



A methodology for business process discovery and diagnosis based on indoor location data: Application to patient pathways improvement

Sina Namaki Araghi

► To cite this version:

Sina Namaki Araghi. A methodology for business process discovery and diagnosis based on indoor location data: Application to patient pathways improvement. Other [cs.OH]. Ecole des Mines d'Albi-Carmaux, 2019. English. NNT : 2019EMAC0014 . tel-03037128

HAL Id: tel-03037128

<https://theses.hal.science/tel-03037128>

Submitted on 3 Dec 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



THÈSE

en vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

délivré par

IMT – École Nationale Supérieure des Mines d'Albi-Carmaux

présentée et soutenue par

Sina NAMAKI ARAGHI

le 12 novembre 2019

A methodology for business process discovery and
diagnosis based on indoor location data:
Application to patient pathways improvement

École doctorale et discipline ou spécialité :

EDSYS : Génie industriel et informatique

Unité de recherche :

Centre Génie Industriel, IMT Mines Albi

Directeurs de thèse :

Frédéric Bénaben, Professeur, IMT Mines Albi (Directeur)

Franck Fontanili, Maître-assistant, IMT Mines Albi (Co-directeur)

Autres membres du jury :

Thomas Lux, Professor, Hochschule Niederrhein University of Applied Sciences Allemagne
(Rapporteur)

François Charoy, Professeur, Université de Lorraine (Rapporteur)

Maria Di Mascolo, Directrice de Recherche, Laboratoire G-SCOP Grenoble (Présidente)

Chihab Hanachi, Professeur, Université Toulouse 1 Capitole (Examineur)

Elyes Lamine, Maître de conférences, Institut National Universitaire Champollion, Albi (Examineur)

Mário José Monteiro de Macedo, Associate Professor, Univ. Nova de Lisboa Portugal (Examineur)

François Babin, CEO and Co-founder, MAPLE High Tech Toulouse, (Invité)

Abstract

Business processes are everywhere and, as such, we must acknowledge them. Among all of them, hospital processes are of vital importance. Healthcare organizations invest huge amount of efforts into keeping these processes under control, as the allowed margin of error is so slight.

This research work seeks to develop a methodology to endorse improvement of patient pathways inside healthcare organizations. It does so by using the indoor location data of patients.

This methodology is called **DIAG** (Data state, Information state, Awareness, Governance). It is constructed of several different functions. The most important ones are as follows: (i) location data interpreting, (ii) automatic discovery of business process models, (iii) business process analyzing for evaluating the performance and quality of processes, and finally, (iv) automatic diagnosing of business processes.

Along the former functions, the contribution of this thesis are:

- (i) the DIAG methodology which, through four different states, extracts knowledge from location data;
- (ii) the DIAG meta-model which supports both the interpretation of location data (from raw data to usable information) and the alignment of the domain knowledge (which are used for the diagnosing methods);
- (iii) two process discovery algorithms which explore statistical stability in event logs.
- (iv) application of Statistical Process Control (SPC) for the “enhancement notation” of Process Mining;
- (v) the ProDIST algorithm for measuring the distance between process models;
- (vi) two automatic process diagnosing methods to detect causes of structural deviations in individual cases and common processes;

The state of the art in this dissertation endorses the necessity for proposing such solutions. A case study within this research work illustrates the applicability of the DIAG methodology and its mentioned functions and methods.

Keywords: Process Mining, Indoor Localization Systems, Business Process Management, Business Process Diagnosis, Healthcare Processes

Resumé

Dans chaque organisation, les processus métier sont aujourd'hui incontournables. Cette thèse vise à développer une méthode pour les améliorer. Dans le domaine de la santé, les organisations hospitalières déploient beaucoup d'efforts pour mettre leurs processus sous contrôle, notamment à cause de la très faible marge d'erreur admise. Les parcours des patients au sein des structures de santé constituent l'application qui a été choisie pour démontrer les apports de cette méthode. Elle a pour originalité d'exploiter les données de géolocalisation des patients à l'intérieur de ces structures. Baptisée **DIAG**, elle améliore les parcours de soins grâce à plusieurs sous-fonctions :

(i) interpréter les données de géolocalisation pour la modélisation de processus, (ii) découvrir automatiquement les processus métier, (iii) évaluer la qualité et la performance des parcours et (iv) diagnostiquer automatiquement les problèmes de performance des processus. Cette thèse propose donc les contributions suivantes :

- (i) la méthode DIAG elle-même qui, grâce à quatre différents états, extrait les informations des données de géolocalisation ;
- (ii) le méta-modèle DIAG qui a deux utilités : d'une part, interpréter les données de géolocalisation et donc passer des données brutes aux informations utilisables, et, d'autre part contribuer à vérifier l'alignement des données avec le domaine grâce à deux méthodes de diagnostic décrites plus bas ;
- (iii) deux algorithmes de découverte de processus qui utilisent la stabilité statistique des logs d'événements ;
- (iv) une nouvelle approche de process mining utilisant SPC (Statistical Process Control) pour l'amélioration ;
- (v) l'algorithme proDIST qui mesure les distances entre les modèles de processus ;
- (vi) deux méthodes de diagnostic automatique de processus pour détecter les causes des déviations structurelles dans des cas individuels et pour des processus communs.

Le contexte de cette thèse confirme la nécessité de proposer de telles solutions. Une étude de cas dans le cadre de ce travail de recherche illustre l'applicabilité de la méthodologie DIAG et des fonctions et méthodes mentionnées.

Keywords: Process Mining, Systèmes de localisation en intérieur, Gestion des processus métiers, Diagnostic des processus métiers, Processus de soins

Cette thèse a été co-financée par la Région Occitanie et la société MAPLE High Tech. Nous les remercions pour leur soutien.

Acknowledgement

I'm lingering on this page far too long and thinking how to start. There is no perfect start, and it doesn't matter how I begin it. What matters is to persevere from start to finish. One of the most blissful times of my life, the journey, the 4 AM's, the late nights. Understanding why and how to do this and many other crazy circumstances to deal with. It required a state of mind—mental toughness—to not get too high or too low no matter what. I learned the hard way to focus on the task in hand, and I have several persons to thanks for it. First and foremost I have to thank my basketball coaches for directing me into a path of grasping the idea of hard work. They showed me the fact that hard work beats talent when the talent doesn't work hard. They helped me to understand the difference between being confident and being cocky.

Looking back at the day I started this Ph.D. life, I can not thank enough my team of directors. Franck who trusted me and sat by my side and taught me many many and many techniques and approaches to analyze business processes. Franck was the person who introduced me to process mining and made me fall in love with this field, thank you Franck.

Elyes, you had always time for my questions and my concerns even though you were always busy with your schedule. Thank you Elyes for all the encouragements you gave me to overcome my doubts.

The great Jim Valvano once said the greatest gift anyone can give another person is believing in him/her. Well, I was lucky enough to get that gift from my thesis director. Fred thank you for all those meetings that we had to build upon what has become this thesis. It wouldn't be possible without those small pep talks you gave me each time that you came to my office.

I'd like to thank other jury members of my thesis for their time and insightful comments. I was inspired by the way you've conducted your research and honored to defend my dissertation in front of you.

My friends and colleagues who survived or are surviving their Ph.D. journey, Jiang, Andres, Laura, Shadan, Rafael, Audrey, Guillaume, Julien, Eva, Aurélie.C, Aurélie.M, Ibrahim, Abdou, Delphine, Quentin, Manon.F, Manon.G, Robin, Jiyao, Alex, Rania and others who I may forget due to writing a thesis distortion phenomenon, thank you all. I have saved a special thanks for Nicolas Salatge, Sébastien Rebière, Paul Gaborit, and Julien Lesbegueries who supported me with the technical stuff of my thesis as well as tennis and badminton sessions.

Isa I'd never forget you, thank you for ... thank you for just being Isa. Rest in power.

Yass during these three years you managed, helped, encouraged me, and simply were there for me always. Thanks for being patient with my crazy drive to finish this thing.

My parents, I can not thank you enough for guiding and endorsing me all the way. Dad, with 4000 hours of flying beyond enemy lines you showed me fighters fight no matter what. Mom, you taught me that normal people do the normal thing, and for me to not be normal, I have to go to the extreme at what I do. My nemesis and most lovable persons of my life, my brothers and sister. Houman, thank you for setting the bar so high that no one can reach it. Saba, you are the kindest person I ever have known in my life and you know at the bottom of my heart I love you more than pasta. Honey, nothing is possible nor manageable without you being there and of course, you are the joy of our lives. For those little ones of yours, Shamim, Gabriel, Kimia, and Romi, I love you more than anything and I hope to stay your cool uncle.

Preface

I believe that each of us as a part of the society deserve to receive the best possible service or product; in essence, I do believe that receiving an error-free healthcare is a human right and not a privilege for moneyed people.

I deem that every organization—hospitals in particular—have visions to obtain the best possible outcomes from their operations.

Patient pathways is an important topic and a riddle to be mastered by healthcare organizations. Many patients start their pathways in hospitals and unfortunately, because of inefficiencies within non-medical operations lose their lives even before seeing a physician.

Association of Indoor Localization Systems (ILS) and Process Mining can provide valuable insights about such inefficiencies.

Our research examined different methodologies, approaches, and methods to make this association possible and efficient.

This dissertation presents a methodology to support data-driven decision making by exploring the real-time location and business process models. The application domain is the healthcare processes and mainly the improvement of patient pathways.

It includes three parts. The first part presents the context, problematic and the background of the research work. The second part presents the mentioned methodology, and the developed function. Finally, the third part illustrates the applicability of this methodology and its corresponding methods by a case study.

Within each chapter important notes and information are highlighted. The complementary information such as, examples and extra descriptions, are presented in gray-colored frames.

The yellow boxes present important decisions taken from the provided information.

I do hope this research work can be an enhancement and a tipping point for the way we experience business processes.

