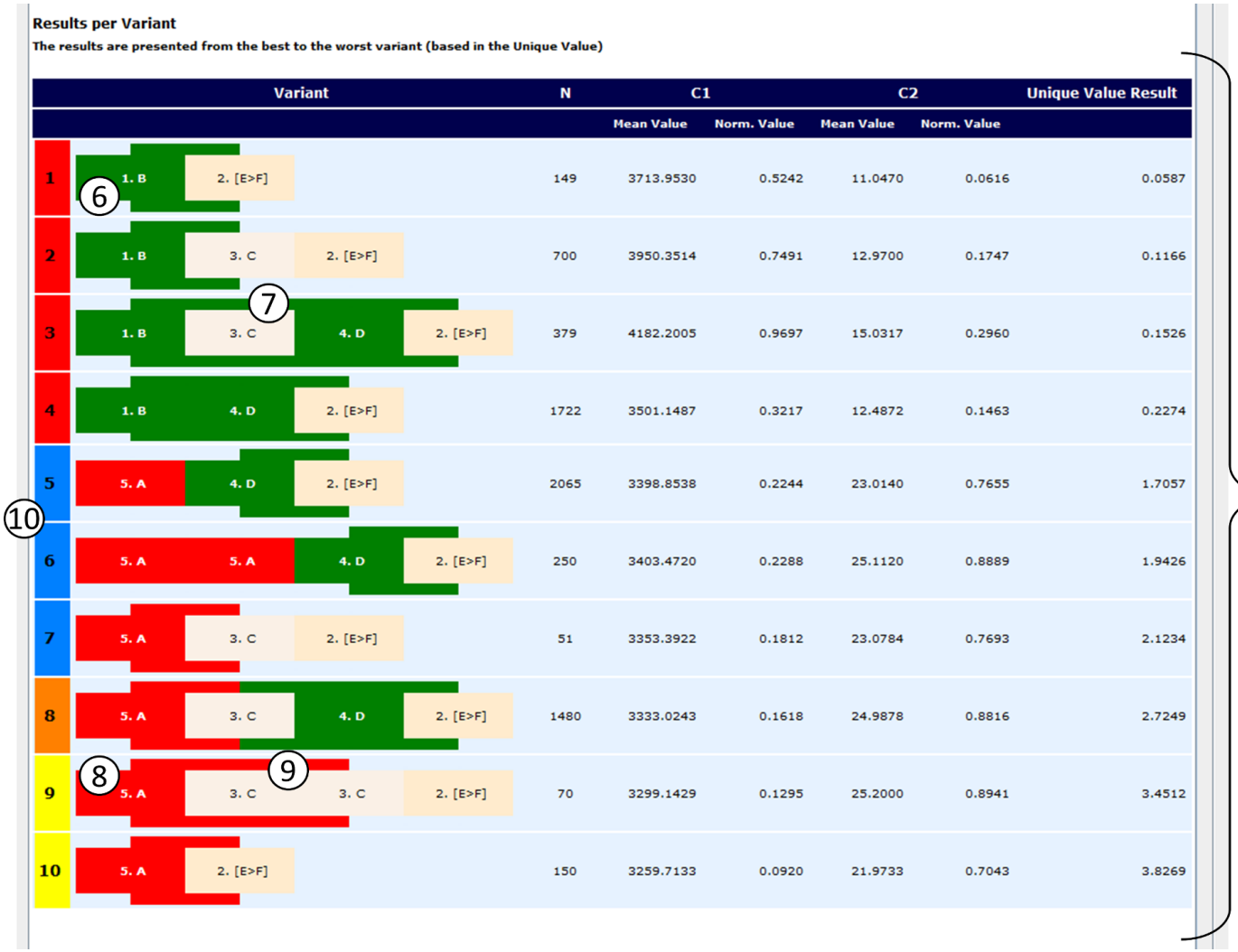
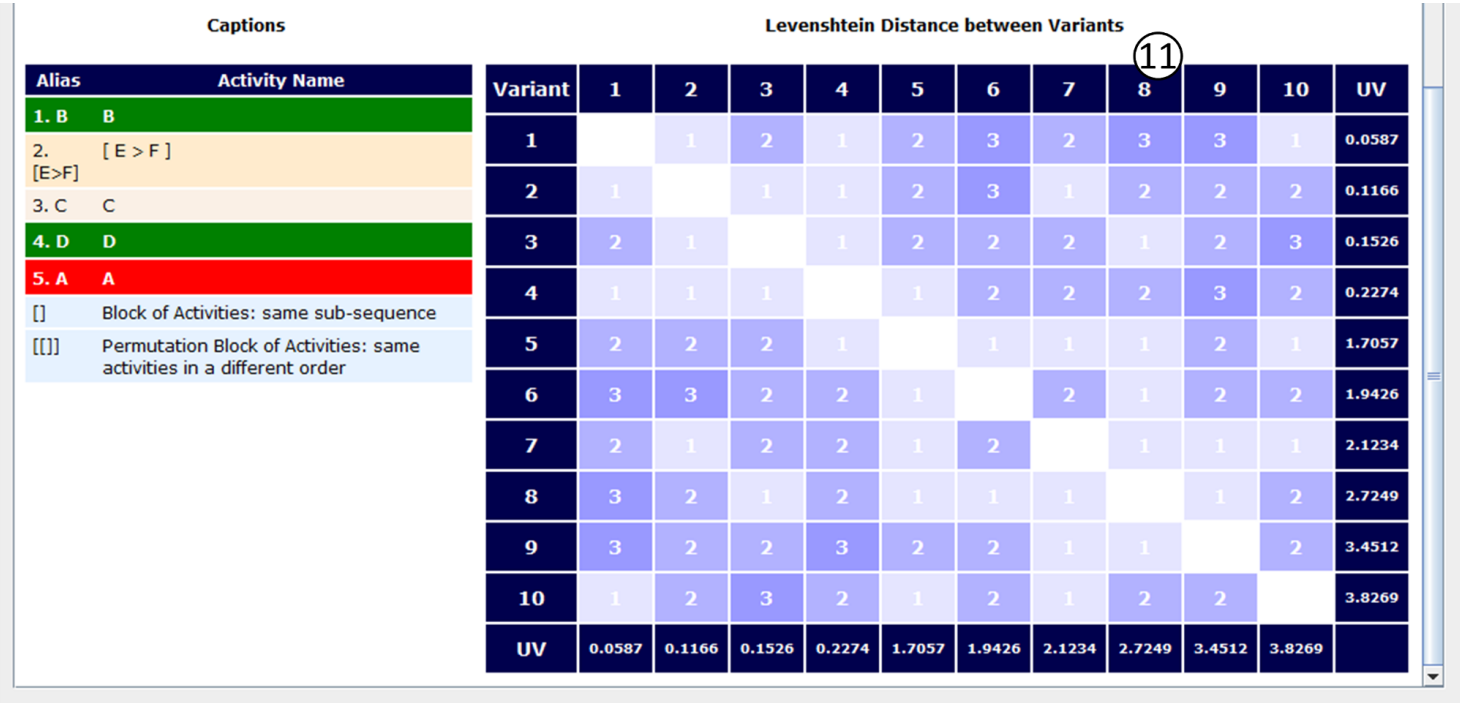


Figure 5.6: Result report screen from Multi-CAT: variants table.



Captions

Alias	Activity Name
1. B	B
2. [E>F]	[E > F]
3. C	C
4. D	D
5. A	A
[]	Block of Activities: same sub-sequence
[[]]	Permutation Block of Activities: same activities in a different order

Levenshtein Distance between Variants

Variant	1	2	3	4	5	6	7	8	9	10	UV
1		1	2	1	2	3	2	3	3	1	0.0587
2	1		1	1	2	3	1	2	2	2	0.1166
3	2	1		1	2	2	2	1	2	3	0.1526
4	1	1	1		1	2	2	2	3	2	0.2274
5	2	2	2	1		1	1	1	2	1	1.7057
6	3	3	2	2	1		2	1	2	2	1.9426
7	2	1	2	2	1	2		1	1	1	2.1234
8	3	2	1	2	1	1	1		1	2	2.7249
9	3	2	2	3	2	2	1	1		2	3.4512
10	1	2	3	2	1	2	1	2	2		3.8269
UV	0.0587	0.1166	0.1526	0.2274	1.7057	1.9426	2.1234	2.7249	3.4512	3.8269	

Figure 5.7: Result report screen from Multi-CAT: the Levenshtein distance between variants matrix and captions of activities.

type of measure (mean, variance, standard deviation) and normalized value for each criterion, and its associated UV value. For example, figure 5.6 shows for Variant 2: 1. the sub-sequence is "1.B > 3.C > 2.[E>F]" ([E>F] is a *sub-sequence block*); 2. its number of cases is 700; 3. criterion C1 (profit) has a mean value of 3,950.3514, and a normalized value of 0.7491; 4. criterion C2 (total time) has a mean value of 12.9700, and a normalized value of 0.1747; 5. the UV is 0.1166. We can note that Variant 1 "1.B > 2. [E>F]" provided the best outcomes since it is the first one in the *variants table* (UV 0.0587), and that Variant 10 "1.A > 2. [E>F]" provided the worst outcomes since it is the last one (UV 3.8269);

– Positive activities are highlighted in green (observe item 6 in figure 5.6) in the *variants table*. The figure shows that activities B and D promoted better outcomes in the process;

– In the *variants table*, positive sub-sequences are highlighted as green bars located in top and bottom of pairs of activities (see item 7 in figure 5.6). In the figure, we can observe that the sub-sequences "B > [E>F]", "B > C", "C > D", "D > [E>F]" and "B > D" contributed to better results;

– Negative activities are highlighted in red (see item 8 in figure 5.6) in the *variants table*. We can note in the figure that activity A promoted worse outcomes to the process;

– In the *variants table*, negative sub-sequences are highlighted as red bars located in top and bottom of pairs of activities (see item 9 in figure 5.6). In figure 5.6, we can observe that the sub-sequences "A > C", "C > C" and "A > [E>F]" contributed to worse outcomes;

– Identification of created clusters (observe item 10 in figure 5.6). Variants with the same label color are members of the same cluster. We can notice in figure 5.6 that Variants 1, 2, 3 and 4 are members of the red cluster and that Variants 9 and 10 are members of the yellow cluster;

– The Levenshtein distance between variants matrix (see item 11 in figure 5.7). The number presented in each cell is the distance from one variant (line) to another one (column). The cell colors change according to its distance value, from light blue (small distance) to dark blue (large distance). The matrix also shows the UV of each variant in its last column and line. The matrix aids the user in identifying pairs of variants with a small distance and that at the same time have a significant difference in their UV, and, analyzing these pairs of variants, the user can identify activities and sub-sequences that mostly contributes to improve their process. We can observe that the distance from Variant 10 (worst

variant) to 1 (best variant) is 1 indicating that changing activity A by B can contribute significantly to improve the process (the UV varied from 3.8269 to 0.0587).

5.4 Final considerations

In this chapter, we presented the development details and usage of Multi-CAT. We described the implementation of each of the steps of our approach presented in section 4.2 and we characterized the sequence of activities to use the tool.

In the next chapter, we will evaluate the tool in two use cases. The first one is the loan request example we presented in Section 4.1, and the second one is the real sepsis clinical pathway presented in Chapter 3. In addition, we will verify Multi-CAT performance applying the tool to nine different event logs, with different process complexities and different number of cases.

6
Validation of Multi-CAT

In this chapter we present the set of tests we executed to evaluate Multi-CAT. We first introduce the evaluation of the tool using two event logs from two different processes: 1. the loan request example (presented in Chapter 4); 2. the real process execution of a sepsis clinical pathways (presented in Chapter 3). Later we check the performance of the tool in distinct contexts.

The execution of these set of tests helped us to validate that the Multi-criteria analysis technique and tool (Multi-CAT) work as expected (providing optimization insights) and to identify their contributions and limitations. We summed up our conclusions regarding our evaluation at the end of this chapter.

6.1
First test scenario: loan request

In this section, we present our tests of Multi-CAT using the loan request process described in Chapter 4. First, we will show the process and event log description. Second, we will present the test parameters, following the Multi-CAT results. Finally, we will show some process recommendations based on the results.

6.1.1
Process and event log description

In this test we used the loan request process. Its model and description can be found in section 4.1.

The event log presents only cases that started the process with activities A or B, and finished it executing activity F. All cases that were rejected (executed activity G) or canceled (executed activity H) were removed from the event log as the bank wants to identify recommendations to reduce the case duration and increase the profit for approved cases. The resulting event log is composed of 8,049 cases, with 35,255 events, 20 variants and 6 activities (A, B, C, D, E, and F).

6.1.2 Test parameters

In this test scenario, we want to get optimization insights to reduce the case duration and increase the profit for approved cases. Table 6.1 details the complete list of parameters used in this test.

Table 6.1: Parameters for the first test scenario: loan request.

Parameter	Value
Criterion 1	
Attribute	Profit
Weight	8 - Extremely important
Direction	Maximize
Type of Measure	Mean
Criterion 2	
Attribute	Total time (time from the start of the process until the execution of the current activity)
Weight	4 - Above average
Direction	Minimize
Type of Measure	Mean
Other Parameters	
Method to remove outliers	Interquartile rule
Method to normalize	Min-Max
Aggregation Type	Additive
Minimal N	50
Cluster Method	K-means - 2 clusters
Presentation of Activities	Alias
Group of sub-sequences	Sub-sequence and Permutation Sub-sequence blocks

6.1.3 Results

Figure 6.1 presents the Multi-CAT report of this test scenario. Analyzing the report together with the normative model (see Figure 4.2) we got the following conclusions:

1. Activity B, highlighted in green in the *variants table* of Figure 6.1, contributed to better outcomes. We can note that the activity is present in the best variants (from 1 to 4);

2. Activity A, which is highlighted in red in the *variants table*, contributed to worse outcomes. We can observe that the activity is not present in the best variants (from 1 to 4);

Multi-Criteria Analysis (Multi-CAT)

Unique Value Formula = ((4 * Mean(Total time))) / ((8 * Mean(Profit)))

Global Unique Value: 0.8227

List of Criteria

Criterion ID	Attribute	Weight	Direction/Target	Type of Measure
C1	Profit	8 - Extremely Important	Maximize	Mean
C2	Total time	4 - Above Average	Minimize	Mean

Parameters

Parameter	Value
Method to remove outliers	Interquartile rule
Method to normalize	Min-Max
Aggregation type	Additive
Minimal Number of Elements per Variant (N)	50
Clustering Method	K-Means - 2 Clusters
Group of sub-sequences	Sequence with Permutation

Results per Variant

The results are presented from the best to the worst variant (based in the Unique Value)

	Variant	N	C1		C2		Unique Value Result
			Mean Value	Norm. Value	Mean Value	Norm. Value	
1	B [E>F]	149	3713.9530	0.5242	11.0470	0.0616	0.0587
2	B C [E>F]	700	3950.3514	0.7491	12.9700	0.1747	0.1166
3	B C D [E>F]	379	4182.2005	0.9697	15.0317	0.2960	0.1526
4	B D [E>F]	1722	3501.1487	0.3217	12.4872	0.1463	0.2274
5	A D [E>F]	2065	3398.8538	0.2244	23.0140	0.7655	1.7057
6	A A D [E>F]	250	3403.4720	0.2288	25.1120	0.8889	1.9426
7	A C [E>F]	51	3353.3922	0.1812	23.0784	0.7693	2.1234
8	A C D [E>F]	1480	3333.0243	0.1618	24.9878	0.8816	2.7249
9	A C C [E>F]	70	3299.1429	0.1295	25.2000	0.8941	3.4512
10	A [E>F]	150	3259.7133	0.0920	21.9733	0.7043	3.8269

Captions

Alias	Activity Name
B	B
[E>F]	[E > F]
C	C
D	D
A	A
[]	Block of Activities: same sub-sequence
[[]]	Permutation Block of Activities: same activities in a different order

Levenshtein Distance between Variants

Variant	1	2	3	4	5	6	7	8	9	10	UV
1		1	2	1	2	3	2	3	3	1	0.0587
2	1		1	1	2	3	1	2	2	2	0.1166
3	2	1		1	2	2	2	1	2	3	0.1526
4	1	1	1		1	2	2	2	3	2	0.2274
5	2	2	2	1		1	1	1	2	1	1.7057
6	3	3	2	2	1		2	1	2	2	1.9426
7	2	1	2	2	1	2		1	1	1	2.1234
8	3	2	1	2	1	1	1		1	2	2.7249
9	3	2	2	3	2	2	1	1		2	3.4512
10	1	2	3	2	1	2	1	2	2		3.8269
UV	0.0587	0.1166	0.1526	0.2274	1.7057	1.9426	2.1234	2.7249	3.4512	3.8269	

Figure 6.1: Multi-CAT report for the first test scenario: loan request.