Sina Namaki Araghi

Contents

Abstract	iii
Resumé	v
Acknowledgement	ix
Preface	xi
Contents	xiii

Part I First part: context, problematic, and background of the thesis 1

1 Introduction	3
1.1 Context and the social problematic relevant to patient pathways . .	3
1.1.1 Challenge No. 1 – “Structural”	3
1.1.2 Challenge No. 2 – “Duration & distance”	5
1.2 Focus and scope	6
1.3 Scientific challenges and questions	7
1.3.1 What data science-oriented procedure is suitable for extracting knowledge form location data?	7
1.3.2 How to interpret location data?	8
1.3.3 How to extract a descriptive reference process model?	9
1.3.4 How to acquire a holistic analytic approach for patient pathways?	9
1.3.5 How to measure the distance of a descriptive model from the normative model?	9
1.3.6 How to automatically diagnose the causes of deviations? . .	10
1.3.7 How to generate different scenarios for choosing the optimum solution?	10
1.4 Contribution of this thesis	10
1.5 Outline of the dissertation	11
Important terminology of this dissertation	12
2 State of the art	13
2.1 Introduction	13
2.2 Data Science	14
2.2.1 Data science project requirements	15
2.2.2 Data science in healthcare	16

2.2.2.1	Levels and types of healthcare data	17
2.2.2.2	The challenges of data science in healthcare	17
2.3	Process Mining	19
2.3.1	Where process mining has started?	19
2.3.2	Used modeling languages	20
2.3.3	Process mining in healthcare	22
2.4	Process mining notations (activities) in healthcare	24
2.4.1	Works related to the business process discovery and modeling notation in healthcare	25
2.4.2	Enhancement notation (activity) of business processes in healthcare	28
2.4.3	Conformance checking activity in healthcare	30
2.5	Indoor Localization Systems (ILS)	30
2.6	Data mining and process mining have met location data before: . .	32
2.6.1	Preparation and interpretation of location data	35
2.6.2	Use of location data for knowledge extraction	37

Part II	Second part: DIAG Methodology	41
	DIAG methodology statement	43
	Introduction	43
	Background	43
	Objectives	45
3	From the Data to the Information state	49
3.1	Introduction	49
3.1.1	Why Data state is defined in the DIAG methodology?	50
3.1.2	How the connection between the location event logs and the actual concepts in a process can be constructed?	51
3.2	Function 1: Configuring the environment and systems	52
3.2.1	DIAG meta-model: a framework to support process model discovery from location data	52
3.2.1.1	The Process and Location Event-log packages	54
3.2.1.2	The Function and Healthcare Functions packages	54
3.2.1.3	The Organization, Healthcare Resources, and Objectives packages	56
3.2.1.4	DIAG meta-model (first version)	57
3.2.2	R.IO-DIAG initialization	59
3.3	Function 2: Location data gathering	65
3.3.1	Objects movements	65
3.4	Function 3: Location data interpreting function	69
3.4.1	Add start-event	71
3.4.2	Add end-event	72
3.4.3	Add task-event	73
3.4.4	Add knowledge on start-event	76
3.4.5	Add knowledge on end-event	77
3.4.6	Add knowledge on task-event	78
3.4.7	Operation Process Charts (OPC)	78
3.5	Recap	82

4 The Business process modeling function within the Information state	85
4.1 Introduction	87
4.2 Background	88
4.2.1 Automatic business process discovery-Existing approaches	88
4.2.1.1 Characteristics of process discovery algorithms	89
4.2.2 The classic heuristic miner	91
4.2.3 Statistical stability	92
4.3 Stable Heuristic Miner V1	94
4.3.1 An imaginary hypothesis	94
4.3.2 Preliminaries	94
4.3.2.1 Assumptions	95
4.3.2.2 Definitions and the sequence of functions in the algorithm	96
4.4 Stable Heuristic Miner V2	105
4.4.1 Preliminaries	105
4.4.2 Order of calculation	108
4.4.3 Comparing the results of the algorithms	114
4.5 Recap	118
5 Awareness: analyzing and diagnosing patient pathways	121
5.1 Introduction	122
5.2 Business Process Analyzing Function	125
5.2.1 Pre-processing the information	125
5.2.1.1 Extracting a data table from the observed behaviors	126
5.2.1.2 Sort the data by ID of cases	126
5.2.1.3 Add the duration of activities	126
5.2.1.4 Calculate the total duration and distances of processes	127
5.2.1.5 Generate the second data table for sampling the duration of processes	127
5.2.1.6 Generate the third data table for sampling the distances of processes	127
5.2.2 Statistical process control (SPC) application to support the enhancement activity of process mining	129
5.2.2.1 Control charts	129
5.2.2.2 Process Capability Ratio (C_p)	130
5.2.3 An example: Applying control charts and process capability ratio analyses	132
5.3 Business Process Diagnosing Function	137
5.3.1 Automatic diagnosing approach 1: Miniscule Movements of Processes (MMP)	137
5.3.1.1 Miniscule Movements: An approach to discover planets	137
5.3.1.2 MMP method: Definitions and Hypothesis	139
5.3.1.3 Logic of MMP method	141
5.3.1.4 Deducing GM's (generated models)- DIAG meta-model is revisited	142
5.3.1.5 ProDIST algorithm: a novel method for measuring the distance between two process models	147
5.3.2 Automatic diagnosing approach 2: DIAG method	152
5.3.2.1 An example for the second automatic diagnosing approach	155
5.4 Recap	158
5.4.1 SWOT analysis of the chapter	158

Part III Third part: the experimentation	161
6 A case study: experimental results of the thesis	163
6.1 Motivation	163
6.2 Presentation of the case study	164
6.3 Preparation for interpreting the location data	166
6.4 Discovering and modeling patient pathways	171
6.4.1 Results of discovery algorithms	171
6.5 Analyzing the quality and performance of patient pathways	182
6.6 Automatic diagnosing of patient pathways	188
6.6.1 Miniscule Movements of Processes (MMP) method	188
6.6.2 Application of DIAG method	190
Conclusion	193
6.7 Limitations	196
6.8 Future perspective	196
Résumé étendu en français	199
Contexte et problématique sociales pertinentes pour les parcours des patients	200
Challenge No. 1 - "Structure et organisation"	200
Challenge No. 2 - "Durée et distance"	201
Objet et portée de ce travail	202
Défis et questions scientifiques	202
Quelle procédure orientée science des données convient pour extraire des données de localisation de formulaire de connaissances ?	203
Comment interpréter les données de localisation ?	203
Comment extraire un modèle de processus de référence descriptif ?	203
Comment acquérir une approche analytique exhaustive pour les parcours des patients ?	204
Comment mesurer la distance d'un modèle descriptif par rapport au modèle normatif ?	204
Comment diagnostiquer automatiquement les causes des écarts ?	204
Comment générer différents scénarios pour choisir la solution optimale ?	204
Les contributions de la thèse	205
Aperçu de la thèse	206
Terminologie importante de cette thèse	206
Revue internationale avec comité de sélection	207
Articles soumis (ou en cours de préparation)	207
Les limites	207
Perspectives d'avenir	208
List of Figures	209
List of Tables	215
Bibliography	217

*They ask us why we mutilate
each other like we do, and
wonder why we hold such little
worth for human life.*

*To ask us why we turn from bad
to worse, is to ignore from
which we came.*

*You see, you wouldn't ask why
the rose that grew from the
concrete had damaged petals.*

*On the contrary, we would all
celebrate its tenacity. We would
all love its will to reach the sun.*

*Well, we are the roses. This is
the concrete, and these are my
damaged petals. Don't ask me
why. Thank god. . . Ask me
how. . .*

TUPAC AMARU SHAKUR

To my parents, my brothers (Honey, Houman), and my beloved sisters (Saba) for their endless love, support and encouragement.

I

**First part: context,
problematic, and background
of the thesis**

1

Introduction

“If you are aiming to be the best at what you do, you can’t worry about whether your actions will upset other people, or what they’ll think of you. We’re taking all the emotions out of this, doing whatever it takes to get to where you want to be. Selfish? Probably. Egocentric? Definitely. If that’s a problem for you, you chose a wrong sport.”

The coach-Tim Grover

1.1	Context and the social problematic relevant to patient pathways	3
1.1.1	Challenge No. 1 – “Structural”	3
1.1.2	Challenge No. 2 – “Duration & distance”	5
1.2	Focus and scope	6
1.3	Scientific challenges and questions	7
1.3.1	What data science-oriented procedure is suitable for extracting knowledge from location data?	7
1.3.2	How to interpret location data?	8
1.3.3	How to extract a descriptive reference process model?	9
1.3.4	How to acquire a holistic analytic approach for patient pathways?	9
1.3.5	How to measure the distance of a descriptive model from the normative model?	9
1.3.6	How to automatically diagnose the causes of deviations?	10
1.3.7	How to generate different scenarios for choosing the optimum solution?	10
1.4	Contribution of this thesis	10
1.5	Outline of the dissertation	11
	Important terminology of this dissertation	12

1.1 Context and the social problematic relevant to patient pathways

1.1.1 Challenge No. 1 – “Structural”

Approximately 6.9 million outpatient hospital appointments are missed in the UK every year and each missed appointment costs around £108 (Pinchin, 2015). This article

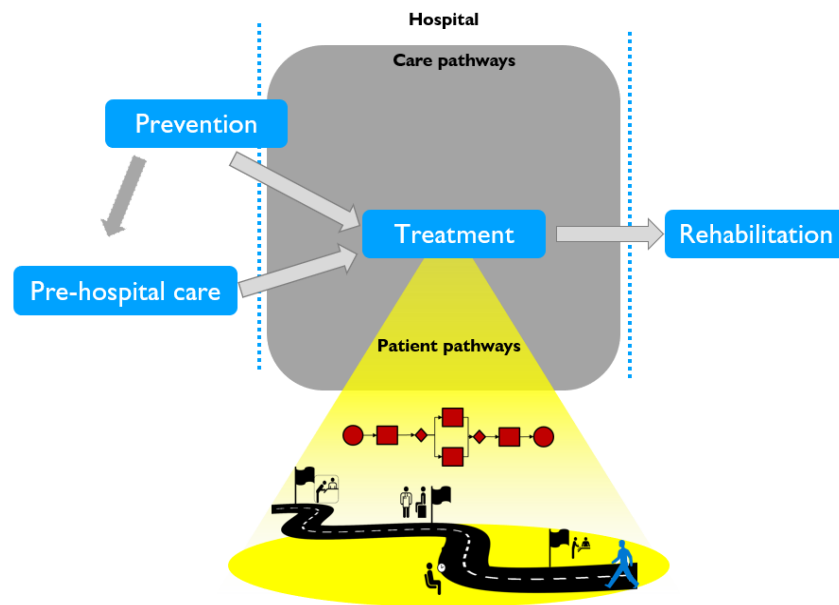


Figure 1.1 – Positioning the definition of **patient pathways** in accordance with previous research works such as the one in (W. Yang et al., 2014).

highlights that patients often get lost in hospitals and this issue can lead to many problems, such as missed appointments.

This costly issue has been mentioned in other research works as well. Ullah et al. (2019) mentioned that the USA healthcare system is affected by a \$ 150 billion annual cost of missed appointments or the so-called “no-shows”. They have highlighted several causes such as **structural or organization problems** and **personal barriers**. The present research work is focusing on the *structural problems* and not the personal barriers.

Basically, it considers that the missed appointments are due to difficulties patients experience in finding their way around hospitals, with kilometers of near identical corridors leading from one door to another with similar names. There is clearly a necessity to reduce such expensive defect costs.

Several research works referred to this issue by conducting reports about the placements of patients in different hospitals around the globe; UK (Commission et al., 2000), Spain (Alameda et al., 2009), New Zealand (Creamer et al., 2010), and France (Lepage et al., 2009).

The deviations in the structure of patient pathways is a common challenge for health-care organizations (Alameda et al., 2009; Ashdown et al., 2003; Gilligan et al., 2008; Wolstenholme et al., 2004). Accordingly, the patients that are not following the **normative pathway** (the desired pathway defined by the domain experts) are considered as “outliers”, “boarders”, or “sleep-outs” (Goulding et al., 2012).

If it would be possible for a hospital to acquire an overview of the **common pathways** that shows the **usual routes taken by patients** when moving around the hospital and conducting their activities, then healthcare experts would be able to detect variations in patients processes. **This could be achieved by comparing such a model with the normative model.** Thereby, such a vision could help to avoid inefficient behaviors in these processes, or as they are called here, these **patient pathways**.

Many authentic research works addressed the subject of patient pathways (Campbell et al., 1998; Probst et al., 2012). For instance, Hall (2013) used “patient flow”¹ to express the sequence of activities patients do in a healthcare facility.

This work considers a patient pathway as the journey that a patient starts from the first contact with hospital staff through tasks such as consulting on their health problems, to different diagnostic and administrative actions inside the healthcare facility until they leave the premises. These patient pathways are a sub-class of care pathways and can be considered as **business processes** (c.f. figure 1.1).

To elaborate on this statement, the definition of care pathways by Vanhaecht et al. (2010) should be revisited, where the authors defined the care pathways as a “complex intervention for the mutual decision-making and organization of care processes for a well-defined group of patients during a well-defined period”.

In addition, it has been signified that the aim of these care pathways is to enhance the quality of care across the continuum by improving risk-adjusted patients’ outcomes, promoting patients’ safety, increasing patients’ satisfaction, and optimizing the use of resources. Based on these definitions and applied analyzing methods in this research work, it could be inferred that patients pathways can be categorized as **care pathways** too.

On the other hand, each patient pathway consists of sequences of *events*, several *steps*, *decision points*, *actors*, and *activities* with the *objective* of delivering health care to the patient. These are the mutual elements with the definition of **business processes** provided by Dumas et al. (2018). Consequently, one could conceive these pathways as **business process models**. As Dumas et al. (2018) mentioned, a business process could be seen as a mean that organizations use to deliver a service or product to clients.

With these being said, to answer the structural problem, finding the common pathways from all of the patient pathways is a primary task. The common pathway helps to increase the awareness of healthcare organizations regarding patients’ processes and activities. This research work will present a new approach to capturing a **descriptive reference process model** that could serve as a solution.

Figure 1.2 presents the differences among the mentioned notions, such as the normative model which is defined by the domain experts and the descriptive pathways that should be extracted (these descriptive pathways can be the process model of only one patient).

The extraction of descriptive models—patient pathways—and the common patient pathways will be addressed in this research work by using indoor location data and process mining activities.

1.1.2 Challenge No. 2 – “Duration & distance”

Imagine this scenario:

- You have an appointment in a medical clinic and by default the clinic has asked you to be there *20 minutes before the actual appointment*. So, you respect this rule and arrive on-time. However, due to certain problems you have to wait *40 minutes*.

¹ Both terms of patient pathways and patient flows are used in the literature. In both cases the researchers evoked approximately the same context.




Figure 1.2 – A representation of the normative, descriptive, and common pathways.

In this scenario, what disturbs you? The problem for you is not the waiting time; in contrast, you’ve respected the 20-minutes delay that they have mentioned. However, you’d be annoyed by the fact that you don’t know what to expect from this process and you have lost 20 minutes of your time. Should you respect the 20-minutes delay for the second appointment or not?

Well, in this case, the **uncertainty of the process outcomes** makes the patient to be unsatisfied and confused.

In view of such problematic, this research work aims to propose different methods to **monitor and detect** the variations and uncertainty of process outcomes.

Such an approach endorses the quality improvement of patient pathways.

 **The social question**

In the view of these mentioned challenges, this research work molded to answer this social question:

How to improve patient pathways?

This improvement can be directed toward structural problems, such as “removing deviating behaviors in patients processes” and “stabilizing duration and the taken distance of patient pathways” .

1.2 Focus and scope

In order to answer the above question, healthcare experts ² need a course of action to monitor patients’ activities.

According to Dumas et al. (2018), the healthcare experts can choose one of these following approaches:

²A domain—healthcare—expert can be a nurse, physician, department director, or any other administrative person in charge to provide a set of primary information about the environment.

1. **Interview-based:** As the name indicates, this approach is based on interviewing patients and staff to capture an image of the AS-IS situation of processes. The objective is to model the information and to see where are the hidden problems .
As one can imagine, the results of this approach can consist of several uncertainty, since each process actor (patient or staff) can perceive the process differently. Moreover, it is significantly a time-consuming approach.
2. **Workshop-based:** In this approach, both the domain expert and the business process modeling expert will work together to devise the process models. The design of these models can undergo several meetings and discussion sessions. Similar to the interview-based, it can be a sluggish approach.
3. **Evidence-based:** This approach is emerging by the fast evolution of the process mining research discipline. Experts by applying this approach go through the existing evidence (e.g. information systems) in the organizations and try to extract and design process models. The modeling can be done by hand. However, the process discovery activity of process mining can be a prompt and accurate solution for the experts to map out the AS-IS state of processes.

Against this background, the present research work evokes the idea of associating the **Indool Localization Systems (ILS)** and **Process Mining** research fields.

The ILS permits to track patients movements inside the premises. This technological solution provides several **advantages**; such as: (i) *accuracy* in tracking patients activities. (ii) *real-time awareness* of patients situations. (iii) *improving the safety* of patients and staff.

However, it can provide ambiguity due to its ability to generate voluminous data sets. In fact, each location tag can emit an event per second. It takes only couple of minutes to encounter thousands lines of events for several patients.

This being said, process mining became a proper solution for extraction of meaningful information from the location event logs (Dogan et al., 2019b; Fernandez-Llatas et al., 2015). Figure 1.3 illustrates how this research work is going to benefit from associating process mining notations (or so-called process mining activities) and ILS to answer the earlier-mentioned social question: *how to improve patient pathways?*

As shown in figure 1.3, the location data generated will be enriched with a primary **domain knowledge**. Next, this information will be used for three process mining notations that will be explained within the chapters of this dissertation.

1.3 Scientific challenges and questions

Although such application seems intriguing, there are multiple **scientific questions** that emerged during the development of this research work.

The following will introduce these issues.

1.3.1 What data science-oriented procedure is suitable for extracting knowledge form location data?

According to Provost et al. (2013), when one is dealing with a data-oriented project, it is extremely important to identify the so-called **data mining procedure**.

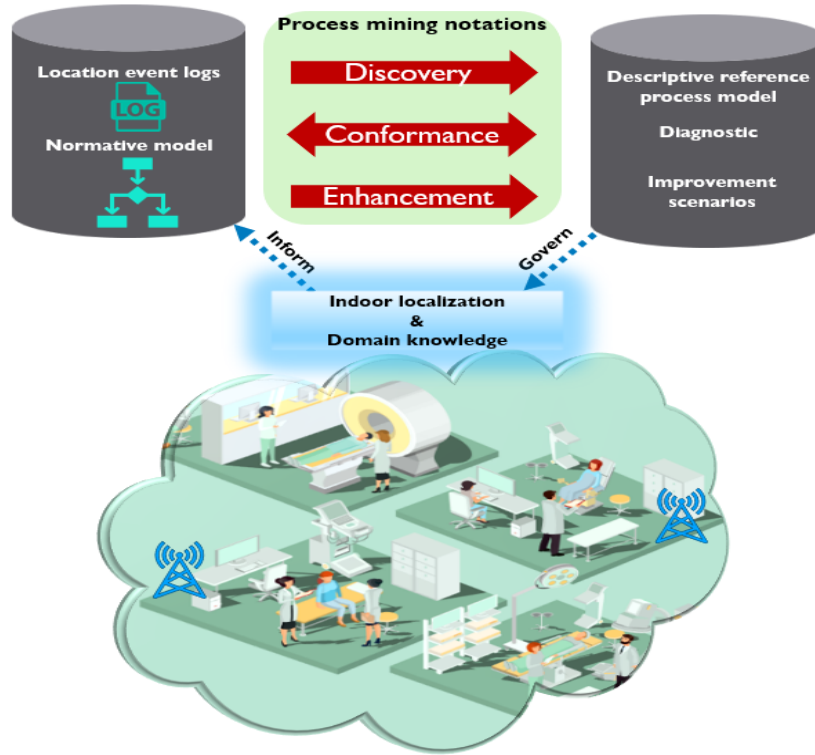


Figure 1.3 – Associating process mining activities with the indoor localization solution.

At the very beginning of this research work, the need for such a formal procedure that addressed “**knowledge extraction**” from the location event logs has emerged.

Such a procedure clarifies the required steps and their orders to extract knowledge from the data. For instance, it helps to understand which is the primary step; the data treatment or the process modeling.

The **DIAG methodology** has been introduced in the second part of this research work to target this requirement. This methodology will be exclusively described in the second part of this dissertation.

1.3.2 How to interpret location data?

One of the most important challenges was to provide a solid foundation to interpret and enrich the information recorded within location event logs. For instance, the location event logs can include information about the temperature, humidity, change of positions, and zones. It is required to interpret these data and to understand which *event represents a process activity*. This research work refers to this requirement as the **sense-making** procedure.

This decision was mainly due to the format of these data. Chapter 3 will explain this issue.

1.3.3 How to extract a descriptive reference process model?

As it will be discussed in chapter 2 and 4, based on the previous methods in the literature of process mining in healthcare, it was not a feasible task to extract a process model which indicates the **common pathway** of patients.

To illustrate this problem, imagine tracking of 100 patients inside a hospital premises. These patients could have a unique pathway if they all follow the normative model. However, in reality, this is not the case. Patients can follow different pathways and execute different activities (c.f. figure 1.2).

As a result, it is difficult for the domain experts to refer to a certain process—patient pathway—as the common pathway. The common pathway can indicate the activities that normally are present in any patient pathway.

Current process mining methods provide multiple process models and they let the experts to make the decision to indicate the common pathway. Consequently, this decision and mining of processes are completely arbitrary.

The methods presented in chapter 4 of this dissertation aim to overcome this challenge.

1.3.4 How to acquire a holistic analytic approach for patient pathways?

According to the literature of business process analyzing, a holistic analytic approach is constructed on top of two main pillars: **qualitative analyses** like visualizing process models, and **quantitative analyses** (Vergidis et al., 2008). The quantitative analyses should be endorsed by solid mathematical methods which are able to detect the **deviations** in the outcome of processes and not only visualizing a basic statistical methods.

Previous quantitative methods, such as visualizing duration of processes by histograms, are not capable to evaluate the performance of the processes. Those methods only provide —once again— qualitative measures. This is due to their inability for detecting inefficiencies and deviations in processes.

This is a non-trivial issue which will be investigated in the fifth chapter of this research work.

1.3.5 How to measure the distance of a descriptive model from the normative model?

As mentioned previously, comparing the normative model with the descriptive model can be a useful approach to evaluate patients processes. To do so, the distance of an extracted patient pathway from the normative one is considered.

With this in mind, this research work introduces an algorithm which is able to measure the distance between process models. The ProDIST algorithm will be discussed within chapter 5. Additionally, it is used for performing an **automatic business process diagnosis**.