3. Sub-sequences "B > [E>F]"[1], "B > C" and "B > D" contributed to better outcomes. All these sub-sequences are highlighted with green bars in the *variants table* of Figure 6.1 and they are present only in the best variants (from 1 to 4);

4. Sub-sequences "A > C", "A > A", "C > C", "A > [E > F]" and "A > D" contributed to worse outcomes. All these sub-sequences are surrounded with red bars in the *variants table* and they are present in the variants that did not generate the best outcomes;

5. In the *Levenshtein distance between variants* matrix, we can observe that the distance from variant 10 (worst variant) to 1 (best variant) is 1 indicating that changing activity A by B can contribute significantly to improve the process (the UV decreased from 3.8269 to 0.0587);

6.1.4 Process optimization recommendations

Based on the analysis of the results from Multi-CAT and our knowledge about the process, we can provide the following recommendations to the bank to improve their process:

1. Encourage the client to request loan in person (execution of activity B). Probably the bank attendant provides clear and complete information regarding the process to the client. In addition, the client has the opportunity to get their questions answered at one time. This recommendation was identified based on the analysis of items 1 and 2 from section 6.1.3;

2. Improve the content of the website (add more information or highlight important information). Adding an online chat can also help, in this way, the client can get their questions answered during the loan request (activity A). This recommendation was identified based on the analysis of items 1 and 2 from section 6.1.3;

3. If the client requests the loan online (execution of activity A), the bank could call them as soon as possible to review all instructions and answer possible questions (inclusion of a new activity in the process). This recommendation was identified based on the analysis of items 1 and 2 from section 6.1.3;

[1] "B > [E>F]" means "Activity B directly followed by *sub-sequence block* [E>F]".

4. Analyze and identify the reasons for clients not sending their documents on time (based on the sub-sequence "C > C" from variant 9, that means that the bank called the client 2 times to request their documents). Is there any action that the bank could do to help the client for collecting all necessary documents? This recommendation was identified based on the analysis of item 4 from section 6.1.3;

5. Verify if there is room to improve the quality of the call from activity C (for variant 9, the activity C happened 2 times). This recommendation was identified based on the analysis of item 4 from section 6.1.3;

6. Identify the reasons for the frequent non-occurrence of activities C and D (as observed in variants 1, 2, 4, 5, 6, 7, 9 and 10). Is there a problem in the system for not registering correctly the activities? Were the activities not executed?;

7. Identify the reasons for activity A (in variant 6) be executed 2 times. Is there a problem in their website?

6.2 Second test scenario: sepsis clinical pathway

In this section, we demonstrate our Multi-CAT tests employing the sepsis clinical pathway process described in Chapter 3. First, we will describe the process and event log. Second, we will present the test parameters, following the tool results. To conclude, we will show some process recommendations based on Multi-CAT results.

6.2.1 Process and event log description

In this test we used the sepsis treatment process from the emergency department. The normative process description (clinical pathway) is presented in section 3.2.1.

For this test, we updated the event log splitting the activity "Start Formal Sepsis Pathway" in two. If the activity was registered by a nurse (as a result of the evaluation during the triage) its name was changed to "Start Formal Sepsis Pathway A". If the activity was registered by a physician (as a result of the medical evaluation) its name was changed to "Start Formal Sepsis Pathway B".

The resulting event log is composed of 1,710 cases, with 20,605 events, 235 variants and 15 activities.

6.2.2 Test parameters

In this test scenario, we want to get optimization insights to reduce the time to give antibiotics since the faster the patient receives the antibiotics, the greater is their survival probability. Table 6.2 details the complete list of parameters used in this test.

Table 6.2: Parameters for the second test scenario: sepsis clinical pathways.

Parameter	Value
Criterion 1	
Attribute	Total time antibiotic (time from the start of the process until the time to give the antibiotics)
Weight	8 - Extremely important
Direction	Minimize
Type of Measure	Mean
Other Parameters	
Method to remove outliers	5% Extremes
Method to normalize	Min-Max
Aggregation Type	Additive
Minimal N	30
Cluster Method	Iterative Farthest First
Presentation of Activities	Number and Alias
Group of sub-sequences	Sub-sequence and Permutation Sub-sequence blocks

6.2.3 Results

Figure 6.2 presents the Multi-CAT report for this test scenario. Analyzing the report together with the normative model (see Figure 3.1) we got the following conclusions:

1. Activity "3. Start Formal Sepsis Pathway A" (3. SFSPA), highlighted in green in the *variants table* of Figure 6.2, contributed to better outcomes. That means that the identification of a sepsis suspicion patient and the start of the sepsis pathway should be preferably done sooner, during triage. We can note that this activity is mostly seen in the variants with better outcomes (1, 2, 4, 5 and 6);

2. Activity "8. Start Formal Sepsis Pathway B" (8. SFSPB), which is highlighted in red in the *variants table*, provided worse results. Meaning that the identification of a sepsis suspicion patient and the start of the sepsis pathway should be preferably not done during the medical

Multi-Criteria Analysis (Multi-CAT)

Unique Value Formula = (8 * Mean(total_time_antibiotic))

Global Unique Value: 3.2851

List of Criteria

Criterion ID	Attribute	Weight	Direction/Target	Type of Measure
C1	total_time_antibiotic	8 - Extremely Important	Minimize	Mean

Parameters

Parameter	Value
Method to remove outliers	5% extremes
Method to normalize	Min-Max
Aggregation type	Additive
Minimal Number of Elements per Variant (N)	30
Clustering Method	Iterative Farthest First - SE
Group of sub-sequences	Sequence with Permutation

Results per Variant

The results are presented from the best to the worst variant (based in the Unique Value)

	Variant	N	C1		Unique Value Result
			Mean Value	Norm. Value	
1	1. [GQN>RoP] 2. FMoVS **3. SFSPA** 4. RoT 5. [[BCR>LR]] 6. [PoA>BCC>L... 7. RoCN	33	30.5152	0.3309	2.6471
2	1. [GQN>RoP] 2. FMoVS **3. SFSPA** 4. RoT 5. [[BCR>LR]] 6. [PoA>BCC>L... 7. RoCN	96	31.3125	0.3438	2.7500
3	1. [GQN>RoP] 2. FMoVS 4. RoT 5. [[BCR>LR]] 6. [PoA>BCC>L... 7. RoCN **8. SFSPB**	84	32.8690	0.3689	2.9508
4	1. [GQN>RoP] 2. FMoVS **3. SFSPA** 4. RoT 7. RoCN 5. [[BCR>LR]] 6. [PoA>BCC>L...	102	33.9216	0.3858	3.0867
5	1. [GQN>RoP] **3. SFSPA** 2. FMoVS 4. RoT 7. RoCN 5. [[BCR>LR]] 6. [PoA>BCC>L...	54	34.5370	0.3958	3.1661
6	1. [GQN>RoP] 2. FMoVS **3. SFSPA** 4. RoT 7. RoCN 5. [[BCR>LR]] 6. [PoA>BCC>L...	344	36.1453	0.4217	3.3736
7	1. [GQN>RoP] 2. FMoVS 4. RoT 7. RoCN **8. SFSPB** 5. [[BCR>LR]] 6. [PoA>BCC>L...	63	37.2540	0.4396	3.5166
8	1. [GQN>RoP] 2. FMoVS 4. RoT 7. RoCN **8. SFSPB** 5. [[BCR>LR]] 6. [PoA>BCC>L...	338	37.3432	0.4410	3.5282

Captions

Alias	Activity Name
1. [GQN>RoP]	[Get Queue Number > Registry of Patient]
2. FMoVS	First Measure of Vital Signs
3. SFSPA	**Start Formal Sepsis Pathway A**
4. RoT	Registry of Triage
5. [[BCR>LR]]	[[Blood Culture Request > Lactate Request]]
6. [PoA>BCC>L...	[Prescription of Antibiotic > Blood Culture Collection > Lactate Collection > Administration of Antibiotic]
7. RoCN	Registry of Clinical Notes
8. SFSPB	**Start Formal Sepsis Pathway B**
[]	Block of Activities: same sub-sequence
[[]]	Permutation Block of Activities: same activities in a different order

Levenshtein Distance between Variants

Variant	1	2	3	4	5	6	7	8	UV
1			2	2	4	2	4	4	2.6471
2			2	2	4	2	4	4	2.7500
3	2	2		4	4	4	4	4	2.9508
4	2	2	4		2		2	2	3.0867
5	4	4	4	2		2	2	2	3.1661
6	2	2	4		2		2	2	3.3736
7	4	4	4	2	2	2			3.5166
8	4	4	4	2	2	2			3.5282
UV	2.6471	2.7500	2.9508	3.0867	3.1661	3.3736	3.5166	3.5282	

Figure 6.2: Multi-CAT report for the second test scenario: sepsis clinical pathways.

evaluation or later. We can observe that this activity is not present in the two best variants (1 and 2);

3. The sub-sequence "2. First Measure of Vital Signs > 3. Start Formal Sepsis Pathway A > 4. Registry of Triage" ("2. FMoVS > 3. SFSPA >

4. RoT"), highlighted with a top and bottom green bars in the *variants table* of Figure 6.2, subsidized better outcomes. That means that it is a good approach to identify a sepsis suspicion patient and start the sepsis pathway during triage. In general, this behavior is mostly seen in the variants with better outcomes (1, 2, 4 and 6);

4. The sub-sequence "4. Registry of Triage > 5. [[Blood Culture Request > Lactate Request]]"[2] ("4. RoT > 5. [[BCR>LR]]"), surrounded by green bars, contributed to better results. That means that the request of blood culture and lactate exams should be done immediately after the registry of triage. We can note that this sub-sequence is only presented in the variants with best outcomes (1, 2 and 3);

5. The sub-sequence "6. [Prescription of Antibiotic > Blood Culture Collection > Lactate Collection > Administration of Antibiotic] > 7. Registry of Clinical Notes" ("6. [PoA>BCC>L... > 7. RoCN"), highlighted with green bars, provided better outcomes. So, the registry of clinical notes should be executed after the prescription and administration of the treatment. This behavior is only seen in the variants with better outcomes (1, 2, and 3);

6. The sub-sequence "2. First Measure of Vital Signs > 4. Registry of Triage" ("2. FMoVS > 4. RoT"), surrounded by red bars in the *variants table* of Figure 6.2, contributed to worse outcomes since the identification of a sepsis suspicion patient and the start of the sepsis pathway did not happen during triage;

7. The sub-sequence "7. Registry of Clinical Notes > 8. Start Formal Sepsis Pathway B > 5. [[Blood Culture Request>Lactate Request]]" ("7. RoCN > 8. SFSPB > 5. [[BCR>LR]]"), highlighted with red bars, provided worse results as the physician prescribed the treatment after registering the clinical notes (activity 7 eventually followed by 5). We can observe that this behavior is not presented in the three best variants (1, 2 and 3);

8. Analyzing the Levenshtein distance matrix, we can identify that moving the registry of clinical notes to after the prescription and administration of medicines could improve significantly the process (comparison of variant 6 with variant 1, with a distance of 2 and a difference of 5.63 minutes in the time to give the antibiotics).

[2] "[[Blood Culture Request > Lactate Request]]" means *permutation sub-sequence block* of activities "Blood Culture Request" and "Lactate Request".

We can note in Figure 6.2 that variants 1 and 2 have exactly the same sub-sequence of activities. This behavior happened since the only difference between both variants is the order of the activities "Blood Culture Request" and "Lactate Request", and in our analysis we considered the creation of *permutation sub-sequence blocks*, generating consequently the activity "5.[[BCR>LR]]". For Variant 1 the order of both activities is "Lactate Request > Blood Culture Request", and for Variant 2 is the inverse order. The same behavior happened with variants 7 and 8. Figure 6.3, shows the *variants table* without selecting the creation of *permutation sub-sequence blocks*. "Blood Culture Request" and "Lactate Request" activities are highlighted with a black box in variants 1, 2, 7 and 8.

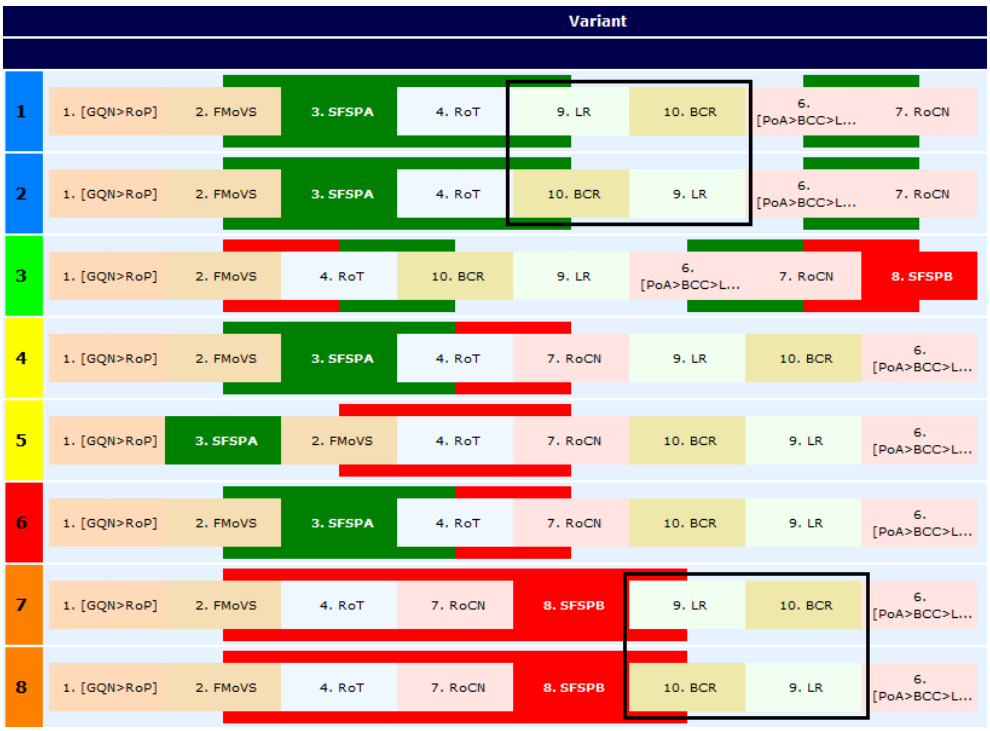


Figure 6.3: *Variants table* of the sepsis clinical pathways test scenario without applying the creation of *permutation sub-sequence blocks*.

6.2.4 Process optimization recommendations

With the results from Multi-CAT and our knowledge about the process, we could identify the following recommendations to the hospital improve their sepsis clinical pathway:

1. To preferably start the formal sepsis pathway during triage. The hospital could provide additional training to nurses to better identify a patient

with sepsis suspicion. Another option could be the deployment of a system that, based on the patient's vital signals collected during triage, could alert nurses to start the formal sepsis pathway. This recommendation was identified based on the analysis of items 1, 2, 3 and 6 from section 6.2.3;

2. Prescribe the sepsis treatment before registering the clinical notes. If the physician prescribes earlier the treatment, the pharmacy process to deliver the medication will start earlier. As a consequence, the patient can receive the sepsis treatment in a shorter time. This recommendation was identified based on the analysis of items 4, 5 and 7 from section 6.2.3. This recommendation is the same we identified in our previous manual analysis (see Chapter 3).