1.3.6 How to automatically diagnose the causes of deviations?

Within the literature of business process management, **BPA** (business process analysis) addresses the evaluation and improvement of processes; in addition, the **enhancement activity** of process mining focuses on this issue as well (Dumas et al., 2018; Vergidis et al., 2008). However, the current methods fail at discovering the causes of inefficiencies. Consequently, this research work in chapter 5 introduces two novel methods to perform automatic business process diagnosis.

1.3.7 How to generate different scenarios for choosing the optimum solution?

After detecting the deviation causes, it is required to remove these inefficiencies to improve the process. Indeed, such an action embraces the definition of continuous improvement and objectives of quality engineering (Montgomery, 2019), (Dumas et al., 2018).

To do so, one should ensure about the effects of such modifications in the process outcomes. This can be defined as **prognostic actions**.

Discrete event simulation is an appropriate candidate for accompanying the introduced methods of this research work to propose several improvement solutions.

As mentioned in (W. M. P. v. d. Aalst, 2018) this can be a “match made in heaven”. This proposition is identified as a future perspective of this thesis, however, it is defined as the seventh step of the DIAG methodology.

Table 1.1 provides a summary of the scientific challenges of this work and the corresponding solutions that are developed during this thesis.

1.4 Contribution of this thesis

Scientific questions	Proposed Solutions	In which chapter
1. What procedure should be chosen to extract knowledge from location event logs?	DIAG methodology	Part 2 of the dissertation
2. How to interpret location event logs?	DIAG meta-model & Location data interpretation rules	
3. How to extract a descriptive reference process model?	Stable Heuristic Miner V1 & V2 algorithms	Chapter 3
4. How to acquire a holistic analytic approach for patient pathways?	Application of SPC (Statistical Process Control)	Chapter 4
5. How to measure the distance of a descriptive model from the normative model?	ProDIST algorithm	Chapter 5
6. How to automatically diagnose the causes of deviations?	MMP method & DIAG method	Chapter 5
7. How to generate different scenarios for choosing the optimum solution?	Future perspective	

Table 1.1 – Scientific questions and the proposed solutions.

In accordance with the mentioned scientific challenges in the previous section (1.3), this section briefly mentions the relevant proposed solutions by this thesis. These solutions are defined as the contributions of this work. They are listed as:

1. **DIAG methodology.**
2. **DIAG meta-model.**
3. **Location data interpretation rules.**
4. **Stable Heuristic miner algorithms (V1 & V2).**
5. **Application of statistical process control (SPC) for the enhancement activity of process mining.**
6. **ProDIST algorithm.**
7. **Miniscule Movements of Processes (MMP) method.**
8. **DIAG process diagnosing method.**

The first contribution is the DIAG methodology which has one main goal; to help domain experts in their (location) data-driven decision-making procedures.

This methodology will be explored in the **second part** of this thesis. It has four main states, **data**, **information**, **awareness**, and **governance**. Each state includes one or several functions. Every functions is defined to address the mentioned scientific problems.

Other listed contribution of this thesis are defined within the functions of this methodology.

For instance, the first function presents the **DIAG meta-model** as the second contribution of the thesis. The second functions relates to the data gathering. The third function, introduces the **location data interpretation rules**. The fourth one, presents the two **process discovery algorithms** to extract the descriptive reference process model.

The fifth function focuses on the **application of SPC** (Statistical Process Control) and applying quantitative analyses. The sixth function, introduces **two methods (MMP and DIAG)³ to automatically detect the causes of deviations in process models**.

Finally, the seventh one suggests discrete event simulation as a solution to propose multiple improvement scenarios.

1.5 Outline of the dissertation

The **first part** of this dissertation focuses on introducing the [background](#) and the [state of the art](#) of this research work.

Accordingly, as shown in figure 1.4, the second chapter presents the state of the art. This chapter explores the literature of the most relevant research disciplines to this work: data science, process mining, and the application of location data in data and process mining projects.

The **second part** of the thesis begins by introducing the **DIAG methodology**. Then, [chapter 3](#) focuses on presenting the first three functions of the DIAG methodology. Mainly, it shows how the location data are received and interpreted.

[Chapter 4](#) introduces stable heuristic miner algorithms (V1 and V2) for extracting the descriptive reference process model.

³MMP: Miniscule Movements of Processes.

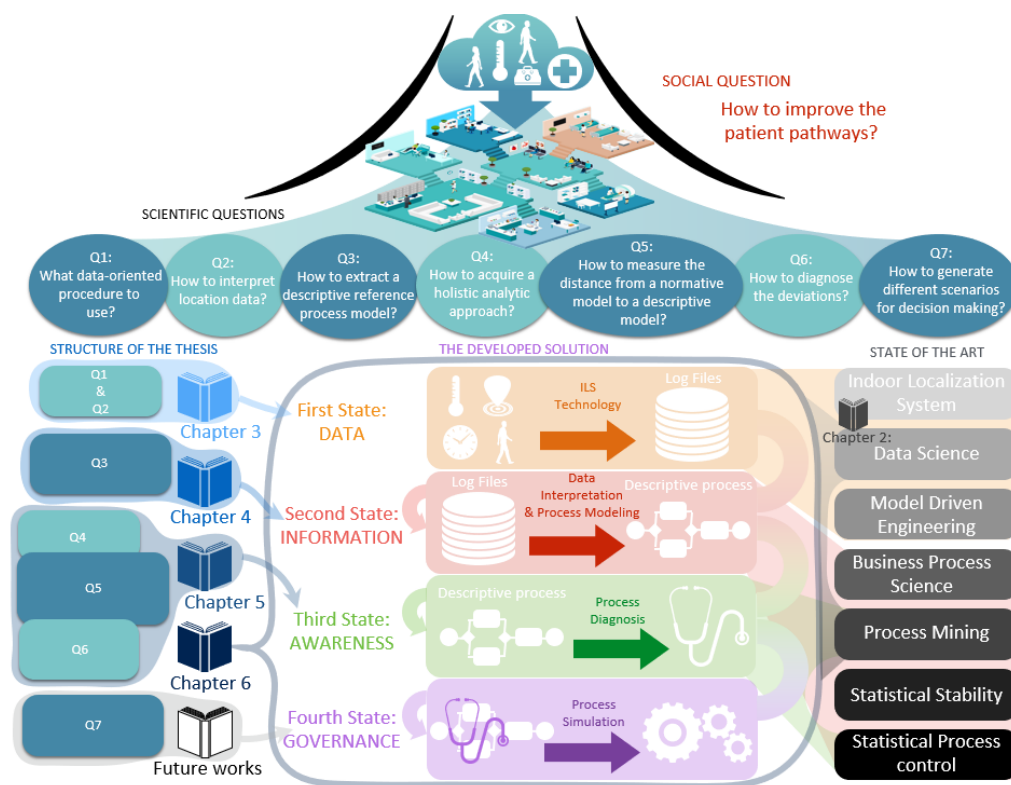


Figure 1.4 – An overall view of this dissertation.

[Chapter 5](#) addresses the application of SPC, ProDIST algorithm, and the two automatic business process diagnosing methods.

Finally, in the **third part** of the dissertation, [chapter 6](#) presents a case study, where the introduced methods of this research work are put into action. The [conclusion](#) is the finishing element of the thesis.

Important terminology of this dissertation

Table 1.2 clarifies some used terms in this dissertation.

Process Model name	Acronym	Source (defined by experts or extracted from data)
Reference (or normative) model	RM	This model is defined by the domain—healthcare—expert.
Descriptive Model	DM	It can be extracted from data by process mining. It represents a random amount of information for a process-like pattern in an event log.
Descriptive Reference Process Model	—	This model presents the common and stable process (pathway). It is mainly used in chapter 4.
Generated Model	GM	This is a model generated by injecting potential assignable causes. It will be addressed in chapter 5.

Table 1.2 – This table clarifies some of the important terms that will be seen by the readers. These terms are related to the processes (patient pathways) that are either designed by the experts or discovered from event logs.

2

State of the art

2.1	Introduction	13
2.2	Data Science	14
2.2.1	Data science project requirements	15
2.2.2	Data science in healthcare	16
2.2.2.1	Levels and types of healthcare data	17
2.2.2.2	The challenges of data science in healthcare	17
2.3	Process Mining	19
2.3.1	Where process mining has started?	19
2.3.2	Used modeling languages	20
2.3.3	Process mining in healthcare	22
2.4	Process mining notations (activities) in healthcare	24
2.4.1	Works related to the business process discovery and modeling notation in healthcare	25
2.4.2	Enhancement notation (activity) of business processes in healthcare	28
2.4.3	Conformance checking activity in healthcare	30
2.5	Indoor Localization Systems (ILS)	30
2.6	Data mining and process mining have met location data before:	32
2.6.1	Preparation and interpretation of location data	35
2.6.2	Use of location data for knowledge extraction	37

"The greats never stop learning. Instinct and talent without technique just makes you reckless, like a teenager driving a powerful, high-performance vehicle. Instinct is raw clay that can be shaped into a masterpiece, if you develop skills that match your talent. That can only come from learning everything there is to know about what you do."

Tim Grover

2.1 Introduction

Recall the social and scientific questions, this chapter is introduced here to illustrate how different **decisions** construct the path toward answering those questions. As mentioned before, several research fields were analyzed to devise the DIAG methodology.

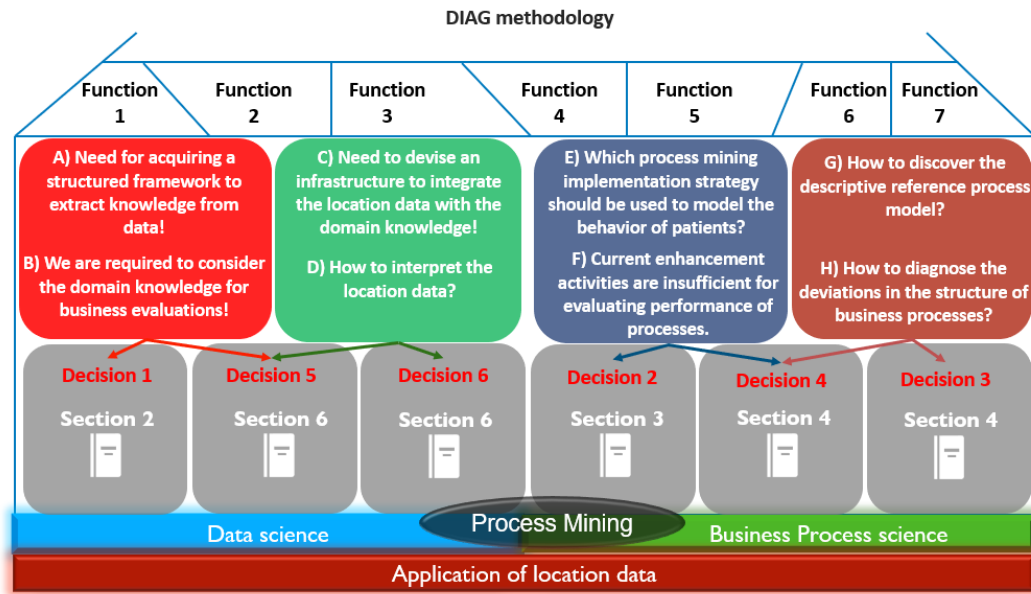


Figure 2.1 – The structure by which two main keywords of this research work have been explored. Several decisions are taken based on the gaps found in the literature. The DIAG methodology and its functions are devised toward addressing these gaps.

Accordingly, **six decisions** are highlighted in this chapter as the result of analyzing the state of the art. The presented decisions here are embraced by the context of data science and process mining.

Figure 2.1 presents the structure of this chapter in a way by illustrating the **research questions** that came to mind while analyzing the literature for extracting knowledge from location data.

Section 2.2 illustrates some aspect and requirements of data science applications. Section 2.3 explores the process mining overall concepts. Then, section 2.4 illustrates the recent trends for process mining activities in the healthcare sector. Section 2.5 introduces a general overview for ILS technology. Finally, section 2.6 presents the literature and the recent trends for applying data and process mining-like methods for ILS applications.

2.2 Data Science

The vast volume of existing data around organizations has led researchers and industries to be more eager in learning different means and ways of extracting beneficial information from data. A simple look on search engines would show the relentless need of recruiting data scientists and engineers by industries to gain a competitive edge over their competitors.

The data-analytic thinking enables organizations to detect the opportunities and threats regarding their operational, decisional and technical processes. Avoiding the exploration of data is a fatal choice in today's business. However, this is not an obligation nor keeping the competitive edge in the market.

Throughout this dissertation, we will discuss several scientific methods for extraction of knowledge from **location event logs**. In this vein, it is important to recognize the prerequisite actions for data science projects.

2.2.1 Data science project requirements

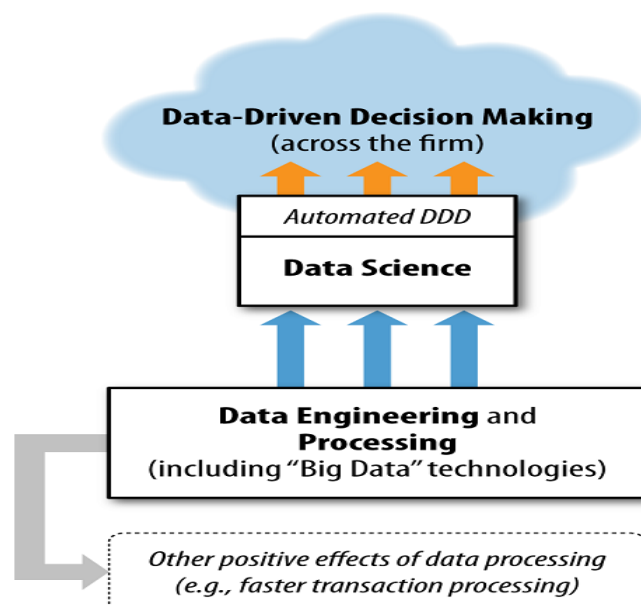


Figure 2.2 – The relationship among data science, data engineering, and data-driven decision making in business (Provost et al., 2013).

According to figure 2.2, Provost et al. (2013) illustrates that any data-oriented projects should start by data engineering and processing. Later on, the data science methods for extraction of the desired information should be employed. However, one essential point here is to avoid neglecting the **business analytic** mindset as well.

Extracting patterns of data and performing data-driven analyses without adequate understanding of the business would not lead to a coherent knowledge extraction. Many data-oriented projects fail due to the fact that they are well addressing the extraction of patterns from data, but, they suffer to extract the **knowledge** from these information and their comprehension of the business problem stays at the **information level**.

Provost et al. (2013) addressed this issue and highlighted the **importance of engaging the business analytic techniques** for making data-driven decisions.

The mentioned criteria pushed this research work toward proposing methods which consider **the domain knowledge** as the base for extracting useful knowledge from location data. For instance, the methods presented in chapter 5 use the domain knowledge to perform **automatic business process diagnosis**.

Another requirement to ensure about the efficiency of the data science projects is to acquire a **data mining procedure**. This could be a procedure in which every step is well defined and planned. In this order, at each step of the project the experts know what actions should be taken and what methods and tools are necessary. One of the known data mining procedure is the **Cross Industry Standard for Data Mining** procedure (**CRISP-DM**) proposed by Shearer (2000).

As shown in figure 2.3, in this framework the experts are continuously gaining more information from each function. It has been indicated that the data mining process needs to address the **business understanding or domain knowledge** at two functions.

- First, *when the primary data is received.*
- Second, *when the experts want to evaluate the performance.*

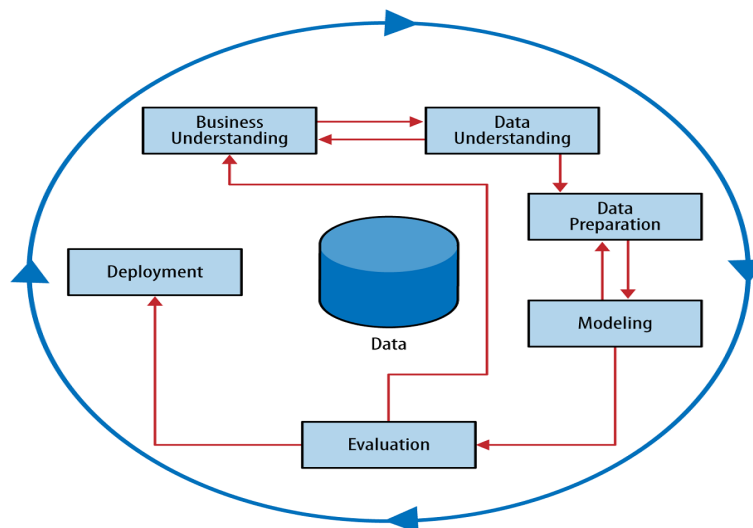



Figure 2.3 – The Cross Industry Standard for Data Mining proposed in (Shearer, 2000); cited from (Provost et al., 2013).

 **#1-A decision based on the state of the art**

With respect to the above two criteria, this research work aimed to use the domain knowledge in this fashion to **interpret the location data** and to **diagnose** the processes.

Stems from the fact that it is highly important to respect a certain procedure for extraction of knowledge from data, this research work has proposed the **DIAG methodology**, which will be studied in the second part of this dissertation.

2.2.2 Data science in healthcare

The essential goal of using *data science application for health informatics* is to take-in real world medical data from all levels of human existence to advance our understanding of medicine and medical practice (Herland et al., 2014).

Jothi et al. (2015) have analyzed 50 research articles and highlighted that the majority of interests for applying data mining methods are related to discovery of **predictive models rather than descriptive models**.

Consequently, out of 50 articles, 45 of them were related to artificial intelligence and machine learning applications, 3 were addressing statistical methods (which can be used for predictive decision makings), and only 2 were relevant to the probability.

Most of data mining projects in the literature were applied in healthcare while addressing the **EMR** (Electronic Medical Records) (Prodel, 2017). EMR covers several types and levels of data.

2.2.2.1 Levels and types of healthcare data

Herland et al. (2014) have conducted a research regarding research works oriented to health informatics and the application of big data in healthcare. They have devised a framework to classify healthcare data. There have defined 4 **levels of data** which are used in different sections:

1. **Molecular**: using micro level data-molecules.
2. **Tissue**: using tissue level data.
3. **Patient-level**: using patient-level data.
4. **Population**: using population-level data like social media.

Prodel (2017) has examined the **patient-level data** and separated this level into 4 types of health care that are commonly used for data-driven analyses.

- Data directly related to a patient (diagnoses, administrative information, characteristic)
- Data related to a care activity (medication, surgeries, medical imaging, biology test)
- Data related to a care event (data, duration, severity, cost)
- Data related to the organization (appointments, human and material resources, number of beds, work schedule)

Many researchers focused on the patient-level data to extract knowledge for the organization. For instance, Zolfaghar et al. (2013) monitored a hospital data over a year in order to reduce the risk of re-hospitalization. To demonstrate the results they have applied a k-mean clustering method.

Others like the work in (Hess et al., 2012) tried to monitor the diagnosis procedure of certain types of diseases.

In a similar fashion, this dissertation explores patients' processes or as called here **patient pathways**. As a result, the **non-medical information** is considered.

Hence, the work presented here copes with the **patient-level data** and examines 3 types of patient data: *data related to the care activity*, *data related to the care event*, and *data related to the organization*.

2.2.2.2 The challenges of data science in healthcare

This section presents the motivations for associating process mining and indoor localization technologies.

Mans et al. (2015) highlighted several challenges for data mining to propose the appropriate models in healthcare.

1. The application of data mining techniques in healthcare requires skills beyond abilities of users in the organization. In general the application of data mining techniques has been conceived as a tough task to handle for healthcare domain users.

2. The supervised techniques have low accuracy to be used in a complex organization such as hospitals. This is due to the misleading information within hospital information systems. On the other hand, the autonomous level of data production is quite high.
3. The other challenge is the availability of data for the experts. Oftentimes, it's not easy to acquire large sets of data for the experiments, therefore, the case studies are not substantial enough in comparison with the whole existing data set.
4. If one is addressing an objective of improving the performance of operational processes at any level, one needs to address the **process like patterns** hidden in the data. However, data mining addresses the patterns in the data that does not consider the end to end processes.
5. Considering that healthcare cost is one of the biggest financial challenges in the U.S, therefore, optimizing healthcare processes while improving the quality of health is extremely important (Institute of Medicine (US) Committee on Assuring the Health of the Public in the 21st Century, 2002). In this regard, activities related to the enhancement of processes are lacking the sufficient visibility. **Process mining** is addressing such an issue.

As it has been mentioned, it is challenging for data mining methods to address end to end process-like patterns in the event logs. As a result, **there are certain questions that can not be answered by data mining:**

1. **What happened? (Analyses):** these questions are relevant to understanding for example what is the typical treatment action for patients? How and when are patients transferred to a healthcare center? What is the typical process of staff?
2. **Why did it happen? (Diagnosis):** this is related to **extracting the cause and root of an unexpected result**. For example, what caused the unusual amount of incidents in the department? Why was the service level agreement not reached? Why did people stop performing the "checkout" activity? What caused the long waiting time? Currently, the **diagnosis methods are missing** for healthcare processes and diagnosing actions completely depend on the expert decision. This is a motive for this research work to seek for a solution to overcome this challenge.
3. **What will happen? (Prognosis):** this is linked to examples such as when will this patient be dismissed?" Is this patient likely to deviate from the normal treatment plan? How many beds are needed tomorrow? Is it possible to handle these five new cases in time?
4. **What is the best that can happen?(Prognosis):** this is related to simulation of scenarios for making the best decision. Such actions would give an ability to the experts to answer questions such as which check should be done first to reduce the flow time? How many physicians are needed to reduce the waiting list by 50? How to redistribute the workload for the staff?