Figure 6.4 presents the Clinical Pathway (normative model) updated considering the insights gain from Multi-CAT. The main updates are: A.the activity "5. Start Sepsis Pathway (A)" should be performed during triage; B. the activity "6. Registry of Clinical Notes" should be executed after the prescription of the treatment.

6.2.5 Global UV evaluation over time

To check the evolution of the hospital regarding its outcomes over time, we divided the event log described in subsection 6.2.1 in four semesters. We executed Multi-CAT for each event log using the same parameters described in Table 6.2. For each analysis we collected its resulting Global UV. Table 6.3 presents the number of cases and the Global UV per semester.

Table 6.3: Global UV evaluation over time for the sepsis CP use case.

Semester	Number of Cases	Global UV
1	401	3.0238
2	338	3.2157
3	603	3.1294
4	368	3.4519

Analyzing Table 6.3, we can note that the hospital had a slight increase in its Global UV during the 2 years period, varying from 3.0238 to 3.4519. These results indicate that the hospital decreased its outcomes and should reevaluate the execution of its process.

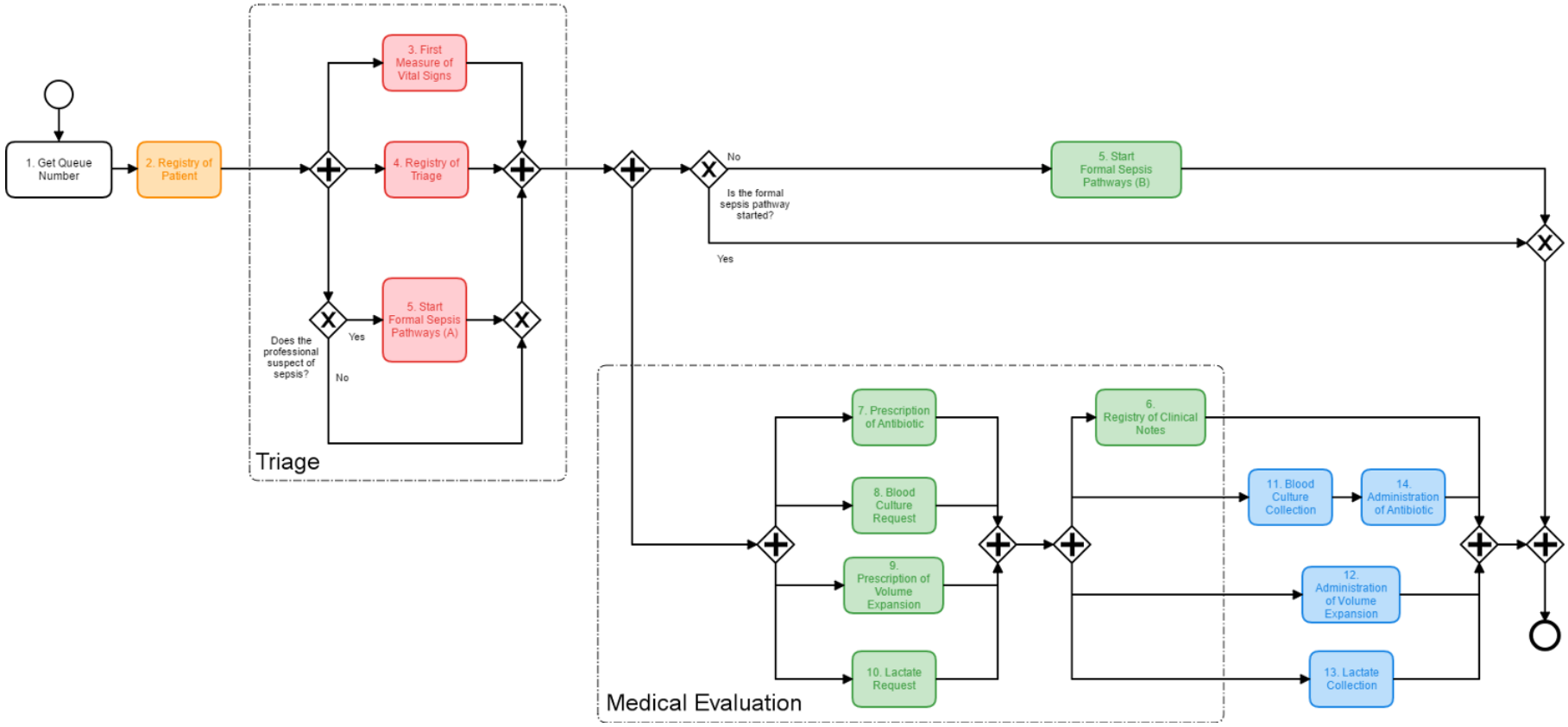


Figure 6.4: Updated sepsis Clinical Pathway (normative model) considering the insights from Multi-CAT. The orange activity is performed by a receptionist; pink activities are executed by nurses; green activities are performed by physicians; blue activities are executed by nurse technicians.

6.3 Time performance tests

With respect to the performance tests, our aim was to verify how much time the tool takes in distinct contexts, like the number of activities per variant or the number of cases in the event log.

6.3.1 Processes and event logs description

We generated three random processes using the Process Log Generator tool (PLG) [Burattin, 2016, 2017]. Each process had a different complexity (low, medium, high). The complexity was determined according to the number of activities and gateways. The gateways could be parallel or exclusive choice. All processes had three variables (A, B and C) with numeric values varying from 1,000 to 50,000. Table 6.4 presents the details of each generated process. For each process we produced three different event logs with a different number of cases (1,000, 10,000 and 50,000). We added noise in the generation of the event logs to make them as real as possible. The possible types of noise were to shift the order of activities, to duplicate activities, to remove some activities or to generate unexpected values for the three variables. Appendix H shows the parameters used for the event logs generation. To sum up, we created nine different event logs, each of them with a different complexity and number of cases. Table 6.5 presents the details of each created event log like the number of events, number of variants and number of activities for the largest variant.

Table 6.4: Characteristics of each process created for the performance tests.

Process Complexity	Number of Activities	Number of Gateways
Low	7	4
Medium	20	12
High	58	34

6.3.2 Test parameters

For each of the nine event logs we created, we ran Multi-CAT using always the same parameters. We selected the parameters options that could demand a high performance complexity like for example the iterative K-Medoids clustering method or the creation of *sub-sequence* and *permutation sub-sequence blocks*. Table 6.6 shows the Multi-CAT parameters used in all performance tests. All tests were executed using an Intel® Core™ i5-4300M

Table 6.5: Details of the nine event logs created for the performance tests.

Event log ID	Process Complexity	Number of Cases	Number of Activities	Number of events	Number of Variants	Number of Activities in the Largest Variant
A	Low	1,000	7	3,981	20	6
B	Low	10,000	7	39,839	49	6
C	Low	50,000	7	199,202	58	6
D	Medium	1,000	20	10,866	50	18
E	Medium	10,000	20	109,552	198	19
F	Medium	50,000	20	545,316	528	19
G	High	1,000	58	36,751	705	71
H	High	10,000	58	365,274	4,225	71
I	High	50,000	58	1,825,588	14,451	81

processor (@ 2.60Hz) with 8.00Gb of primary memory (RAM - Random Access Memory), and Microsoft Windows 10 operating system.

Table 6.6: Multi-CAT parameters used in all performance tests.

Parameter	Value
Criterion 1	
Attribute	Variable A
Weight	1 - Little Importance
Direction	Minimize
Type of Measure	Mean
Criterion 2	
Attribute	Variable B
Weight	2 - Above Average
Direction	Minimize
Type of Measure	Variance
Criterion 3	
Attribute	Variable C
Weight	8 - Extremely Important
Direction	Maximize
Type of Measure	Mean
Other Parameters	
Method to remove outliers	Interquartile rule
Method to normalize	Min-Max
Aggregation Type	Additive
Minimal N	20
Cluster Method	Iterative K-Medoids
Presentation of Activities	Number and Alias
Group of sub-sequences	Sub-sequence and Permutation Sub-sequence blocks

6.3.3 Results

Table 6.7 shows the results from the performance tests. Each table line shows the results from each event log. In the table, we present the time in milliseconds (ms) of each executed step of our approach (the same steps presented in Section 4.2), the total processing time (considering all data processing and data presentation), and the ProM time to present the user interface. Table 6.8 shows the smallest variant size and number of variants per event log after removing outliers and removing infrequent variants (post-processing). A detailed report of the performance tests can be found in Appendix I.

Analyzing the results from both tables we can note:

1. Without considering the time to create *permutation sub-sequence blocks*, the more complex a process (number of activities and gateways) was, the more time it took to Multi-CAT to generate the results. In our tests, the complexity was correlated to the number of variants: the more complex the process was, the more variants the event log had. The number of variants can affect the time to perform some processing steps. For example, during the clustering step each variant is a data point, or during the Levenshtein Distance matrix creation the number of pair comparisons is correlated to the number of variants;

2. Without considering the time to create *permutation sub-sequence blocks*, the more cases an event log had, the more time it took to Multi-CAT provide the analysis results. The number of cases can affect the time to execute some processing steps, like for example, grouping cases into variants, normalization of values and valuation of variants. In our tests, the number of cases was also associated to the number of variants: the more cases an event log had, the more variants it had;

3. The creation of *permutation sub-sequence blocks* was in most tests the step that consumed the most part of the total time (about 63%). The time to create these blocks is associated to the size of the smallest variant. The higher is the size of the smallest variant, the more time it will take to create all possible sub-sequences based on permutations of activities. Even though scenario G had a more complex process than D, G had a lower permutation process time than D because G created previously 2 *sub-sequence blocks*, reducing the size of activities to permute to a maximum of 8. The same happened to scenarios H and E. H had a

Table 6.7: Results from the performance tests.

Event log ID	Time per Step (ms)						Total processing time[a] (ms)	Total processing time without permutation[b] (ms)	ProM Interface Processing Time[c] (ms)
	3	**5**	**6**	**7**	**8**	**9**			
A	15	16	0	2	0	0	88	88	13
B	100	38	0	5	0	0	205	205	19
C	615	178	22	45	0	0	997	997	155
D	18	16	0	18	46,448	8	46,571	126	135
E	406	21	3	16	40,193	27	40,805	613	271
F	1,252	115	16	35	228	0	1,820	1,592	229
G	84	0	0	13	3,385	23	3,574	190	165
H	763	31	0	13	16,706	21	17,664	1,030	7,031
I	7,637	99	9	56	46,811	238	55,147	8,514	67,413

[a]Total processing time: time considering all steps from the multi-criteria approach + time to consolidate results for interface creation.
[b]Total processing time without permutation: total processing time - time to create *permutation sub-sequence blocks.*
[c]ProM Interface Processing Time: time for ProM to present the results to the user.

Table 6.8: Smallest variant size and number of variants per event log.

Event log ID	Post-processing Smallest Variant Size (1)	Pre-processing Number of Variants (2)	Post-processing Number of Variants (3)
A	3	20	3
B	2	49	4
C	1	58	15
D	10	50	7
E	10	198	11
F	8	528	14
G	23	705	8
H	23	4,225	69
I	23	14,451	200

lower permutation process time than E because H created 2 *sub-sequence blocks*, reducing the size of activities to permute to a maximum of 8;

4. The removal of outliers and infrequent process variants reduced significantly the total number of variants per event log (see Table 6.8, columns 2 and 3). This reduction in the number of variants contributes for better performance in Multi-CAT;

5. Multi-CAT demonstrated in our tests a good performance. The most complex test scenario (I), with the highest complexity and 50,000 cases, took 2.04 minutes to execute the complete analysis.

6.4 Discussion

We successfully applied Multi-CAT in a real sepsis treatment process, and with the execution of this test scenario, we acquired more optimization insights that in our previous manual analysis (see Chapter 3). For example, during our tests with Multi-CAT, we found out that the hospital could improve their triage process to increase the chances of an earlier identification of sepsis suspicion patients (e.g. the hospital could prepare more its nurses or deploy a system to help them in the identification process).

In Multi-CAT, the UV is an important artefact to rank variants and clusters, as well as, to cluster variants based in their performance. The unique value, as a composite indicator, allows the simplification of complex measures (set of simultaneous criteria) and, as a consequence, makes the analysis simpler. On the other hand, it may be challenging for users to identify the correct importance weights and the criteria aggregation method, and it also may

provide lack of transparency regarding the method to create the composite indicator [Barclay et al., 2018].

The creation of *sub-sequence* and *permutation sub-sequence blocks* turns out to be a good option to simplify the analyses, making it easier to identify important process behaviors. For example, in the sepsis CP process, the number of activities was reduced in 4 (from 13 activities to 7). The creation of *permutation sub-sequence blocks* should be used carefully. Sometimes using *permutation sub-sequence blocks* can hide a sub-sequence that contributed to better or worse results. For example, for the given sub-sequences "A > B > C" and "B > A > C" the *permutation sub-sequence blocks* could hide that the swap of activities A and B could affect the outcomes. In addition, the creation of the *permutation sub-sequence blocks* was, in most performance tests, the step that consumed the most part of the total processing time.

The Levenshtein distance matrix proved to be helpful in the identification of the set of activities and sub-sequences that can add higher benefits with smaller process changes. For example, in the sepsis clinical pathways test scenario, analyzing the Levenshtein distance matrix we could identify that moving the registry of clinical notes to after the prescription of medicines could improve significantly the process (comparison of variant 6 with variant 1, with a distance of 2 and a difference of 5.63 minutes in the time to give the antibiotics).

We demonstrated that the Global UV can be used to check an organization performance evolution regarding the execution of a given process. We identified that the hospital slightly decreased its sepsis CP outcomes during a period of 2 years, serving as a signal for them to reevaluate the execution of their process. In our tests, we only considered as outcomes the time to give antibiotics, but a combination of different measures can be used. For example, minimizing mortality, LOS and costs simultaneously.

Our performance tests indicated a good performance of Multi-CAT. In our most complex test scenario (with the highest process complexity, 50,000 cases and the largest variant size of 81 activities) the complete processing time took 2.04 minutes. If the same analysis was performed manually, it would take more time, with a higher probability of human error.

It is important to note that the identification/understanding of the optimization recommendations provided by Multi-CAT was possible as we had a good understanding of both processes we used for the validation (including their normative models). Thus, the analysis of Multi-CAT results should be done with/by a specialist in the process. In general, the specialist knows the reasons of process behaviors, and this is important to support the identification

of process optimization recommendations.

With the execution of all tests presented in this chapter, we can conclude that Multi-CAT has high potential to help in the optimization of processes with a good performance. The tool provided the expected results in two completely different use cases, indicating potential capability of the tool to be applicable to different business areas.

7
Conclusions

In this thesis, we generated a broad Brazilian sepsis population epidemiology report from 2006 to 2015 for SUS hospitalizations. The incidence increased 50.5% during the study period. The overall lethality rate of sepsis was 46.3%, and for hospitalizations with admission to the ICU, it was 64.5%. From 2006 to 2015, the lethality rate for children/teenagers decreased, but for all other age groups, it became worse with an increase of 11.4%. The lethality rate in public hospitals (55.5%) was higher than in private hospitals (37.0%). This report evidenced that the treatment of sepsis needs close attention in Brazil.

The employment of Clinical Pathways (CP) is one possible solution to improve sepsis treatment outcomes. However, the management and evaluation of the execution of CPs is not an easy task and has several challenges. We have successfully applied process mining techniques to evaluate the execution of a sepsis treatment in a large private hospital. We identified the real treatment process, its adherence to their CP, performance indicators, bottlenecks and recommendations for process optimization. We consider these results very promising since they can help the hospital in the management of their sepsis treatment and can reduce their burden in the extraction of KPIs as was confirmed through a structured interview with a panel of experts. Due to limitations in the process mining area, all our analyses regarding the identification of recommendations to improve the CP were done manually, taking a significant amount of time.

To tackle these limitations, we proposed, implemented and validated a novel process mining technique that supports users to improve their processes, considering a set of criteria. We validated the tool in 11 different test scenarios providing the expected results and good performance. In the most complex test scenario, with 58 activities, 50,000 cases and 14,451 variants, Multi-CAT took 2.04 minutes to execute the analysis. The technique has potential to be applied in different business areas. For the sepsis CP test scenario, we obtained more optimization recommendations than was previously manually identified.