Other considerable research works such as the one in (Nambiar et al., 2013) also investigated the challenges of data science in the healthcare sector. These questions and challenges could be investigated thanks to other data-driven research disciplines known as **business process science** and **process mining**.

2.3 Process Mining

Process mining is concerned with the extraction of knowledge from event logs registered in an information system. Process mining has three primary notations: **process discovery**, **conformance checking**, and **enhancement of business processes** (W. M. P. v. d. Aalst, 2016).

In practice the definition of process mining is misinterpreted by data mining denotations. Although these two paradigms share a similar purpose, to extract knowledge from data, as van der Aalst points out, process mining and data mining both start from data, but **data mining techniques are not typically "process-centric"** (W. M. P. v. d. Aalst, 2016). In addition, the three process mining activities defined above are not well-addressed by data mining techniques. Also, process mining uses a transformed version of data that are recorded as **event logs**.

A precise definition of an event log is provided in (W. M. P. v. d. Aalst, 2016). To explain an event log regarding patients' pathways, a **patient** should be considered as a case which can run a **process** by a *trace* of activities. These **activities** are being recorded thanks to certain *events*. Therefore, **an event log contains one or multiple traces** that are executed by different cases, and **each trace includes a series of events**. Each case (patient), trace (sequence of activities), and event can have one or several **attributes**. For instance, an event could have attributes such as *case-id*, *time-stamp*, *activity*, *resources* and so on. These attributes are used to define certain **classifiers**, which helps to discover the relevant patterns of processes. These concepts are further investigated in chapter 3.

2.3.1 Where process mining has started?

Agrawal et al. (1998) were early pioneers of process mining. Their algorithmic approach to process mining allowed the construction of process flow graphs from execution logs of a workflow application.

Cook et al. (1998) also attempted to find the software process models from the data contained in event logs.

W. v. d. Aalst et al. (2004) compared the method of extracting process models from data with a refining procedure. In terms of business process mining, van der Aalst states that almost any transactional information system can provide suitable data (W. v. d. Aalst et al., 2004).

According to Rebugue et al. (2012), as organizations depend on their information systems to support and execute their work, these systems can record a wide range of valuable data such as which tasks were performed, who performed the task and when. For example, as patients arrive in an emergency room the system could record the time and the needed health treatments, the nurse who performed the tasks and the required materials. Process mining offers an interesting approach to extract such knowledge in an efficient way and to solve the mentioned shortcomings of business process analyses.

Essentially, process mining sits between data science paradigms (like: artificial intelligence, statistics, machine learning, data mining, and etc.) and process science paradigms (like: lean management, BPM, industrial engineering, optimization, stochastic, and etc.) and it plays a bridge-like role to establish links between the actual processes and their data on the one hand, and process models on the other hand. W. M. P. v. d. Aalst (2016) presented the relationship among data science, business process science, and process mining in figure 2.4.

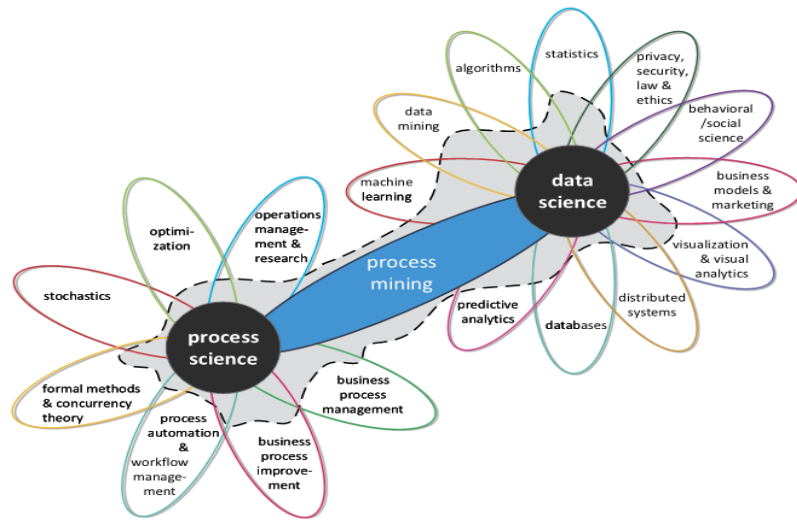


Figure 2.4 – The relationship among data science, business process science, and process mining defined in (W. M. P. v. d. Aalst, 2016)

Nowadays, process mining is highly applied in industries to capture an image of what is the state of the operational processes (which is not a full-usage of process mining potentials). In this order, industries use this research discipline to facilitate the discovering and modeling of their business processes. This research work presents a novel modeling language for discovering and modeling business processes to not only observe the map of processes, but, it is also capable to be simulated.

Therefore, the next section covers the traditional and existing business process modeling languages.

2.3.2 Used modeling languages

W. M. P. v. d. Aalst et al. (2003) has identified two types of workflow meta models: (i) Graph oriented and (ii) Block-Oriented models. Each of these models has their own languages and graphical representations.

However, Aguilar-Savén (2004) adds NET-BASED languages to this definition, but van der Aalst does not make a difference between net-based and graph oriented models in (W. M. P. v. d. Aalst et al., 2003). Also, he describes net-base models such as petri-net as a form of graph-oriented models.

The most common form of graph-oriented meta-models is the **directed graph**. Agrawal et al. (1998) were one of the first ones to use directed graph in process mining. In this way, the authors described a number of constructs involved in the actual graph. Activities, usually mentioned in boxes or circles, and the arrows between the activities that indicate the direction of flow are known as edges. The article presented by Agrawal et al. (1998) was an inspiration for this thesis to classify value of each activity in a process model.

One of the most prominent works that has been used in this research work is the review provided by Augusto et al. (2019a). The authors contributed a complete overview of the existing discovery approaches, modeling methods, and existing applications. They have presented another approach to classify modeling methods. They identified three

modeling types: **declarative**, **procedural**, and **hybrid**. These different types of models are being represented by several modeling languages such as:

1. *Process tree*: The Inductive Miner (S. J. J. Leemans et al., 2014a; S. J. J. Leemans et al., 2014b) and the Evolutionary Tree Miner (Buijs et al., 2014) are both based on the extraction of process trees from an event log. Many different variants have been proposed during the last years, but its first has been presented in (S. J. J. Leemans et al., 2013). Immediately, an enhanced version of this was presented in (S. J. J. Leemans et al., 2014b) due to the inability to deal with infrequent behaviors in the previous version. Many research works followed this approach and focused on process-tree mining (M. Leemans et al., 2017; S. J. J. Leemans et al., 2015).
2. *Petri-nets*: A petri-net is a modeling language which consists of several connecting arcs, places, transitions, and tokens. Places might hold tokens. The movements of the tokens by the transitions from one place to another leads to the change in the state of the process.

The authors in (Huang et al., 2011) described an algorithm to extract block-structured petri nets from event logs. This technique works by first building an adjacency matrix between all pairs of tasks and then analyzing the information in it to extract block-structured models consisting of basic sequence, choice, parallel, loop, and self loops. This method has been implemented in a standalone tool called HK.

In (Rubin et al., 2007), the alpha algorithm has been presented by using this modeling language. This algorithm which is considered as a pioneering method for discovering concurrency can extract the invisible tasks within a non-free-choice construct. Non-free choice constructs are the situations where there is a mixture of relations and synchronizations among activities in an event log.

The authors in (H. M. W. Verbeek et al., 2017) also used petri-nets as the modeling representation. The authors presented a generic framework for the discovery of process models from high volume event logs. Their method suggests the split of large data sets into smaller event logs. This could address the issue of incompleteness. Several studies such as (H. M. W. (Verbeek et al., 2013) and (W. M. P. v. d. Aalst, 2013a) extensively illustrate the idea of splitting a large event log into collection of smaller logs to improve the performance of a discovery algorithm while representing processes by petri-nets.

Other research works used this modeling method as well (W. Zheng et al., 2019), (Breuker et al., 2016), (He et al., 2019), (Ferilli, 2014).

3. *Causal nets*: The authors in (Greco et al., 2015; Greco et al., 2012) proposed a discovery method that returns causal nets. A causal net is a net where only the causal relations among activities in a log is represented. The proposed method illustrates the causal relations gathered from an event log. Several methods and plug-ins permit for extraction of causal nets as well. For instance, ProDiGen presented in (Vázquez-Barreiros et al., 2014; Vázquez-Barreiros et al., 2015) applied genetic algorithms for extraction of causal nets. Another example is the Proximity Miner described in (Yahya et al., 2013; Yahya et al., 2016).
4. *State machine*: This method is discussed in (M. L. v. Eck et al., 2017; M. L. v. Eck et al., 2016). The relevant algorithms discover state machines from event logs. Instead of focusing on the events or activities that are executed in the context of a particular process, this method focuses on the states of the different process

perspectives and discover how they are related with each other. These relations are expressed in terms of Composite State Machines (CSM).

5. BPMN: In (Conforti et al., 2014) Conforti et al presented the BPMN Miner, a method for the automated discovery of BPMN models containing sub-processes, activities, loops, and etc. Later on, the method has been improved in (Conforti et al., 2016) for dealing with noisy event logs. Another methods to discover BPMN models has been presented in (Augusto et al., 2019b; Kalenkova et al., 2017; Molka et al., 2015).
6. Declare: applying declarative approaches for modeling processes is linked to removing the execution semantic from the process models. The extracted models will only show the relationship among activities which are shown by basic shapes (no exclusive semantic). This statement does not indicate the declarative approaches are not appropriate for further analyses. **In many cases the domain experts are not familiar with the procedural modeling languages.** Therefore, it is relevant to use declarative modeling in such cases. Many research works addressed the declare modeling language (Bernardi et al., 2014; Maggi et al., 2011; Maggi et al., 2013; Zahoor et al., 2019). In (Aa et al., 2019), and (Mendling et al., 2019) the authors tried to extract declarative process models by applying natural language processing techniques.

Respectively, in this research work the **declarative and procedural approaches are used in parallel**. As a new procedural modeling language **Operation Process Chart (OPC)** is proposed in chapter 3. The declarative approach is used as well to present the results of the new process mining algorithms presented in chapter 4.

2.3.3 Process mining in healthcare

Rojas et al. (2016) provided a general and complete overview on the current situation of process mining in healthcare. They define process mining as a practice to extract knowledge from data generated and stored in corporate information systems in order to analyze the executed processes. Their article shows different use cases that proved process mining is an applicable research discipline in the healthcare domain.

Their approach is toward first defining healthcare processes. They indicated these processes are series of activities aimed to diagnose, treat, and prevent any diseases in order to improve patient's health. They have defined two types of healthcare processes :

- Clinical processes
- Non-clinical processes or administrative processes

This classification is quite varies from the previously expressed classification presented by Mans et al. (2013).

Based on literature multiple strategies have been used to treat and improve healthcare processes: Lean, Evidence Based Medicine (EBM) (Straus et al., 2019), business process redesign, process mining.

According to their research, most used algorithms in healthcare process mining are: **Heuristic Miner**, Fuzzy Miner, and Trace clustering. Heuristic Miner is a process discovery algorithm that can generate process models and is very robust in dealing with noises in event logs (A. Weijters et al., 2006).