7.1 Thesis contributions

This thesis contributes to the scientific and technological context in three main aspects:

1. Presents a Brazilian sepsis population epidemiology from 2006 to 2015. We believe this report can contribute significantly since, differently from previous studies, it characterizes the epidemiology of sepsis using a national database, including all hospitalizations (not only intensive care unit cases), with a total of 724,458 cases from 4,271 public and private Brazilian hospitals, all severities of sepsis, and all patient ages (not only adults), thus, providing a broad overview of sepsis in Brazil. We expect that this work may help organizations in planning policies to improve the sepsis Brazilian scenario;

2. Presents a research study to evaluate a sepsis clinical pathway using process mining techniques. We believe this work can contribute significantly since, differently from previous studies, we selected and applied process mining techniques to tackle real needs and challenges identified from a hospital in the evaluation of their sepsis clinical pathway. In addition, we provided a novel technique to optimize a CP. We expect that this work may help hospitals in the difficult task of evaluating clinical pathways;

3. Proposes, implements and tests a novel process mining technique that supports users to optimize their processes considering multiple simultaneous criteria. Differently from previous approaches this technique: A. identifies and highlights a set of activities and sub-sequences that provided positive or negative outcomes; B. is designed to work with multiple simultaneous criteria considering different importance levels defined by the user; C. generates outputs considering a group of different variants instead of a unique variant; D. provides results that allow the user to re-evaluate constraints (rules) predefined in their normative process; E. helps the user to identify the set of behaviours that mostly contribute to improve their process outcomes; F. supports the performance evaluation with respect to the outcomes in the execution of a process in different time frames or different organizations.

7.2 Future perspectives

The execution of this research study opened new possible ideas for future study:

1. DATASUS is currently deploying the Minimal Data Set (CMD - *Conjunto Mínimo de Dados*) system (form) to collect data from different types of Brazilian medical encounters (e.g. hospitalization, emergency assistance, outpatient care) [DATASUS, 2018a]. This system contains data from private and public health care facilities and will unify the available health care data in one place. The CMD data is available for open access (like the SIHSUS data) and it will allow a broader epidemiology view. In this way, the creation of a new sepsis population-level epidemiology report using CMD data will benefit the Brazilian government and international communities in planning policies to improve the sepsis scenario;

2. The evaluation of the sepsis CP using process mining techniques presented in Chapter 3 demonstrated to be useful for the hospital. We suggest the replication of the research study in different healthcare facilities and for different diseases to check its generalization. This process mining analysis could also be used by communities (e.g. the Latin American Sepsis Institute) to have an overview of the treatment of the diseases executed in different healthcare facilities supporting the update, management and deployment of clinical guidelines and clinical pathways. A pre-requisite for the replication of this research study in different contexts is that the registration of healthcare information (e.g. triage, medical evaluation, prescription) is done using integrated systems, allowing the creation of the event log. Sadly, the adoption of HIS and EHR in Brazil is still small [Computer World, 2012; iMedicina, 2017].

With respect to Multi-CAT, we suggest:

1. To test Multi-CAT using processes from different business areas for better verification of its generalization. We are currently applying the tool to identify optimization recommendations for an on-line customer service process of a company that attend three different restaurant chains;

2. Implement new criterion type of measures to support the analysis of skewed distributions (e.g. median, IQR);

3. Test the sensitivity of the tool. This is important to understand the relationship between the input parameters (e.g. method to remove outliers,

method to normalize data, minimal number of cases per variant, method to cluster variants, criteria aggregation method, importance of each criterion) and the output results. In this way, it is possible to identify which input parameters contribute significantly to generate different outcomes and as consequence require more attention from the user in its definition;

4. Implement a functionality that automatically run different input parameters combinations and presents to the user the recurrent recommendations (output results). This functionality will simplify the tool usage as it will reduce the number of parameters requested to the user;

5. Implement a functionality that provides a recommendation to the user of the minimal number of cases per variant;

6. All tool tests were performed only by us, who designed and created Multi-CAT. In this way, it is important that the tool is used by other users to evaluate its usage and results interpretation. Basically, the following questions should be answered: A. What is the (expected) learning curve to use the tool?; B. What is the learning effect in interpreting the tool's recommendations?;

7. To adapt Multi-CAT to consider resources, like professionals, beds, rooms, and equipment, in the analysis. The good execution of a process is not exclusively associated with the definition of its normative model, but also to the correct allocation of resources. We believe that the combination of Discrete Event Simulation (DES) with process mining techniques can support this aim. DES can identify the adequate number of resources to be allocated for each activity. One possible challenging regarding this topic is that an event log not always presents all events executed by resources (e.g. the sepsis event log from Hospital Samaritano has a sub-set of the events executed by the staff from the ED), and in this way, it is not possible to identify in an automatic way the current allocation of resources;

8. Even though we put strong effort to create a simple tool, we understand that it is not ready to be used by clinicians. Thus, we suggest the evaluation of Multi-CAT in a hospital environment. This would allow the collection of user feedback providing raw material for adapting the tool for the health care context. In addition, parts of the multi-criteria analysis could be considered to be automated to simplify its usage as much as possible, for example, the extraction of data and preparation of the event log (Step 1) could be totally automated;

9. To test the Global UV in the comparison of the outcome performance of different organizations executing the same process. For example, to compare 2 or more hospitals regarding the execution of their sepsis CP. This analysis would support sepsis communities (e.g. ILAS) to identify CPs that provided good outcomes, identify differences in the CPs that possibly can contribute to better outcomes, and identify the health care facilities that require priority in the improvement of their CP;

10. To implement manual clustering of variants. Multi-CAT automatically cluster variants according to their UV. The idea is to allow the user to select the variants they want to consider in each cluster. This would provide a more flexible and personal analysis.

Bibliography

Acosta, C. D., Knight, M., Lee, H. C., Kurinczuk, J. J., Gould, J. B., and Lyndon, A. (2013). The continuum of maternal sepsis severity: incidence and risk factors in a population-based cohort study. *PloS one*, 8(7):e67175.

Adrie, C., Alberti, C., Chaix-Couturier, C., Azoulay, É., de Lassence, A., Cohen, Y., Meshaka, P., Cheval, C., Thuong, M., Troché, G., et al. (2005). Epidemiology and economic evaluation of severe sepsis in france: age, severity, infection site, and place of acquisition (community, hospital, or intensive care unit) as determinants of workload and cost. *Journal of critical care*, 20(1):46–58.

Aggarwal, C. C. (2015). *Data mining: the textbook*. Springer.

Angus, D. C., Linde-Zwirble, W. T., Lidicker, J., Clermont, G., Carcillo, J., and Pinsky, M. R. (2001). Epidemiology of severe sepsis in the united states: analysis of incidence, outcome, and associated costs of care. *Critical care medicine*, 29(7):1303–1310.

Apache Commons (2018). Apache commons text. Available at `https://commons.apache.org/proper/commons-text/`. Accessed 15 June 2018.

Arefian, H., Heublein, S., Scherag, A., Brunkhorst, F. M., Younis, M. Z., Moerer, O., Fischer, D., and Hartmann, M. (2017). Hospital-related cost of sepsis: A systematic review. *Journal of Infection*, 74(2):107–117.

Augusto, V., Xie, X., Prodel, M., Jouaneton, B., and Lamarsalle, L. (2016). Evaluation of discovered clinical pathways using process mining and joint agent-based discrete-event simulation. In *Proceedings of the 2016 Winter Simulation Conference*, pages 2135–2146. IEEE Press.

Baker, K., Dunwoodie, E., Jones, R. G., Newsham, A., Johnson, O., Price, C. P., Wolstenholme, J., Leal, J., McGinley, P., Twelves, C., et al. (2017). Process mining routinely collected electronic health records to define real-life clinical pathways during chemotherapy. *International Journal of Medical Informatics*, 103:32–41.

Barclay, M., Dixon-Woods, M., and Lyratzopoulos, G. (2018). The problem with composite indicators. *BMJ Qual Saf*, pages bmjqs–2018.

Beale, R., Reinhart, K., Brunkhorst, F. M., Dobb, G., Levy, M., Martin, G., Martin, C., Ramsey, G., Silva, E., Vallet, B., et al. (2009). Promoting global research excellence in severe sepsis (progress): lessons from an international sepsis registry. *Infection*, 37(3):222–232.

BPIC (2017). Business process intelligence challenge (bpic). Available at `https://www.win.tue.nl/bpi/doku.php?id=2017:challenge`. Accessed 25 January 2018.

Burattin, A. (2016). Plg2: Multiperspective process randomization with online and offline simulations. In *BPM (Demos)*, pages 1–6.

Burattin, A. (2017). Plg - process log generator. Available at `http://plg.processmining.it/`. Accessed 05 July 2018.

Caron, F., Vanthienen, J., Vanhaecht, K., Van Limbergen, E., De Weerdt, J., and Baesens, B. (2014). Monitoring care processes in the gynecologic oncology department. *Computers in biology and medicine*, 44:88–96.

Carrillo-Esper, R., Carrillo-Córdova, J. R., and Carrillo-Córdova, L. D. (2009). Epidemiological study of sepsis in mexican intensive care units. *Cirugia y cirujanos*, 77(4):301–308.

Chalupka, A. N. and Talmor, D. (2012). The economics of sepsis. *Critical care clinics*, 28(1):57–76.

Chawla, A., Westrich, K., Matter, S., Kaltenboeck, A., and Dubois, R. (2016). Care pathways in us healthcare settings: current successes and limitations, and future challenges. *The American journal of managed care*, 22(1):53–62.

Cheng, B., Xie, G., Yao, S., Wu, X., Guo, Q., Gu, M., Fang, Q., Xu, Q., Wang, D., Jin, Y., et al. (2007). Epidemiology of severe sepsis in critically ill surgical patients in ten university hospitals in china. *Critical care medicine*, 35(11):2538–2546.

CIHI (2009). In focus: a national look at sepsis. *Ottawa, Ont: CIHI*. Available at `http://publications.gc.ca/collections/collection_2009/icis-cihi/H118-60-2009E.pdf`. Accessed 26 February 2018.

Claes, J. and Poels, G. (2012). Process mining and the prom framework: an exploratory survey. In *International Conference on Business Process Management*, pages 187–198. Springer.

Clarivate Analytics (2018). Web of science. Available at `https://webofknowledge.com/`. Accessed 01 February 2018.

Commission, J. R. C.-E. et al. (2008). *Handbook on constructing composite indicators: Methodology and user guide*. OECD publishing.

Computer World (2012). Adoção do prontuário eletrônico no brasil ainda é baixa. Available at `https://computerworld.com.br/2012/07/16/adocao-do-prontuario-eletronico-no-brasil-ainda-e-baixa/`. Accessed 05 October 2018.

Conde, K. A. P., Silva, E., Silva, C. O., Ferreira, E., Freitas, F. G. R., Castro, I., Rea-Neto, A., Grion, C. M. C., Moura, A. D., Lobo, S. M., et al. (2013). Differences in sepsis treatment and outcomes between public and private hospitals in brazil: a multicenter observational study. *PLoS One*, 8(6):e64790.

DATASUS (2018a). Cmd - conjunto mínimo de dados. Available at `https://conjuntominimo.saude.gov.br`. Accessed 24 July 2018.

DATASUS (2018b). Portal da saúde - transferência de arquivos. Available at `http://www2.datasus.gov.br/DATASUS/index.php?area=0901`. Accessed 08 January 2018.

Dees, M., de Leoni, M., and Mannhardt, F. (2017). Enhancing process models to improve business performance: A methodology and case studies. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 232–251. Springer.

Dramski, M. (2017). Inductive mining in modeling of the ship's route. In *Marine Navigation: Proceedings of the 12th International Conference on Marine Navigation and Safety of Sea Transportation*, pages 51–55.

Elsevier (2018). Scopus. Available at `https://www.scopus.com/`. Accessed 01 February 2018.

Fernandez-Llatas, C., Lizondo, A., Monton, E., Benedi, J.-M., and Traver, V. (2015). Process mining methodology for health process tracking using real-time indoor location systems. *Sensors*, 15(12):29821–29840.

Fluxicon (2018). Disco. Available at `https://fluxicon.com/disco/`. Accessed 15 January 2018.

Fujino, Y., Kubo, T., Muramatsu, K., Murata, A., Hayashida, K., Tomioka, S., Fushimi, K., and Matsuda, S. (2014). Impact of regional clinical pathways on the length of stay in hospital among stroke patients in japan. *Medical care*, 52(7):634–640.

Gaieski, D. F., Edwards, J. M., Kallan, M. J., and Carr, B. G. (2013). Benchmarking the incidence and mortality of severe sepsis in the united states. *Critical care medicine*, 41(5):1167–1174.

Gan, G., Ma, C., and Wu, J. (2007). *Data clustering: theory, algorithms, and applications*, volume 20. Siam.

Gobatto, A. L. N., Besen, B. A. M. P., and Azevedo, L. C. P. (2017). How can we estimate sepsis incidence and mortality? *Shock*, 47(1S):6–11.

Gonzalez, T. F. (1985). Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293–306.

Greene, S. E. and Nash, D. B. (2009). Pay for performance: an overview of the literature. *American Journal of Medical Quality*, 24(2):140–163.

Group, E. S. et al. (2004). Episepsis: a reappraisal of the epidemiology and outcome of severe sepsis in french intensive care units. *Intensive care medicine*, 30(4):580–588.

GTRH (2010). Grupo de trabalho sobre remuneração dos hospitais. sistemáticas de remuneração dos hospitais que atuam na saúde suplementar: diretrizes e rumos. Available at `http://www.ans.gov.br/images/stories/Particitacao_da_sociedade/2016_gt_opme/grupo5_orteses_proteses_materiais_especiais_rodadarj_2010.pdf`. Accessed 26 July 2018.

Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

HIPAA (2012). Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability act (hipaa) privacy rule. Available at `https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html`. Accessed 26 July 2018.

Hospital Samaritano (2018a). Hospital samaritano, com 121 anos de atividades, o hospital samaritano de são paulo destaca-se pela excelência e humanização no atendimento à saúde. Available at `http://samaritano.com.br/institucional/institucional/`. Accessed 15 January 2018.

Hospital Samaritano (2018b). Hospital samaritano de são paulo. Available at `http://samaritano.com.br/`. Accessed 15 January 2018.

IBGE (2018a). Brasil em síntese, distribuição da população por grandes grupos de idade. Available at http://brasilemsintese.ibge.gov.br/populacao/distribuicao-da-populacao-por-grandes-grupos-de-idade.html. Accessed 08 January 2018.

IBGE (2018b). Ipca - índice nacional de preços ao consumidor amplo. Available at http://www.ibge.gov.br/home/estatistica/indicadores/precos/inpc_ipca/defaultinpc.shtm. Accessed 08 January 2018.

IBGE (2018c). Projeção da população por sexo e idades. Available at http://www.ibge.gov.br/home/estatistica/populacao/projecao_da_populacao/2013/default_tab.shtm. Accessed 08 January 2018.

IEEE (2016). 1849-2016 xes standard. Available at http://www.xes-standard.org/. Accessed 22 January 2018.

Ikeda, M., Otaki, K., and Yamamoto, A. (2014). Formal concept analysis for process enhancement based on a pair of perspectives. In *CLA*, pages 59–70.

ILAS (2016). O que é sepse. Available at http://www.ilas.org.br/o-que-e-sepse.php. Accessed 28 January 2018.

ILAS (2018a). Instituto latino americano de sepsis. Available at http://www.ilas.org.br/. Accessed 04 January 2018.

ILAS (2018b). Relatório nacional - protocolos gerenciados de sepse - sepse e choque séptico 2005-2016. Available at http://www.ilas.org.br/assets/arquivos/relatorio-nacional/relatorio-nacional-final.pdf/. Accessed 05 January 2018.

iMedicina (2017). Prontuário eletrônico: 7 estatísticas reveladoras que vão mudar sua opinião. Available at http://blog.imedicina.com.br/prontuario-eletronico-estatisticas-artigo-pe-topo/. Accessed 05 October 2018.

Japiassú, A. M., Amâncio, R. T., Mesquita, E. C., Medeiros, D. M., Bernal, H. B., Nunes, E. P., Luz, P. M., Grinsztejn, B., and Bozza, F. A. (2010). Sepsis is a major determinant of outcome in critically ill hiv/aids patients. *Critical Care*, 14(4):R152.

Java-ML (2012). Java machine learning library. Available at http://java-ml.sourceforge.net/. Accessed 12 February 2018.

Jolley, R. J., Quan, H., Jetté, N., Sawka, K. J., Diep, L., Goliath, J., Roberts, D. J., Yipp, B. G., and Doig, C. J. (2015a). Validation and optimisation of an icd-10-coded case definition for sepsis using administrative health data. *BMJ open*, 5(12):e009487.

Jolley, R. J., Sawka, K. J., Yergens, D. W., Quan, H., Jetté, N., and Doig, C. J. (2015b). Validity of administrative data in recording sepsis: a systematic review. *Critical care*, 19(1):139.

Kaufman, L. and Rousseeuw, P. (1987). *Clustering by means of medoids*. North-Holland.

Kempker, J. A. and Martin, G. S. (2016). The changing epidemiology and definitions of sepsis. *Clinics in chest medicine*, 37(2):165–179.

Khalifa, M. and Alswailem, O. (2015). Clinical pathways: Identifying development, implementation and evaluation challenges. In *ICIMTH*, pages 131–134.

Knoop, S. T., Skrede, S., Langeland, N., and Flaatten, H. K. (2017). Epidemiology and impact on all-cause mortality of sepsis in norwegian hospitals: A national retrospective study. *PloS one*, 12(11):e0187990.

Kumar, A., Roberts, D., Wood, K. E., Light, B., Parrillo, J. E., Sharma, S., Suppes, R., Feinstein, D., Zanotti, S., Taiberg, L., et al. (2006). Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. *Critical care medicine*, 34(6):1589–1596.

Kumar, G., Kumar, N., Taneja, A., Kaleekal, T., Tarima, S., McGinley, E., Jimenez, E., Mohan, A., Khan, R. A., Whittle, J., et al. (2011). Nationwide trends of severe sepsis in the 21st century (2000-2007). *Chest Journal*, 140(5):1223–1231.

Kumar, M. et al. (2013). An optimized farthest first clustering algorithm. In *Engineering (NUiCONE), 2013 Nirma University International Conference on*, pages 1–5. IEEE.

Lagu, T., Rothberg, M. B., Shieh, M.-S., Pekow, P. S., Steingrub, J. S., and Lindenauer, P. K. (2012). Hospitalizations, costs, and outcomes of severe sepsis in the united states 2003 to 2007. *Critical care medicine*, 40(3):754–761.

Lakshmanan, G. T., Rozsnyai, S., and Wang, F. (2013). Investigating clinical care pathways correlated with outcomes. In *Business process management*, pages 323–338. Springer.

Lehto, T., Hinkka, M., and Hollmén, J. (2016). Focusing business improvements using process mining based influence analysis. In *International Conference on Business Process Management*, pages 177–192. Springer.

Lehto, T., Hinkka, M., and Hollmén, J. (2017). Focusing business process lead time improvements using influence analysis. In *SIMPDA*.

Liao, X., Du, B., Lu, M., Wu, M., and Kang, Y. (2016). Current epidemiology of sepsis in mainland china. *Annals of Translational Medicine*, 4(17):324.

Lismont, J., Janssens, A.-S., Odnoletkova, I., vanden Broucke, S., Caron, F., and Vanthienen, J. (2016). A guide for the application of analytics on healthcare processes: A dynamic view on patient pathways. *Computers in biology and medicine*, 77:125–134.

Low, W. Z., De Weerdt, J., Wynn, M. T., ter Hofstede, A. H., van der Aalst, W. M., and vanden Broucke, S. (2014). Perturbing event logs to identify cost reduction opportunities: A genetic algorithm-based approach. In *Evolutionary Computation (CEC), 2014 IEEE Congress on*, pages 2428–2435. IEEE.

Low, W. Z., vanden Broucke, S. K., Wynn, M. T., ter Hofstede, A. H., De Weerdt, J., and van der Aalst, W. M. (2016). Revising history for cost-informed process improvement. *Computing*, 98(9):895–921.

Machado, F. R., Cavalcanti, A. B., Bozza, F. A., Ferreira, E. M., Carrara, F. S. A., Sousa, J. L., Caixeta, N., Salomao, R., Angus, D. C., Azevedo, L. C. P., et al. (2017). The epidemiology of sepsis in brazilian intensive care units (the sepsis prevalence assessment database, spread): an observational study. *The Lancet Infectious Diseases*, 17(11):1180–1189.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Mangia, C. M. F., Kissoon, N., Branchini, O. A., Andrade, M. C., Kopelman, B. I., and Carcillo, J. (2011). Bacterial sepsis in brazilian children: a trend analysis from 1992 to 2006. *PLoS One*, 6(6):e14817.

Mans, R. S., Van der Aalst, W. M., and Vanwersch, R. J. (2015). *Process mining in healthcare: Evaluating and exploiting operational healthcare processes*. Springer.

Ministério da Saúde (2018). Terminologia básica em saúde. Available at `http://bvsms.saude.gov.br/bvs/publicacoes/0112terminologia1.pdf`. Accessed 08 January 2018.

Navarro, G. (2001). A guided tour to approximate string matching. *ACM computing surveys (CSUR)*, 33(1):31–88.

Neira, R. A. Q., de Vries, G.-J., Caffarel, J., and Stretton, E. (2017). Extraction of data from a hospital information system to perform process mining. In *MedInfo*, pages 554–558.

On Target (2018). How to consolidate multiple kpis. Available at `http://www.cbsolution.net/ontarget/how_to_consolidate_multiple_kpis`. Accessed 14 October 2018.

Oracle Corporation (2018). Java. Available at `https://www.java.com`. Accessed 08 February 2018.

Palleschi, M. T., Sirianni, S., O'connor, N., Dunn, D., and Hasenau, S. M. (2014). An interprofessional process to improve early identification and treatment for sepsis. *Journal for Healthcare Quality*, 36(4):23–31.

Panella, M., Marchisio, S., and Di Stanislao, F. (2003). Reducing clinical variations with clinical pathways: do pathways work? *International Journal for Quality in Health Care*, 15(6):509–521.

Papali, A., Verceles, A. C., Augustin, M. E., Colas, L. N., Jean-Francois, C. H., Patel, D. M., Todd, N. W., McCurdy, M. T., West, T. E., et al. (2017). Sepsis in haiti: Prevalence, treatment, and outcomes in a port-au-prince referral hospital. *Journal of critical care*, 38:35–40.

Profit, J., Typpo, K. V., Hysong, S. J., Woodard, L. D., Kallen, M. A., and Petersen, L. A. (2010). Improving benchmarking by using an explicit framework for the development of composite indicators: an example using pediatric quality of care. *Implementation Science*, 5(1):13.

R core team (2018). R: a language and environment for statistical computing. Available at `https://www.r-project.org/`. Accessed 08 January 2018.

Revista Exame (2018). Os hospitais brasileiros de excelência em 2014. Available at `https://exame.abril.com.br/seu-dinheiro/os-hospitais-brasileiros-de-excelencia-em-2014/`. Accessed 15 January 2018.

Rodríguez, F., Barrera, L., De La Rosa, G., Dennis, R., Dueñas, C., Granados, M., Londoño, D., Molina, F., Ortiz, G., and Jaimes, F. (2011). The epidemiology of sepsis in colombia: a prospective multicenter cohort study in ten university hospitals. *Critical care medicine*, 39(7):1675–1682.

Rojas, E., Munoz-Gama, J., Sepúlveda, M., and Capurro, D. (2016). Process mining in healthcare: A literature review. *Journal of biomedical informatics*, 61:224–236.

Saaty, R. W. (1987). The analytic hierarchy process—what it is and how it is used. *Mathematical modelling*, 9(3-5):161–176.

Salluh, J. I., Soares, M., and Keegan, M. T. (2017). Understanding intensive care unit benchmarking. *Intensive Care Medicine*, pages 1–5.

Sanjoy Dasgupta, P. L. (2002). Performance guarantees for hierarchical clustering. In *15th Annual Conference on Computational Learning Theory*, pages 351–363. Springer.

Santos, A. C. d. (2009). Sistema de informações hospitalares do sistema único de saúde: documentação do sistema para auxiliar o uso das suas informações. Master's thesis, Fundação Oswaldo Cruz. Available at `https://www.arca.fiocruz.br/bitstream/icict/2372/1/ENSP_Disserta%C3%A7%C3%A3o_Santos_Andr%C3%A9ia_Cristina.pdf`. Accessed 26 February 2018.

SCCM (2018). Surviving sepsis campaign. Available at `http://www.survivingsepsis.org/`. Accessed 04 January 2018.

Seo, S. (2006). *A review and comparison of methods for detecting outliers in univariate data sets*. PhD thesis, University of Pittsburgh. Available at `http://d-scholarship.pitt.edu/7948/1/Seo.pdf`. Accessed 20 February 2018.

Silva, E., de Almeida Pedro, M., Sogayar, A. C. B., Mohovic, T., Silva, C. L. O., Janiszewski, M., Cal, R. G. R., de Sousa, É. F., Abe, T. P., de Andrade, J., et al. (2004). Brazilian sepsis epidemiological study (bases study). *Critical Care*, 8(4):R251.

Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G. R., Chiche, J.-D., Coopersmith, C. M., et al. (2016). The third international consensus definitions for sepsis and septic shock (sepsis-3). *Jama*, 315(8):801–810.

Sogayar, A. M., Machado, F. R., Rea-Neto, A., Dornas, A., Grion, C. M., Lobo, S. M., Tura, B. R., Silva, C. L., Cal, R. G., Beer, I., et al. (2008). A multicentre, prospective study to evaluate costs of septic patients in brazilian intensive care units. *Pharmacoeconomics*, 26(5):425–434.

SSC (2015). Surviving sepsis campaign bundles - updated bundles in response to new evidence. Available at `http://www.survivingsepsis.org/SiteCollectionDocuments/SSC_Bundle.pdf`. Accessed 15 January 2018.

Stillwell, W. G., Seaver, D. A., and Edwards, W. (1981). A comparison of weight approximation techniques in multiattribute utility decision making. *Organizational behavior and human performance*, 28(1):62–77.

Stoller, J., Halpin, L., Weis, M., Aplin, B., Qu, W., Georgescu, C., and Nazzal, M. (2016). Epidemiology of severe sepsis: 2008-2012. *Journal of critical care*, 31(1):58–62.

Taniguchi, L. U., Bierrenbach, A. L., Toscano, C. M., Schettino, G. P., and Azevedo, L. C. (2014). Sepsis-related deaths in brazil: an analysis of the national mortality registry from 2002 to 2010. *Critical Care*, 18(6):608.

The world bank (2018). The world bank, world development indicators: exchange rates and prices. Available at `http://wdi.worldbank.org/table/4.16`. Accessed 08 January 2018.

Tsertsvadze, A., Royle, P., Seedat, F., Cooper, J., Crosby, R., and McCarthy, N. (2016). Community-onset sepsis and its public health burden: a systematic review. *Systematic reviews*, 5(1):81.

TU/E - Math&CS Department (2018a). Process mining group. Available at `http://www.processmining.org/`. Accessed 14 January 2018.

TU/E - Math&CS Department (2018b). Prom tools. Available at `http://www.promtools.org/doku.php`. Accessed 15 January 2018.

United Nations Development Programme (2017). Human development report, work for human development. Available at `http://hdr.undp.org/sites/default/files/2015_human_development_report.pdf`. Accessed 27 February 2018.

UOL Economia (2018). Cotações, câmbio, dólar comercial. Available at `http://economia.uol.com.br/cotacoes/cambio/dolar-comercial-estados-unidos/?historico`. Accessed 08 January 2018.

van der Aalst, W. (2016). *Process mining: data science in action*. Springer, second edition.

van der Aalst, W. M., Low, W. Z., Wynn, M. T., and ter Hofstede, A. H. (2015). Change your history: Learning from event logs to improve processes. In *Computer Supported Cooperative Work in Design (CSCWD), 2015 IEEE 19th International Conference on*, pages 7–12. IEEE.

Victora, C. G., Barreto, M. L., do Carmo Leal, M., Monteiro, C. A., Schmidt, M. I., Paim, J., Bastos, F. I., Almeida, C., Bahia, L., Travassos, C., et al. (2011). Health conditions and health-policy innovations in brazil: the way forward. *The Lancet*, 377(9782):2042–2053.

Watts, J. (2016). Brazil's health system woes worsen in economic crisis. *Lancet*, 387(10028):1603–4.

WHO (2018). World health organization, global health observatory data repository, life expectancy data by country. Available at `http://apps.who.int/gho/data/node.main.688`. Accessed 08 January 2018.

Xu, X., Jin, T., and Wang, J. (2016). Summarizing patient daily activities for clinical pathway mining. In *e-Health Networking, Applications and Services (Healthcom), 2016 IEEE 18th International Conference on*, pages 1–6. IEEE.

Yang, W. and Su, Q. (2014). Process mining for clinical pathway: Literature review and future directions. In *Service Systems and Service Management (ICSSSM), 2014 11th International Conference on*, pages 1–5. IEEE.

Yoo, S., Cho, M., Kim, E., Kim, S., Sim, Y., Yoo, D., Hwang, H., and Song, M. (2016). Assessment of hospital processes using a process mining technique: Outpatient process analysis at a tertiary hospital. *International journal of medical informatics*, 88:34–43.

Zhang, Y., Padman, R., and Patel, N. (2015). Paving the cowpath: Learning and visualizing clinical pathways from electronic health record data. *Journal of biomedical informatics*, 58:186–197.

Zhou, J., Tian, H., Du, X., Xi, X., An, Y., Duan, M., Weng, L., Du, B., Group, C. C. C. C. T., et al. (2017). Population-based epidemiology of sepsis in a subdistrict of beijing. *Critical Care Medicine*, 45(7):1168–1176.

A
ROC curves from multiple logistic regressions

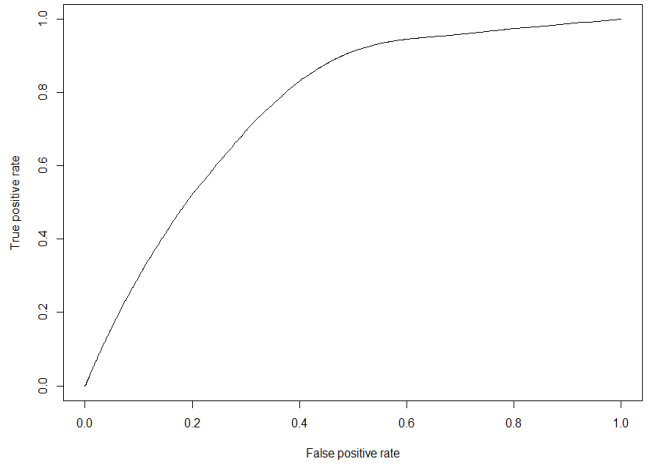


Figure A.1: ROC Curve - death as dependent variable and age, gender, race as independent variables. The prediction error is 29.6% and the area under the ROC curve is 0.77.

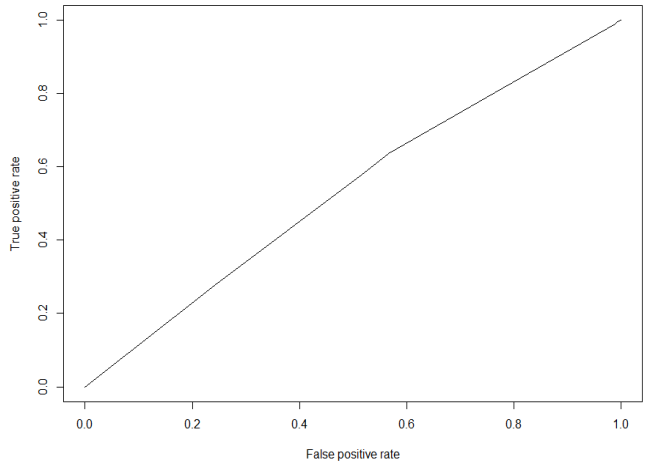


Figure A.2: ROC Curve - death as dependent variable and gender, race as independent variables. The prediction error is 46.8% and the area under the ROC curve is 0.54.