Fuzzy Miner is a configurable discovery algorithm that allows through its parameters the generation of multiple models at different levels of detail. It helps to deal with unstructured processes (Günther et al., 2007). Trace clustering techniques allows the partitioning of the event logs to generate simpler and more structured process models (Song et al., 2009). The presented statistics regarding the most applied algorithms became a reference for this research work to consider the heuristic mining algorithms. This will be further discussed in the fourth chapter.

The general process mining methodologies are using clustering approaches to conduct analysis. Clustering approaches have five phases: log preparation, log inspection, control flow analysis, performance analysis, role analysis (Rojas et al., 2016). Moreover, according to Rojas et al. (2016), three **implementation strategies** for process mining projects are considered relevant to the application of process mining in healthcare:

1. **Direct implementation:** Which is applying directly process mining techniques on a gathered event log from hospital information systems. There are two challenges: first is the data extraction and building correct event log. Second, the need to understand tools, techniques and algorithms available for getting analysis.
2. **Semi-automated:** Event logs are made through specific developments. These developments link several data sources and by using queries provide the event log. This strategy has the disadvantage of being defined in an ad-hoc manner for extracting data. This approach did not receive too much attention.
3. **An integrated suite:** In which data sources are connected and event logs are being created automatically and process mining techniques are applied. The advantages of this approach is that one does not need detailed knowledge of connecting data sources or how process mining techniques work. The disadvantages of this approach is its requirement in using expensive infrastructure. Moreover, it has been developed for special cases and does not support all areas and data sources which do not provide a portable solution.



#2-A decision based on the state of the art

Process mining implementation strategy chosen in this research work:

The direct implementation approach is used in this research work. Consequently, the R.IO-DIAG application is developed exclusively to extract knowledge from location data.

The analytic approach for process mining projects in healthcare are:

- Basic
- Basic + adding a new process mining technique.
- Basic + adding a new process mining technique + applying algorithms to other areas such as statistics.

By this classification, this research work is related to the [third analytic class](#). Here, an approach is used to not only get a control-flow perspective, but, to apply **quantitative analyses** for measuring the performance of processes and to diagnose them too.

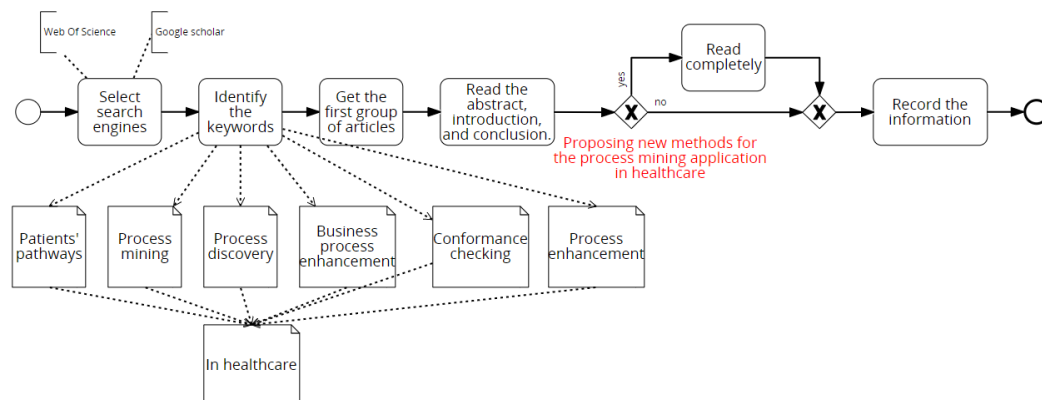


Figure 2.5 – The approach for analyzing the application of process mining activities in the healthcare sector.

2.4 Process mining notations (activities) in healthcare

Process mining has three main notations (or so-called activities); **automated process discovery**, **conformance checking**, and **enhancement of business processes**. This section presents the recent trends for applying these activities in healthcare.

Figure 2.5 presents the approach taken for analyzing the literature for the last 4 years. The reason behind this choice was to analyze the most recent works.

Accordingly, two search engines were used; Google scholar, and Web Of Science. Then, several key words were used **for healthcare applications** such as: *process mining*, *patients' pathways*, *process discovery*, *process enhancement* and *conformance checking*. After this step a first group of articles consisted of 84 research works were selected.

These articles were analyzed rapidly by skimming through abstract, introduction and conclusion to ensure about their closeness with this research work objectives.

Mainly, the research works that only applied already known process mining techniques were put aside. The objective was to extract articles that targeted a gap in the literature of process mining in healthcare and proposed essentially **new approaches and methods** for this field.

Consequently, only 35 articles were chosen. Figure 2.6 presents the share of each process mining notation. According to this figure, the **enhancement of business processes** starts to gain more attention. The process discovery activity still is the primary focus of the researchers. This is due to the huge advantage of applying process mining for extracting a **control-flow perspective** from healthcare data.

Conformance checking activity did not receive adequate attention. This might be due to the fact that existing methods do not suite highly complex environments such as healthcare organizations.

The data quality issues is a subject that should be drawing more attention since it is the primary issue for practitioners.

In the following, brief summaries of these works will be presented.

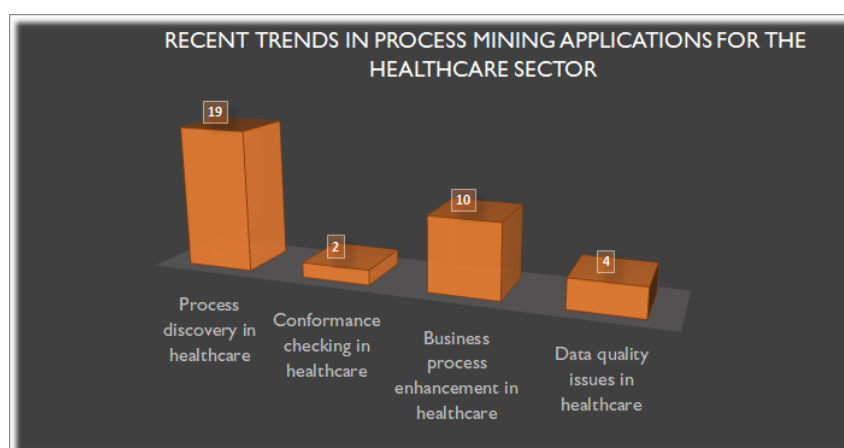


Figure 2.6 – This chart presents the focus of research works during last 5 years for the application process mining in healthcare.

2.4.1 Works related to the business process discovery and modeling notation in healthcare

In (Gurgen Erdogan et al., 2018) the authors proposed a method which comprises several steps for process discovery; defining goals and questions, data extraction, data preprocessing, log and pattern inspection, process mining analyses and generating answers to questions, evaluating results, and initiating proposals for process improvements. This method was used for analyzing a case study in a university hospital in Turkey. The authors mainly focused on the discovering of paths for a certain surgery process while using DISCO. The proposed quantitative analyses were basic analyses on the duration of the processes and cases activities.

The authors in (Weber et al., 2018) have introduced a significant social problem in UK which is related to the conflicts for preparing the prescription for concurrent medications. They have associated process mining approaches with text analytic to extract prescription processes for patients from five primary case studies.

In (Dixit et al., 2018) authors indicated the advantage of process mining for linking data science and process science paradigms, and they have proposed a discovery approach which addresses a gap in the literature by narrowing the distance between domain knowledge and the process experts. This approach was named as ProDiGy. ProDiGy provides an interface for the users to directly modify and interact with the model by proposing new recommendations. These models are executable and it is possible to quantify the results of each recommendation.

In (S. Yang et al., 2018b) authors proposed a clustering-based approach to discover trauma procedures while considering the most frequent behaviors. They have evaluated the association between patients cohorts and the trauma resuscitation procedures. One of the main objectives was to find the main patient cohorts through applying clustering techniques. For the presented algorithm they have used the **domain knowledge** by an interview-based technique and asking the domain experts to define a certain score for a weight attribute. The unsupervised clustering techniques were applied, yet, their approach is highly dependent on this attribute.

The authors in (S. Yang et al., 2018a) presented a very interesting approach by applying a hierarchical hidden Markov model. The authors tried to extract the objective behind

certain surgical and trauma procedures. The reason was the ambiguity for the experts to detect the hidden goals or intentions behind these medical procedures. They have used a robust case study by analyzing 123 trauma cases. In their approach they are able to detect multi-level intentions. In order to be more precise in their approach they have used a state splitting method with a *maximum 'posteriori' probability (MAP)* as the scoring function.

In (Johnson et al., 2019) the authors targeted a difficult task of combining process mining methods and simulation techniques. For this cause, they have used the **clear path method** which is an extension of earlier discussed PM2 methodology. In their research they have addressed a challenge of using EHR (electronic healthcare data) . They have indicated even though EHR is a rich data source it can include many quality problems in the data. To illustrate their contributions they have used the NETIMIS software tool to support healthcare process simulation. This tool is used for academic and commercial purposes as well. In this article their main focus was on presenting the EVIDENCE TEMPLATE for supporting early, low-fidelity models for constructing a more executive process model.

The authors in (Lira et al., 2019) aimed to provide feedback on the performance of processes, and detecting desired and undesired patterns. Therefore, they have proposed a novel approach in using process mining techniques to provide such feedback. Their method is inspired by PM2 methodology. Unlike many process mining case studies, they *did not generate the execution event logs automatically* from an information system. They have used **a human observer to monitor processes**. Their method has five steps: (i) Video recording; the data are gathered from the videos. (ii) Video tagging; the videos were analyzed and tagged by medical doctors for each case, activities, and time-stamps. (iii) Event log generation; the extracted tagged information are used for creation of an event log as an input for process mining activities. (iv) Model discovery. They have used Celonis software¹ for discovering the models. (v) Model analysis. For the analyzing part as well they have used Celonis.

In another approach, the authors in (Rojas et al., 2019) proposed a method for identifying specific patient cohorts from complex "phenotype" as a primary action for applying process mining activities. They have used pattern matching and an abstraction-based digital phenotyping to capture the drug usage patterns of patients.

Traxler et al. (2018) have addressed the challenge driven from the availability of data in healthcare. They used the mode data sources are for process mining in healthcare. As a result, they presented the challenges to find a common approach and language to cope with variety of data and their formats in these data sources.

In a very impressive approach they have listed current frameworks for dealing facilitating process mining studies in healthcare; FHIR , HRM, openEHR, proprietary approaches, and RIM. As a result they have introduced a rule-based information mining method for healthcare data while using FHIR framework, This method helps to extract clinical and patient pathways.

Intriguingly, the authors in (Duma et al., 2018) aimed to find an algorithm to extract a general pathway for patients in emergency departments. Such a pathway should help the experts to predict what are the next possible actions for the patients. The final goal is to simulate these pathways. To evolve their approach they have used a decision-tree-based clustering method for a supervised prediction of patients activities.

¹<https://www.celonis.com/>

However, one question seem to be appropriate to pose, how one can ensure about the applicability of a supervised clustering approach in a high complex and unpredictable environment such as an emergency department with processes changing from one case to another.