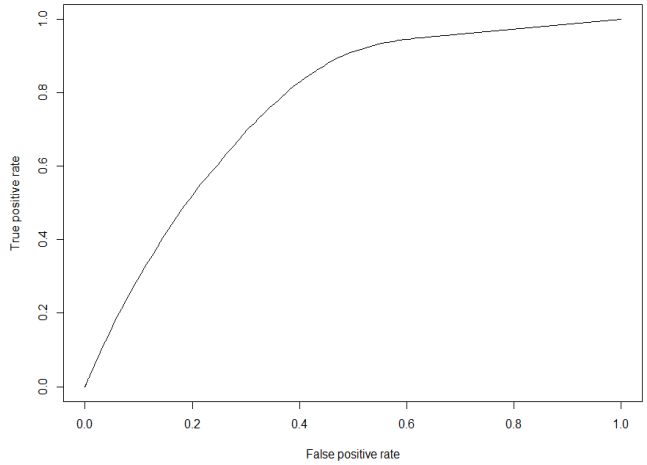


Figure A.3: ROC Curve - death as dependent variable and age as independent variable. The prediction error is 29.6% and the area under the ROC curve is 0.76.

B
Number of hospitals and cases per hospital group from the treatment efficiency matrix for sepsis

Table B.1: Number of cases per hospital group.

Hospital Type	Hospital Size	Number of Cases
private	small	37,111
private	medium	152,771
private	large	143,608
private	very large	27,548
public	small	28,657
public	medium	86,189
public	large	202,781
public	very large	45,322

Table B.2: Number of hospitals per hospital group.

Hospital Type	Hospital Size	Number of Hospitals
private	small	1,111
private	medium	1,282
private	large	368
private	very large	26
public	small	1,304
public	medium	707
public	large	322
public	very large	32

C
Extraction of data from a hospital information system to perform process mining

The content of this appendix was published in the 16th World Congress on Medical and Health Informatics (MedInfo 2017) [Neira et al., 2017]

C.1
Introduction

Process mining consists of a set of techniques that enable the analysis of business processes using system data. Specialized algorithms treat the data identifying patterns and trends. The use of process mining algorithms allows to discover process models (process discovery), identify deviations in a process (conformance checking), identify bottlenecks and performance indicators (performance checking), and identify how information flows between resources using social networks [1,2].

The event log is the raw material for running process mining algorithms. It contains all events used to construct a journey map and has as main attributes a case ID that represent one instance of the process (in healthcare field it could be the patient or hospitalization identification), the activity performed (e.g. "Perform triage", "Discharge of patient"), and the date and time the activity was performed [1,3].

Several research studies have applied process mining techniques for healthcare and they are present in many medical fields such as cardiology, oncology, diabetes and clinical images [4][5]. For example, Forsberg et al. [6] performed a study to identify the reading chest radiograph process in a Picture Archiving and Communication System (PACS). Rattanavayakorn and Premchaiswadi [7] applied the "working together metric" of social miner techniques to understand the behavior of healthcare professionals when treating patients in a hospital in Bangkok. Mans et al. [8] applied the heuristics miner algorithm to discover and compare the stroke treatment process in four Italian hospitals. Also, they applied performance checking algorithms to discover the bottlenecks and performance indicators for the pre-hospital process. Huang et al. [9] presented a technique that creates a summary of the structure of clinical pathways. They applied the approach for four different diseases (bronchial lung

cancer, colon cancer, gastric cancer, and cerebral infarction) discovering essential medical behavior with specific execution order. Another study analyzed the control flow, organizational and performance perspectives of a gynecological oncology process in a Dutch hospital to obtain insights in the care flow [10].

We are performing a research study using process mining techniques to identify deviations and bottlenecks in a sepsis treatment process in a Brazilian hospital. We expect to identify actions (changes in the process) that can improve the sepsis treatment process. Our first step was the extraction of data from a Hospital Information System (HIS).

The aim of this work is to present the steps we followed to extract data from the HIS to perform the process mining work.

The main purpose of this work is to share our experience regarding the data extraction and all the preparation work associated with this task. We hope that our experience can help other researchers to plan and execute the extraction of data for process mining research studies.

C.2 Methods

Below we present all steps we followed to extract the data from the HIS database.

C.2.1 Research definition

First, we defined the research questions we want to answer in our work:

1. Which is the AS-IS (current process) sepsis treatment process of the hospital? How does the hospital staff treat septic patients?

2. Which are the deviations in the process? Do professionals perform activities in a different order than defined in the normative sepsis treatment process?

3. Which are the bottlenecks in the process? Are there activities in the process that are taking more time than expected?

4. What is the workload of each professional in the hospital? Could a heavy workload cause delays in the process?

5. Which actions can improve the process? For example, we would like to identify changes in processes that might reduce the time it takes to administer initial antibiotic therapy.

The research questions were crucial to understanding which type of information we should extract from the database. For example, if we want to know the AS-IS process, we need to create an event log with the case identification (in our case it is the hospitalization identification), the activity type ("Registry of patient", "Triage," "Medical evaluation") and the completion date and time. If we want to analyze the workload of healthcare professionals then in the event log we need to add the health professional identification and extract information about all hospitalizations (not only of sepsis) to get a complete overview of health professionals tasks.

C.2.2 Mapping the process

In a first visit to the hospital, we analyzed the process the hospital applies for treating sepsis patients. We studied the hospital documents (sepsis guidelines and sepsis screening form), performed interviews with health professionals (2 physicians, 2 nurses, 2 nurse technicians, 1 quality analyst, 1 receptionist) and performed shadowing. The number of interviewers was four.

With all the information collected, we designed two models using the Business Process Model and Notation (BPMN). One model represents the sepsis treatment in the Emergency Department and the second the sepsis treatment in the Intensive Care Unit. The models represent the way health professionals should work when treating sepsis patients (normative models). Both models were updated and validated by the staff (4 physicians, 3 nurses, 2 nurse technicians, 1 quality analyst, 1 laboratory technician, 2 pharmacists) in a second visit to the hospital. The number of interviewers was five.

Figure C.1 presents a simple model of a treatment process based in the Emergency Department model. We created it to help the reader to understand the extraction process. We linked the process activities to the next steps of the extraction.

The process models and all the information collected from the hospital visits were very important to guide us in the selection of the attributes to collect from an immense amount of attributes presented in the database. For example, in the "1. Register Patient" activity we needed to know the time and date this event happened, and who performed the registration.

C.2.3 Identification of tables and fields of the database

When we started this research, we had no knowledge about the HIS database structure. We did not know from which tables and fields we should

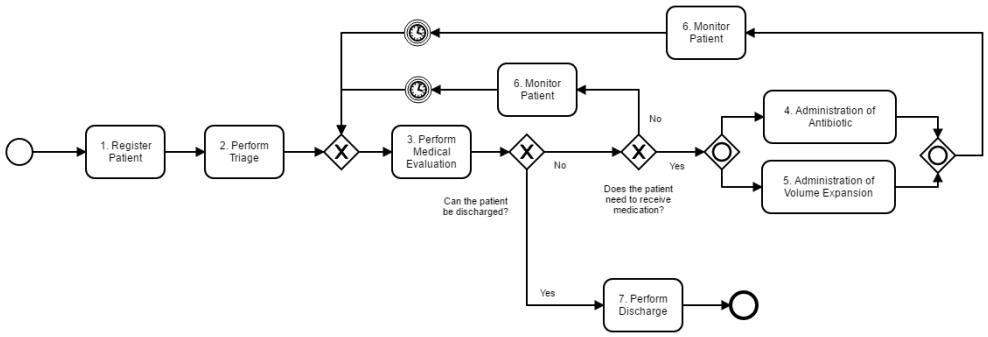


Figure C.1: Simple BPMN model representing a treatment process in the emergency department (*Note: this model was created as an elucidative example. It does not represent the exact and complete process as followed by the hospital).*

extract the data we needed to perform our research.

We asked the HIS development team to give us guidelines of which tables and fields we should collect all the information needed for our research. For this task, we created a spreadsheet containing all attributes that we needed to extract from the database based on the research questions and process models. We sent this spreadsheet together with the process models to the HIS development team and asked them to complete it.

We created two versions of the spreadsheet. The first one presented just one tab with the attributes needed by each step (activity) of the treatment processes. We denominate this tab as "Process Oriented". Since it was difficult and confusing for the development team to fill it, we created a second version of the spreadsheet that contained a new tab, denominated "HIS Modules Oriented", presenting the same attributes needed grouped by functionality of the system. We believe this module oriented view would help them to more easily associate the correct tables and fields of the database, since this is closer to their way of thinking (they do not necessarily know the clinical process that is followed, but they do know the modules used by the users). The main difference between the process and module tabs is the way that the attributes needed are organized. The "Process Oriented" tab presents the attributes grouped by each activity of the process, and the "HIS Modules Oriented" tab presents the same attributes grouped by each module of the HIS. To convert an item from the "Process Oriented" tab to the "HIS Modules Oriented" tab we identified the HIS module used by the clinical user to register the information (attribute) associated with the process activity.

Below we present both tab structures of the spreadsheet.

Process oriented tab

Table C.1 presents a small sample of the first tab. It contains the following columns:

- Step name (A): name of the activity from the BPMN model. E.g., "1. Register Patient", "2. Perform Triage" from Figure C.1;
- Attribute (B): name of the attribute that we need from the process activity. E.g., from the triage step (2. Perform Triage) we need to extract the "temperature", "blood pressure", and "clinical notes". One step name can have many attributes;
- Responsible (C): contact information of the responsible from the HIS development team who filled the spreadsheet line. This is important in case we had doubts about an attribute and thus we could contact directly the professional;
- Table (D): name of the table from the database where the field is located;
- Table field (E): name of the field from the database which contains the attribute information to be extracted;
- Execution - date and time (F): field from the database that contains the date and time that the action was performed in practice. E.g. "What was the time that the administration of the medication was performed for patient John?";
- Execution – role of user (G): field from the database that contains the role of the professional who performed the action. E.g. "Nurse", "Physician";
- Registry - date and time (H): field from the database that contains the date and time that the attribute was entered in the system. E.g. "What was the time that the administration of the medication for patient John was entered in the HIS?";
- Registry - Role of user (I): field from the database that contains the role of the professional who entered the attribute in the system. It is important to also retrieve this information as the person documenting the activity may not be the same as the one executing it (e.g. a doctor may perform an action and ask a nurse to document it in the HIS).

Columns from C to I should be filled by a professional from the HIS development team.

HIS modules oriented tab

Table C.2 presents a small sample of the second tab. Columns B to I are the same from Table C.1. The column "Step Name" (A) was replaced

Table C.1: Sample of process oriented tab of the spreadsheet (*Note: all content data presented is fictitious. The idea is to present the structure of the table. The table presents a subset of the activities from Figure C.1*).

Step Name (A)	**Attribute (B)**	**Responsible (C)**	**Table (D)**	**Table Field (E)**	**Execution**		**Registry**	
					Date and Time (F)	**Role of User (G)**	**Date and Time (H)**	**Role of User (I)**
2. Perform Triage	Clinical Notes	ES	TRIAGE	clinical_notes	exec_date	exec_role	end_date	user_role
2. Perform Triage	Temperature	JC	VITAL_SIGNS	temperature	start_date	perf_role	reg_date	reg_role
2. Perform Triage	Blood Pressure	JC	VITAL_SIGNS	blood_pressure	start_date	perf_role	reg_date	reg_role
3. Perform Medical Evaluation	Clinical Notes	GJV	NOTES	clinical_notes	exec_date	exec_role	end_date	user_role
6. Monitor Patient	Temperature	JC	VITAL_SIGNS	temperature	start_date	perf_role	reg_date	reg_role
6. Monitor Patient	Blood Pressure	JC	VITAL_SIGNS	blood_pressure	start_date	perf_role	reg_date	reg_role

with "Module". This new column presents the module name from the HIS that the information requested can be collected. E.g. "Patient registry", "Electronic Health Record", "Computerized Physician Order Entry", "Imaging", "Transfer of Patient". Columns from C to I should be filled by a professional from the HIS development team.

Table C.1 and Table C.2 present the same content in different views.

Filling the spreadsheet by the HIS development team

We had one main contact person who was in charge to make the communication bridge with the development team. We sent him the spreadsheet and the BPMN process models with clear instructions on how to fill the spreadsheet. This procedure was performed with both versions of the spreadsheet. For the second version, we made it clear that the developers could choose any of the two tabs to fill. During the "filling of the spreadsheet" step, we kept direct contact with the development team to solve their doubts. When receiving a newly filled part of the spreadsheet, we immediately reviewed it to check if there was any white cell (cell not filled) and to check if the cells were with a coherent value (e.g. we asked for "prescription of medication" and we received "prescription_procedure" in the table name – this seems clearly to be not right). In the case of any problem identified, we contacted them to discuss and update the information.

C.2.4 Extraction of data

Based upon the fields identified in the previous step, SQL queries were written to extract the relevant fields from the HIS. For the extraction, we had to anonymize patient and hospitalization data to guarantee that no-one outside the hospital could identify a patient or link the extracted data with the hospital database. In this stage, to anonymize the data:

1. We encrypted any identification code like patient and health professional codes, chart number, hospitalization number, and prescription and administration ids;

2. Rather than storing the date of birth, we calculated the age of patients according to the admission hospitalization date. Patients older than 90 years old have a higher probability of being identified, thus all of these cases ($>$ 90y) were classified as 90 years old;

3. Patients with a weight greater than 130kg also have a greater probability of identification, thus all of them ($>$ 130kg) were classified as 130kg;

Table C.2: Sample of HIS module oriented tab of the spreadsheet *(Note: all content data presented is fictitious. The idea is to present the structure of the table).*

Module (A)	**Attribute (B)**	**Responsible (C)**	**Table (D)**	**Table Field (E)**	**Execution**		**Registry**	
					Date and Time (F)	**Role of User (G)**	**Date and Time (H)**	**Role of User (I)**
Triage	Clinical Notes	ES	TRIAGE	clinical_notes	exec_date	exec_role	end_date	user_role
Vital Signs	Temperature	JC	VITAL_SIGNS	temperature	start_date	perf_role	reg_date	reg_role
Vital Signs	Blood Pressure	JC	VITAL_SIGNS	blood_pressure	start_date	perf_role	reg_date	reg_role
Electronic Health Record	Clinical Notes	GJV	NOTES	clinical_notes	exec_date	exec_role	end_date	user_role

4. All extracted dates were shifted to a given time interval to further remove context which could lead to identification of patient data;

5. For text fields (like clinical notes and discharge summaries) we anonymized names of patients and professionals, specific numbers like chart, hospitalization, telephone, bed numbers.

C.3 Results

Regarding the "mapping the process" step (item 2 of the Methods section), several iterations were required to ensure we understood health professionals correctly and vice versa. For us it was a challenge to identify the commonalities and differences in treatment of different patient groups, as defined per severity, age, or list of comorbidities.

The "identification of tables and fields of the database" (item 3) was performed by 10 developers. All of them filled the module oriented tab. Only one developer filled both tabs. At the end we could successfully fill all cells of the spreadsheet. For this step, it was fundamental to have a single contact point to orchestrate the work.

Regarding the "extraction of data" step (item 4), we extracted 4,516 sepsis hospital encounters for a period of two years. We also extracted all (not only sepsis) 61,260 hospital encounters for a period of 2 months, to collect information regarding the workload of professionals (to answer our fourth research question). All the information is present in 57 tables and more than 600 fields.

C.4 Discussion

Mapping the sepsis processes was not an easy task, mainly because the communication between two different teams (health and information technology) is very challenging. The use of activity cards (cards filled by the hospital staff containing questions regarding each activity of the process; e.g. step description, notification process, tools used) helped us to understand better the processes. In addition, when validating the processes, the hospital staff could easily understand the BPMN notation (after an explanation of its elements). Thus, the BPMN models were important tools in the communication process between our team and the hospital staff.

Regarding the "identification of tables and fields of the database" step (item 3), all developers filled the module oriented tab. In our understanding,

the process oriented tab was difficult for the development team to work with, since they had to search in all spreadsheet for the fields that they were responsible for. Indeed, for one step a clinical user may have to work in multiple modules, meaning that the attributes would be distributed over the system; and the development team is organized in such a way that sub-teams are responsible for individual modules. We believe that the module oriented tab helped them to easily identify the required fields.

It is important to mention that it was only possible to convert the process oriented view to the module oriented view since we had access to the hospital information system and we had shadowed clinical users during their use of the system. In addition, some of our researchers had previous experience in HIS architecture.

Columns F, G, H and I from the spreadsheet are used to collect the name of the fields regarding the time and user role that executed and registered an action. Professionals that execute actions are not necessarily the professionals who document them. These columns are very important for process mining research and they could lead to interesting insights when it is performed analyses using performance checking or social networks techniques. The executed time and user role might not always be available, and then the best alternative for process mining is to use the information of the registration of data, however taking into account the assumption made in subsequent analyses.

The extraction of non-structured fields is very important for process mining research. These fields may have some information that can be converted to process activities or they can help to find answers for some questions. For example, when a deviation is discovered in a treatment process it is important to understand why certain cases took this unexpected path. Some answers can be found studying clinical notes discovering for example, that patients that followed this deviation were in a critical stage of the disease. Friedrich et al. [11] discussed that it is estimated that 85% of the information of companies is stored in non-structure format and this source of information can be important for creating models. Most of process mining techniques need a very structured event log as input, and the use of natural language processing (NLP) techniques takes an important role for process mining research since it allows to convert unstructured data to concrete events.

Performing this work we could understand that to extract data from a HIS system is not an easy task: a lot of pre-work, knowledge about the treatment process but also about the HIS and its use, and collaboration with multi-disciplinary teams are needed. In general, the complete extraction

process took us more time than we expected. Below we present the main reasons:

1. The HIS was not originally designed with the sepsis management pathway, and therefore it was a challenge for us to collect all the information that is relevant for the pathway and to identify where it was located in the system;

2. Limited time availability of the development team to support this initiative;

3. We had to deal with challenges to understanding the data structure of the system;

4. We had to guarantee that all patient and hospitalization information was properly anonymized by our scripts (especially in the non-structured fields);

5. We had limited time to run the scripts to not compromise the use of the system by the hospital.

Our initial attempts of process mining analyses encompassed the evaluation of conformance of a simple sepsis process [12]. It took us great effort to check the quality of the data and to create a simple event log, however this opened up subsequent analyses. Later we started a more complete analysis and we could identify the AS-IS sepsis treatment process in the emergency department, identify deviations, and identify bottlenecks. During this investigation, we had to deal with many challenges (e.g. creating specific events, filtering the right cases, dealing with missing timestamps) and it was very important to have the close participation of health professionals. The results of the research study are very promising. The next step is to validate the results with the hospital staff.

The investigation of healthcare processes is far from a trivial task. Healthcare processes tend to be very flexible and complex. Healthcare professionals play an important role providing health information, analyzing and interpreting results, getting insights and guiding to new analyses. Data scientists know methods and tools to organize and consolidate data in a proper way, and they can propose solutions to optimize resources and improve processes. To reconstruct, analyze and improve healthcare processes it is important to use smart approaches combining informatics/engineering techniques with healthcare knowledge. That means that data scientists and health professionals should work always together in all stages of a research to provide

meaningful results. This approach should also be applied in the extraction phase.

We believe that the extraction steps we presented in this work (applying specific adjustments) could be applicable to other healthcare fields of research that uses, for example, data mining, simulation and neural networks techniques.

The extraction phase requires a lot of attention, active and clear communication with external teams, to guarantee that the extracted data will have quality and will allow to perform the research correctly. Getting the wrong data will result in wrong results for the entire research ("garbage in, garbage out" as they say).

C.5 Conclusions

With this work, we share our experience in extracting data from a hospital information system to perform process mining research. Any errors made in the extraction phase will have implications on subsequent analyses, thus it is essential to devote a great deal of attention in this phase, even if it is time and resource-intensive to perform all activities, with the goal of ensuring high quality of the extracted data.

C.6 Acknowledgements

The authors thank all professionals from the hospital and the HIS development team for supporting us in the extraction phase.

The authors thank CAPES and Philips Research for funding the development of this work.

C.7 References

[1] W. van der Aalst, Process Mining: Data Science in Action, Springer Berlin Heidelberg, Berlin, Heidelberg, 2016. doi:10.1007/978-3-662-49851-4.

[2] Process Mining Group - Math& CS department - Eindhoven University of Technology, Process Mining, (2016). http://www.processmining.org/ (accessed December 11, 2016).

[3] A.O. García, D.P. Alfonso, and O.U.L. Armenteros, Analysis of Hospital Processes with Process Mining Techniques, in: MedInfo, 2015, pp. 310–314.

doi:10.3233/978-1-61499-564-7-310.

[4] E. Rojas, J. Munoz-Gama, M. Sepúlveda, and D. Capurro, Process mining in healthcare: A literature review, J. Biomed. Inform. 61 (2016), 224–236. doi:10.1016/j.jbi.2016.04.007.

[5] W. Yang, and Q. Su, Process mining for clinical pathway: Literature review and future directions, in: 2014 11th Int. Conf. Serv. Syst. Serv. Manag., IEEE, 2014, pp. 1–5. doi:10.1109/ICSSSM.2014.6943412.

[6] D. Forsberg, B. Rosipko, and J.L. Sunshine, Analyzing PACS Usage Patterns by Means of Process Mining: Steps Toward a More Detailed Workflow Analysis in Radiology, J. Digit. Imaging. 29 (2016), 47–58. doi:10.1007/s10278-015-9824-2.

[7] P. Rattanavayakorn, and W. Premchaiswadi, Analysis of the social network miner (working together) of physicians, in: 2015 13th Int.Conf. ICT Knowl. Eng. (ICT Knowl. Eng. 2015), IEEE, 2015, pp. 121–124. doi:10.1109/ICTKE.2015.7368482.

[8] R. Mans, H. Schonenberg, G. Leonardi, S. Panzarasa, A. Cavallini, S. Quaglini, and W. van der Aalst, Process mining techniques: an application to stroke care, Stud. Health Technol. Inform. 136 (2008), 573–578. http://www.ncbi.nlm.nih.gov/pubmed/18487792.

[9] Z. Huang, X. Lu, H. Duan, and W. Fan, Summarizing clinical pathways from event logs, J. Biomed. Inform. 46 (2013), 111–127.

[10] R.S. Mans, M.H. Schonenberg, M. Song, W.M.P. van der Aalst, and P.J.M. Bakker, Application of Process Mining in Healthcare – A Case Study in a Dutch Hospital, in: BIOSTEC, Springer, 2008, pp. 425–438. doi:10.1007/978-3-540-92219-3_32.

[11] F. Friedrich, J. Mendling, and F. Puhlmann, Process model generation from natural language text, in: Int. Conf. Adv. Inf. Syst. Eng., 2011, pp. 482–496.

[12] G.-J. de Vries, R.A.Q. Neira, G. Geleijnse, P. Dixit, and B.F. Mazza, Towards Process Mining of EMR Data - Case Study for Sepsis Management,

in: BIOSTEC, 2017.

D
Research validation questionnaire

D.1 - Participant Questionnaire

Each participant filled out the questionnaire.

List of questions of the participant questionnaire

Code	Question
Q1	For how long have you worked in this hospital?
Q2	In which hospital department do you work?
Q3	For how long have you worked in the current department?
Q4	What is your role?
Q5	Do you work with the sepsis clinical pathway in your current department?
Q6	What is your role in the sepsis clinical pathway?
Q7	For how long have you worked with the sepsis clinical pathway in your current department?

D.2 - Deviation Analysis

Deviation: Registry of clinical notes after the prescription of the therapy
Number of cases: 390 (22.81% of cases)

1. Is this a true deviation?	☐ yes ☐ no					
2. If not, why not?						
If the answer to question 1 is Yes then continue to Item 3. Otherwise, move to the next deviation						
3. What is the reason for this deviation?						
4. How do you rate this deviation *Explain that a deviation is not necessarily something bad. E.g., the deviation can improve the patient safety. Alternatively, the deviation can speed up the administration of medications.*	--	-	+/-	+	++	
	☐ Very bad	☐	☐	☐	☐ Very good	
	Why?					
5. What actions would you recommend regarding this deviation?						

Deviation: Blood culture requested by nurse technician
Number of cases: 130 (7.6% of cases)

1. Is this a true deviation?	☐ yes ☐ no					
2. If not, why not?						
If the answer to question 1 is Yes then continue to Item 3. Otherwise, move to the next deviation						
3. What is the reason for this deviation?						
4. How do you rate this deviation *Explain that a deviation is not necessarily something bad. E.g., the deviation can improve the patient safety. Alternatively, the deviation can speed up the administration of medications.*	--	-	+/-	+	++	
	☐ Very bad	☐	☐	☐	☐ Very good	
	Why?					
5. What actions would you recommend regarding this deviation?						

Deviation: Cases without prescription of volume expansion
Number of cases: 1,523 (89% of cases)

1.	Is this a true deviation?	□ yes □ no
2.	If not, why not?	
If the answer to question 1 is Yes then continue to Item 3. Otherwise, move to the next deviation		
3.	What is the reason for this deviation?	
4.	How do you rate this deviation *Explain that a deviation is not necessarily something bad. E.g., the deviation can improve the patient safety. Alternatively, the deviation can speed up the administration of medications.*	-- - +/- + ++ □ □ □ □ □ Very bad Very good Why?
5.	What actions would you recommend regarding this deviation?	

Deviation: Antibiotic was not administrated until 1 hour since sepsis identification
Number of cases: 65 (3.8% of cases). 16 cases with delay in the administration and 49 cases without administration

1.	Is this a true deviation?	□ yes □ no
2.	If not, why not?	
If the answer to question 1 is Yes then continue to Item 3. Otherwise, move to the next deviation		
3.	What is the reason for this deviation?	
4.	How do you rate this deviation *Explain that a deviation is not necessarily something bad. E.g., the deviation can improve the patient safety. Alternatively, the deviation can speed up the administration of medications.*	-- - +/- + ++ □ □ □ □ □ Very bad Very good Why?
5.	What actions would you recommend regarding this deviation?	

General Questionnaire

1.	How important do you think it is for the hospital to have access to this kind of information about deviations?	-- □ Little important	- □	+/- □	+ □	++ □ Very important
2.	Why?					

D.3 - Bottleneck Analysis

Patients waiting in the reception before triage: mean of 17.74/18.58 minutes (without outliers/with outliers)

1. Is this a true bottleneck for you?	□ yes □ no
2. If not, why not?	
3. What is the reason for the bottleneck?	
4. What would you recommend the hospital do to reduce/remove this bottleneck?	

Prescription of medications and request of exams: mean of 4.83/6.77 minutes (without outliers/with outliers)

1. Is this a true bottleneck for you?	□ yes □ no
2. If not, why not?	
3. What is the reason for the bottleneck?	
4. What would you recommend the hospital do to reduce/remove this bottleneck?	

General Questionnaire

1.	Have you identified other bottlenecks in the process that were not pointed out in our research?					
2.	How important do you think it is for the hospital to have access to performance information like the ones presented above?	-- □ Little important	- □	+/- □	+ □	++ □ Very important
3.	Why?					

D.4 - Solution for improving the process

Prescription before registration of clinical notes
Reduction in 3.4 min (median 3 min) – without outliers
Reduction in 4.9 min (median 4 min) – with outliers

1. Do you believe that it makes sense to implement this deviation in the sepsis clinical pathway of the hospital?	-- □ It makes no sense	- □	+/- □	+ □	++ □ It makes all sense
2. Why?					
3. Are these results associated to urgent patients?	□ yes □ no				

E
Multi-criteria analysis example: loan request process

In this appendix we describe in detail the use case of the loan request process to exemplify the multi-criteria analysis technique. The example was inspired from the process used in the Business Process Intelligence Challenge 2017 [BPIC, 2017], and its main purpose is to help the reader to better understand the process mining multi-criteria technique presented in Chapter 4.

E.1
Use case description

Figure 4.2 (Chapter 4) presents the BPMN normative model of a simple loan request.

The process starts when a customer requests a loan to the bank. The customer can ask for any amount of money for a specific target (e.g. to buy a house or a car). The customer can start the process accessing a website (activity A) or in person by going to the bank (activity B). Once the request is formalized, the customer needs to provide a set of required documents (e.g. birth certificate, proof of income) (activity D). If the costumer takes more than one week to perform the previous activity, the bank calls them to remember to send their documents (activity C). At the moment the institution receives all documentation, the bank validates the loan request (activity E), approving (activity F) or rejecting it (activity G). During activities C, D and E, the customer can cancel the loan request (activity H). The process finishes after the execution of activities F, G or H.

The bank acts in several countries and, in our example, the quality team wants to get insights of which actions can improve the process for the subsidiaries located in the Netherlands. The bank wants to get recommendations to improve their process to reduce the case duration and increase the profit for approved cases.

E.2 Multi-criteria analysis

In this section we will apply the Multi-criteria technique for the loan example. We will follow the same steps presented in Chapter 4.

E.2.1 Step 1 - Data collection

The bank extracted data from 30 different tables from their system and created the event log using the same activities presented in Figure 4.2 (Chapter 4). Table E.1 presents the attributes of the event log.

Table E.1: Attributes from the event log.

Attribute	Description	Type	Format
Case ID	Loan request code	Numeric	NNNNNNNN
Activity	Executed activity	Text	
Completion date	Date that the execution of the activity was finished	Time-stamp	YYYY-MM-DD HH24:MI:SS
Case approved	Case was approved or not	Boolean	0 for approved, 1 for not approved
Case canceled	Case was canceled or not	Boolean	0 for canceled, 1 for not canceled
Profit	Total profit value regarding the loan request	Numeric	NNNNNNNNN.N
Total time	Total time in days until the execution of the current activity. The value from the last executed event is the case duration.	Numeric	NNNN

E.2.2 Step 2 - Case selection

The bank selected only completed and approved cases (that performed the activities A or B and, F) executed in the Netherlands for the period from 2016 to 2018.

E.2.3 Step 3 - Definition of variants

The process was executed in 20 different ways, and as consequence, 20 different groups of cases were created. Table E.2 presents details of all variants.

Table E.2: Variants from the event log.

Variant	Sequence of Activities	Frequency (Number of Cases)	Mean Case Duration (Days)	Mean Profit (€)
1	A >C >D >E >F	1,500	25	3,289
2	B >E >F	149	11	3,714
3	B >C >E >F	700	13	3,950
4	A >C >C >C >D >E >F	10	30	3,190
5	A >C >D >E >F >F	16	31	3,332
6	B >D >E >C >E >F	2	17	3,928
7	B >B >D >E >F	7	28	3,971
8	A >C >E >F	56	28	3,352
9	A >D >E >F	2,100	24	3,399
10	A >A >D >E >F	250	25	3,403
11	B >D >E >F	1,722	12	3,501
12	A >D >E >C >E >F	15	27	3,515
13	A >C >C >D >E >F	65	27	3,631
14	B >C >C >E >F	16	15	3,816
15	B >C >C >C >D >E >F	8	38	3,991
16	B >C >D >E >F >F	13	23	4,072
17	A >E >F	155	24	3,259
18	A >C >C >E >F	94	37	3,314
19	B >C >C >D >E >F	49	17	4,221
20	B >C >D >E >F	1,122	15	4,249

E.2.4 Step 4 - Criteria definition

The bank defined two simultaneous criteria. Table E.3 presents details regarding the criteria definition.

Table E.3: Set of criteria defined by the bank for the analysis.

Priority	Attribute	Target	Type of Measure
Extremely important	Profit	Maximize	Mean
Above average	Total time	Minimize	Mean

E.2.5 Step 5 - Data treatment

The bank defined the interquartile rule for removing outliers and selected the Min-Max method to normalize the values.

Variants 1, 5, 7, 8, 9, 13, 15, 16, 17, 18, 19 and 20 had some cases removed since they had outlier values for case duration or for profit. Table E.4 presents the frequency, mean case duration and mean profit for the variants after removing the outliers. Later, variants 4, 5, 6, 7, 12, 13, 14, 15, 16 and 19

Table E.4: Detail of variants after removing outlier values.

Variant	Sequence of Activities	Frequency (Number of Cases)	Mean Case Duration (Days)	Normalized Mean Case Duration (Min-Max)	Mean Profit (€)	Normalized Mean Profit (Min-Max)
1	A >C >D >E >F	1,480	25	0.8816	3,333	0.1618
2	B >E >F	149	11	0.0616	3,714	0.5242
3	B >C >E >F	700	13	0.1747	3,950	0.7491
4	A >C >C >D >E >F	10	30	-	3,190	-
5	A >C >D >E >F >F	15	27	-	3,332	-
6	B >D >E >C >E >F	2	17	-	3,928	-
7	B >B >D >E >F	6	15	-	3,968	-
8	A >C >E >F	51	23	0.7693	3,353	0.1812
9	A >D >E >F	2,065	23	0.7655	3,399	0.2244
10	A >A >D >E >F	250	25	0.8889	3,403	0.2288
11	B >D >E >F	1,722	12	0.1463	3,501	0.3217
12	A >D >E >C >E >F	15	27	-	3,515	-
13	A >C >C >D >E >F	17	27	-	3,010	-
14	B >C >C >E >F	16	15	-	3,816	-
15	B >C >C >C >D >E >F	6	19	-	3,973	-
16	B >C >D >E >F >F	12	17	-	4,078	-
17	A >E >F	150	22	0.7043	3,260	0.092
18	A >C >C >E >F	70	25	0.8941	3,299	0.1295
19	B >C >C >D >E >F	22	17	-	4,168	-
20	B >C >D >E >F	379	15	0.296	4,182	0.9697

were removed from the analysis as the bank defined a minimal number of 50 cases per variant.

All attribute values from the remaining cases were normalized using the Min-Max method. Table E.4 shows the mean normalized value for case duration and profit.

E.2.6 Step 6 - Valuation of variants

For the analysis, the following formula was used for calculating the UV:

$$\frac{4 \times mean(normalization(total\_time))}{8 \times mean(normalization(profit))} \quad \text{(E-1)}$$

Table E.5 presents the UV per variant. As this is a minimization problem, the smaller is the UV, the better outcomes the variant provides.

Table E.5: Unique Values per variant.

Variant	Sequence of Activities	Unique Value
1	A >C >D >E >F	2.7249
2	B >E >F	0.0587
3	B >C >E >F	0.1166
8	A >C >E >F	2.1234
9	A >D >E >F	1.7057
10	A >A >D >E >F	1.9426
11	B >D >E >F	0.2274
17	A >E >F	3.8269
18	A >C >C >E >F	3.4512
20	B >C >D >E >F	0.1526

E.2.7
Step 7 - Cluster of variants

For the analysis, two clusters were created. Table E.6 presents the clustering results of the variants and the UV per cluster. Cluster 1 was the one with the best outcomes with a UV of 0.1626. The cluster is composed of variants 2, 3, 20 and 11.

Table E.6: Clustering of variants. Variants are sorted by the variant UV.

Variant	Sequence of Activities	Variant Unique Value	Cluster	Cluster Unique Value
2	B >E >F	0.0587	1	0.1626
3	B >C >E >F	0.1166		
20	B >C >D >E >F	0.1526		
11	B >D >E >F	0.2274		
9	A >D >E >F	1.7057	2	2.0927
10	A >A >D >E >F	1.9426		
8	A >C >E >F	2.1234		
1	A >C >D >E >F	2.7249		
18	A >C >C >E >F	3.4512		
17	A >E >F	3.8269		

E.2.8
Step 8 - Simplification of variants

The bank decided to convert repeated sequences of activities to *sub-sequences blocks* and *permutation sub-sequence blocks* to simplify their analysis.

The only *sub-sequence block* that was created was the [E > F], since the sub-sequence "E > F" was present in all variants. No *permutation sub-sequence block* was created. Thus, all sub-sequences "E > F" were replaced to the [E > F] block in all variants. Table E.7 presents the variants after their simplification.

Table E.7: Variants after their simplification. The [E > F] is a *sub-sequence block*.

Variant	Sequence of Activities
2	B >[E >F]
3	B >C >[E >F]
20	B >C >D >[E >F]
11	B >D >[E >F]
9	A >D >[E >F]
10	A >A >D >[E >F]
8	A >C >[E >F]
1	A >C >D >[E >F]
18	A >C >C >[E >F]
17	A >[E >F]

E.2.9
Step 9 - Comparison of clusters and identification of insights

Cluster 1 was compared to cluster 2 (the cluster that contains variants with worse outcome results than cluster 1). Table E.8 presents the differences between both clusters. For example, the execution of (column A) activity B, and sub-sequences "B > C", "B > D", "B > [E > F]" promoted positive outcomes (column B).

Table E.8: Differences from clusters 1 and 2.

Execution of (A)	**Type of Outcome (B)**
Activity B	Positive
Sub-sequence B >C	Positive
Sub-sequence B >D	Positive
Sub-sequence B >[E >F]	Positive
Activity A	Negative
Sub-sequence A >A	Negative
Sub-sequence A >C	Negative
Sub-sequence A >D	Negative
Sub-sequence A >[E >F]	Negative
Sub-sequence C >C	Negative

Table E.9 presents the *Levenshtein distance between variants* matrix. The matrix indicates that the worst variant (10, with UV 3.8269) has only one difference (distance of 1) to the best variant (1, with UV 0.0587). That means that the substitution of activity A to B probably would improve significantly the outcomes.

Table E.9: Levenshtein distance between variants matrix.

Variant	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**UV**
1	0	1	2	1	2	3	2	3	3	1	0.0587
2	1	0	1	1	2	3	1	2	2	2	0.1166
3	2	1	0	1	2	2	2	1	2	3	0.1526
4	1	1	1	0	1	2	2	2	3	2	0.2274
5	2	2	2	1	0	1	1	1	2	1	1.7057
6	3	3	2	2	1	0	2	1	2	2	1.9426
7	2	1	2	2	1	2	0	1	1	1	2.1234
8	3	2	1	2	1	1	1	0	1	2	2.7249
9	3	2	2	3	2	2	1	1	0	2	3.4512
10	1	2	3	2	1	2	1	2	2	0	3.8269

F
Multi-CAT algorithm for the comparison of sub-sequences of directly followed activities

In this appendix we present the algorithm implemented in Multi-CAT for the comparison of sub-sequences of directly followed activities. To explain the algorithm we will use the same example of the two clusters detailed in Table 5.3 (Chapter 5). The algorithm steps are:

1. A directly followed matrix is created for each cluster. Each matrix stores the execution frequency of directly followed activities (sub-sequences). For each variant of a given cluster, the frequencies of all executed sub-sequences (e.g. "A > C", 20 cases) are stored. The directly followed matrix of the Positive cluster from our example is presented in Table F.1 (we denominate it as Matrix Positive), and the directly followed matrix of the Negative cluster is presented in Table F.2 (Matrix Negative). In the example, the sub-sequence "A > B" happened 20 times for the first variant "A > B > C > D", and 10 times for variant "A > B > D". In this way, the cell "A > B" (line A, column B) from Matrix Positive will receive the value 30 (20 + 10);

Table F.1: Matrix Positive: stores the execution frequency of directly followed activities of the Positive cluster.

From/To	A	B	C	D	E
A	0	20 + 10 = 30	20 = 20	0	0
B	0	0	20 = 20	10 = 10	0
C	0	0	0	20 = 20	0
D	0	0	0	0	0
E	0	0	0	0	0

Table F.2: Matrix Negative: stores the execution frequency of directly followed activities of the Negative cluster.

From/To	A	B	C	D	E
A	0	10 = 10	20 + 5 = 25	0	0
B	0	0	0	10 = 10	0
C	0	20 = 20	0	5 = 5	0
D	0	0	0	0	10 = 10
E	0	0	0	0	0

2. Matrix Sum is created as the sum of Matrix Positive and Matrix Negative. Table F.3 presents Matrix Sum;

Table F.3: Matrix Sum: stores the sum of Matrix Positive and Matrix Negative.

From/To	A	B	C	D	E
A	0	40	45	0	0
B	0	0	20	20	0
C	0	20	0	25	0
D	0	0	0	0	10
E	0	0	0	0	0

3. Matrix Positive and Matrix Negative are normalized dividing all their values by Matrix Sum. The resulting values will range from 0 to 1. If the value from a cell from Matrix Sum is equal to zero, then the same cells from the normalized Matrix Positive and Matrix Negative will receive a zero value. Tables F.4 and F.5 present the normalized Matrix Positive_n and Matrix Negative_n;

Table F.4: Matrix Positive_n: stores normalized values from Matrix Positive.

From/To	A	B	C	D	E
A	0	0.75	0.44	0	0
B	0	0	1.00	0.50	0
C	0	0	0	0.80	0
D	0	0	0	0	0
E	0	0	0	0	0

Table F.5: Matrix Negative_n: stores normalized values from Matrix Negative.

From/To	A	B	C	D	E
A	0	0.25	0.56	0	0
B	0	0	0	0.50	0
C	0	1.00	0	0.20	0
D	0	0	0	0	1.00
E	0	0	0	0	0

4. Matrix Diff is created by subtracting Matrix Negative_n from Matrix Positive_n. The resulting values will range from -1 to 1. Table F.6 presents Matrix Diff and it represents the difference from both clusters. Results with value +1 in Matrix Diff indicate sub-sequences that are only present in the Positive Cluster and were beneficial in the execution of the process, and results equals to -1 indicate sub-sequences that are only present in the Negative Cluster and were harmful for the process. With the analysis of the results from the example, we can conclude that the sub-sequence "B > C" contributed positively to the process, and that the sub-sequences "C > B" and "D > E" contributed negatively to the process.

Table F.6: Matrix Diff: all elements from Matrix Negative$_n$ are subtracted from Matrix Positive$_n$.

From/To	**A**	**B**	**C**	**D**	**E**
A	0	0.50	-0.11	0	0
B	0	0	1.00	0	0
C	0	-1.00	0	0.60	0
D	0	0	0	0	-1.00
E	0	0	0	0	0

G
Multi-CAT input screen options

Table G.1: Multi-CAT input screen options.

Field	Options
Criterion Parameters	
Attribute	Show all numeric attributes from the event log
Weight	1 - Little importance
	2 - Under average
	4 - Above average
	8 - Extremely important
Direction/target	Minimize
	Maximize
Type of measure	Mean
	Variance
	Standard deviation
Remove outliers?	Yes, I want to remove them
	No, I don't want to remove them
Other Parameters	
Method to remove outliers	5% extremes
	Interquartile rule
	None
Method to normalize	Min-Max
	Ranking
	None
Aggregation type	Additive
	Multiplicative
Minimal N	Minimal variant size defined by the user
Clustering method	Iterative Farthest First
	Iterative K-Means
	Iterative K-Medoids
	Farthest First - 2 Clusters
	Farthest First - 3 Clusters
	Farthest First - 4 Clusters
	Farthest First - 5 Clusters
	Farthest First - 6 Clusters
	Farthest First - 7 Clusters
	K-Means - 2 Clusters
	K-Means - 3 Clusters
	K-Means - 4 Clusters
	K-Means - 5 Clusters
	K-Means - 6 Clusters
	K-Means - 7 Clusters
	K-Medoids - 2 Clusters
	K-Medoids - 3 Clusters
	K-Medoids - 4 Clusters
	K-Medoids - 5 Clusters
	K-Medoids - 6 Clusters
	K-Medoids - 7 Clusters
Presentation of activities	Alias
	Number
	Number and alias
Group of sub-sequences	Sequence without permutation
	Sequence with permutation
	None

H
Parameters from PLG tool used to generate the nine event logs for the performance tests

Table H.1: Parameters from PLG tool used to generate the nine event logs.

Parameter	Value
Number of traces	1,000, 10,000 or 50,000
Noise	
Trace missing head probability	5%
Trace missing tail probability	5%
Trace missing episode probability	5%
Perturbed event order probability	5%
Double event probability	5%
Integer data object error probability	5%
Noise configuration	
Head max. size	2
Tail max. size	2
Episode max. size	2
Integer data object error delta	1

I
Results from Multi-CAT performance tests

Table I.1 shows the complete results from the Multi-CAT performance tests. In the table, each event log is a combination of a line (A, B, C) with a column (1, 2, 3).

Table I.1: Results from the performance tests.

Number of cases	Evaluation parameter	Step	Process Complexity		
			Low (1)	Medium (2)	High (3)
	Event log ID		**A**	**D**	**G**
1,000 (A)	Sort event log time (ms)	3	0	3	15
	Group variants time (ms)	3	15	15	69
	Outliers removal time (ms)	5	0	16	0
	Removal of infrequent cases time (ms)	5	0	0	0
	Normalization of values time (ms)	5	16	0	0
	Valuation of variants time (ms)	6	0	0	0
	Clustering time (ms)	7	0	0	0
	Valuation of clusters time (ms)	7	2	18	13
	Sub-sequence block creation time (ms)	8	0	3	1
	Sub-sequence permutation block creation time (ms)	8	0	46,445	3,384
	Differences of clusters time (ms)	9	0	7	21
	Update name of activities time (ms)	9	0	1	1
	Levenshtein Distance matrix creation time (ms)	9	0	0	1
	Calculation of Global UV time (ms)	-	1	0	0
	Total processing time (ms)	-	88	46,571	3,574
	ProM Interface processing time (ms) (Visualize HTML text)	-	13	135	165
	Smallest variant size	-	3	10	23
	Largest variant size	-	5	14	31
	Number of variants	-	3	7	8
	Event log ID		**B**	**E**	**H**
10,000 (B)	Sort event log time (ms)	3	0	47	62
	Group variants time (ms)	3	100	359	701
	Outliers removal time (ms)	5	15	12	16
	Removal of infrequent cases time (ms)	5	0	0	0
	Normalization of values time (ms)	5	23	9	15
	Valuation of variants time (ms)	6	0	3	0
	Clustering time (ms)	7	0	10	0
	Valuation of clusters time (ms)	7	5	6	13
	Sub-sequence block creation time (ms)	8	0	1	72
	Sub-sequence permutation block creation time (ms)	8	0	40,192	16,634
	Differences of clusters time (ms)	9	0	15	20
	Update name of activities time (ms)	9	0	11	1
	Levenshtein Distance matrix creation time (ms)	9	0	1	0
	Calculation of Global UV time (ms)	-	5	3	7
	Total processing time (ms)	-	205	40,805	17,664
	ProM Interface processing time (ms) (Visualize HTML text)	-	19	271	7031
	Smallest variant size	-	2	10	23
	Largest variant size	-	5	18	43
	Number of variants	-	4	11	69
	Event log ID		**C**	**F**	**I**
50,000 (C)	Sort event log time (ms)	3	100	116	317
	Group variants time (ms)	3	515	1,136	7,320
	Outliers removal time (ms)	5	84	93	69
	Removal of infrequent cases time (ms)	5	0	0	3
	Normalization of values time (ms)	5	94	22	27
	Valuation of variants time (ms)	6	22	16	9
	Clustering time (ms)	7	16	15	29
	Valuation of clusters time (ms)	7	29	20	27
	Sub-sequence block creation time (ms)	8	0	0	178
	Sub-sequence permutation block creation time (ms)	8	0	228	46,633
	Differences of clusters time (ms)	9	0	0	97
	Update name of activities time (ms)	9	0	0	11
	Levenshtein Distance matrix creation time (ms)	9	0	0	130
	Calculation of Global UV time (ms)	-	21	31	21
	Total processing time (ms)	-	977	1,820	55,147
	ProM Interface processing time (ms) (Visualize HTML text)	-	155	229	67,413
	Smallest variant size	-	1	8	23
	Largest variant size	-	6	18	51
	Number of variants	-	15	14	200