In (Ibanez-Sanchez et al., 2019) researchers mentioned the challenges for expert domains to comprehend fully the benefit of applying data science oriented activities for obtaining useful insights about healthcare processes. To overcome this challenge they have proposed process mining to endorse healthcare professionals with analyzing emergency processes and using the critical timing of stroke treatments. It seems that they have applied PALIA algorithm for discovering patients processes. In this work they have improved a missing part of the PALIA tool which was providing statistical analyses about the duration of processes.

The authors in (Cho et al., 2019) presented a real world case study from an EHR system and tried to create a simulation model for analyzing the schedule planning. The objective was to find solutions for reducing the waiting times for patients. They have covered a major challenge in their work, they have mentioned that collected data from EHR are not a proper input for creating simulation models.

To pass by this challenge they have developed a framework to support simulation of schedules in a clinical case study. To this end, they have applied process discovery techniques and used the *patients' pathways throughput time* for analyzing the service time. In their approach they have used frequency mining for capturing a map of the process. It seems that they have created a footprint matrix from the data and used the direct relation among activities. They aimed to present 100% of information in the event log. Hence, they have captured all the patients pathways for their further analyses.

The authors in (Williams et al., 2019) highlighted the importance of **detecting deviations in healthcare processes**. In this order they have defined a method for extracting a so-called prescribing pathway. This research was used to understand how medical practitioners authorize certain drugs for patients.

Another work focused on proposing a new approach for extracting knowledge from Geographic Information Systems. Their approach named as HGIS (healthcare GIS) is a three layer framework to support the integration of heterogeneous data with spatio-temporal features (Batista et al., 2019). HGIS includes a *data layer* where certain data engineering actions take place to prepare the data for the next layer which is the *processing layer*. Within *processing layer*, the author proposed data mining, statistical analysis and process mining techniques to extract information and visualize them inside the next layer, *visualization layer*.

The authors in (Terragni et al., 2019) set an objective to provide high-level of care for patients. Accordingly, researchers in this project used the **CRISP-DM methodology** and applied process mining techniques to obtain patient pathways and the trajectory patterns for patients from certain accidents to receive definitive care. Within their results they have gathered several pathways; in addition, they have used conformance checking techniques to validate the major pathways that represent patient behaviors.

In order to extract more informative process models, the authors in (Ijcsis et al., 2019) proposed a method which gets the data from a HIS and creates the XES log files, then it uses process mining techniques to extract the process models. The base model (behavior) will be filtered by the unconventional log files and an objective filtering of events to reach an informative model. For instance, in their case study they have used a filter for long-running cases (cases with a duration greater than 380 days). As presented,

the filtering approach is manual. In their study, it seems that they have used DISCO software².



#3-A decision based on the state of the art

In summary, most of the presented works focused on **process discovery** with a vision to get a **control-flow** perspective. Evidently, most of these works are presenting ad-hoc approaches.

However, in this dissertation, the objective was formed around the social question of **how to improve patients' pathways**. For this reason, it was necessary to obtain a **descriptive reference behavior** as the common pathways. The existing methods struggled for fulfilling this need. Therefore, the new process discovery algorithms presented in chapter 4 will target this gap.

2.4.2 Enhancement notation (activity) of business processes in healthcare

In previous surveys it has been seen that this activity of process mining did not receive lots of attentions in the literature. For instance, W. Yang et al. (2014) have addressed this problem by covering 37 studies from 2004 to 2013. According to their research, out of all of the studies only 5 research works addressed enhancement of business processes.

An example of the improvement actions by process mining is the work by Garg et al. (2009). They have presented a non-homogeneous Markov model to efficiently extract the frequent patterns. This method can also identify pathways having high expected cost and high readmission rate; these pathways require more attention to ensure optimum resource allocation (W. Yang et al., 2014). It can help the medical institution to evaluate the effectiveness of the improvement policy. However, it can provide limited information.

In the recent years, the evaluation and improvement actions are regarded as the **re-design of the processes**. However, the **root causes** that triggered the improvement are always ignored, leading to too many difficulties in the improvement phase.

Recently, Van der Aalst discussed the association of process mining and discrete event simulation in (W. M. P. v. d. Aalst, 2018). This work mainly focused on the importance of discovering **executable process models** for discrete event simulation .

Discrete event simulation could be considered as a promising analyzing method in healthcare.

In (Orellana et al., 2018) researchers focused on capturing the different alternative for executing processes in a hospital environment. To do so, they have integrated a new information system named as XAVIA HIS for integrating a plugin in charge of discovering different process variants with the ProM tool.

Duma (2019) identified three health care delivery problems to illustrate and develop an approach toward them. These problems belong to surgical pathways, and emergency care pathways.

Both problems related to the pathways, were related to the uncertainty in processes of emergency care, and surgical care. For instance, in surgical cases it is not clear how

²<https://fluxicon.com/disco/>

long it takes to finish an operation. For emergency care, it is not clear when will be the next event, for example when will the next emergency call occur, or when will be the arrival of the next patient. These uncertainty led the very well structured processes or so-called lasagna processes into spaghetti-like processes during the real execution.

Therefore, in order to manage such challenges they have proposed an ad-hoc process mining approach to extract information from historical data of a case study to monitor the evolution of paths for different cases. Therefore, discrete event simulation was used for providing several scenarios for evaluating their improvement solutions.

In another work in ("[ScholarWorks: A Data Analysis Methodology for Process Diagnosis and Redesign in Healthcare](#)" 2019) researchers addressed the issue of lacking adequate attention to the improvement phase of the processes in healthcare. They indicated the need to acquire a data analysis and diagnosis approach in healthcare.

They have highlighted three challenges: a lack of proper guideline for data analysis to help understand clinical processes, the research gap between clinical data analysis and process redesign in healthcare, and a lack of accuracy and reliability in redesign assessment in healthcare

With this in mind they have proposed a 4-step approach to improve processes. These steps are Data preparation, data preprocessing data analysis and psot-hoc analysis.

In (Boersma et al., 2019) researchers associated process mining activities with the DMAIC approach toward reaching an operational excellence. This work is another example of medical experts reaching out for obtaining proper tools to monitor and improve transparency about medical processes.

Authors in (Dahlin et al., 2019) provided a comparison between **process mining and process mapping**. They highlighted that these two paradigms are sharing a mutual goal of process improvement, however, it is not clear how to go from mining and mapping a process into improvement phase, therefore, they have proposed certain propositions to help future research work to fill this gap.

For instance, they have mentioned the **root-cause analysis** as an important point to detect the cause of problems.

In another work by Miranda et al researchers aimed to monitor the efficiency of operation management in a hospital by a complex network approach. In parallel they have applied process mining techniques to track patient pathways within different departments (Miranda et al., 2019). The presented approach helps to monitor the time evolution of the networks by quantifying certain important parameters about the objectives of each department. They mainly tried to map the relations among different departments not a particular patient per-se. Therefore, process mining helped them to discover the patterns among all the monitored departments.

According to Partington et al. (2015), healthcare processes are time-sensitive. Moreover, the length of waiting times between activities can be a significant driver of cost (Kim et al., 2019). Therefore, the deviations in process cycle time must be monitored.

In addition, the steps involved in healthcare processes are nonlinear and the use of conformance and enhancement process mining techniques seems to be currently underutilized in the healthcare field. Based on a literature review by Partington et al. (2015), most studies focused on control-flow perspective (89%) , only 25 % looked into time-perspective of processes. Roughly, 11 % reported the organizational perspective and only one article proposed the comparative analysis of organizations.



#4-A decision based on the state of the art

It seems that further works are required to help the enhancement activity of process mining. Above all, the **business process diagnosing is completely missing**. This means experts are able to apply process mining to see the AS-IS situation, but, they are not able to detect the causes of deviations in processes. Without recognizing these causes how one can improve a process?

Inspired by these highlighted gaps, this research work has focused on proposing novel methods for enhancement activity of process mining. Chapter 5 will invest heavily to present these methods for evaluating the performance and quality of patient pathways.

2.4.3 Conformance checking activity in healthcare

As shown in figure 2.6, the attention to this subject was considerably low.

However, in (Gatta et al., 2018) the authors evoked three challenges for construction of an effective event log which is suitable for certain conformance checking activities. These challenges are: (i) the ambiguity of involved healthcare data related to different aspects of a patient's process. (ii) The availability of data that are coming from several heterogeneous HIS. (iii) Existence of several communication data sources for data representation.

They have developed several conformance checking techniques in 'R' thanks to a model-driven engineering approach to support generation of event logs from clinical data sources. They also used **PMineR package** for discovering process models.

In addition, the authors in (Duma, 2019) have addressed conformance checking for evaluating the quality of their process models, but **not in a sense to diagnose deviations and improve the efficiency of processes**.

After exploring the literature of process mining notations and activities regarding healthcare data, in the following the indoor localization systems will be briefly introduced.

2.5 Indoor Localization Systems (ILS)

The benefit of ILS application in this work is toward gathering the necessary location event logs for discovering process models which initially have been mentioned as patient pathways.

Indoor localization stands for a procedure in which the location of a device or an object is obtained within an indoor facility (Amundson et al., 2009). It is only less than a decade that the application of ILS has began its growth. This is derived by the growth of IOT (internet of things). ILS is being used in many sectors such as smart cities, smart healthcare management, manufacturing, shops, airports, and the disaster management (Adame et al., 2018; Al Nuaimi et al., 2011). So far, localization systems mainly consist of three components; **transmitters, receivers, and a location engine**. The transmitter could be a device such as a tag which emits the signals. The receivers or reference nodes such as antennas or beacons can receive these signals. The system would identify the location of the objects thanks to the localization techniques developed in the location engine.

There are several prominent review articles regarding the ILS topic for readers who are interested in studying ILS such as the work in (Amundson et al., 2009). They have collected a research regarding the wireless sensor networks for both indoor and outdoor applications. Zafari et al. (2017) presented a collective understanding on localization systems, techniques, and technologies. Al Nuaimi et al. (2011) explained the challenges of ILS and different types of feasible and existing solutions.

There are several localization techniques existing in the literature such as: RSS (Received Signal Strength) can estimate the distance among transmitters and receivers by measuring the strength of signals when they reach the receivers.

CSI (Channel State Information) is almost similar to RSS, however, CSI results in more robust analysis due to its ability to cope with different frequencies and pairing several transmitters and receivers at the same time (Zafari et al., 2017).

AoA (Angle of Arrival) needs more elaborated infrastructure, but it provides accurate results. AoA estimates the angle at which arrays of signals arrive at receivers (Zafari et al., 2017).

ToF (Time of Flight) or also known as ToA (Time of Arrival) uses the transmission time of signals.

TDoA (Time Difference of Arrival) estimates the position of transmitters by measuring the differences among arrival time of signals from a transmitter to multiple receivers.

TWR (Two Way Ranging) uses the ToF and multiplies the propagation time by the speed of light to measure the distance between the transmitter and the receiver.

Triangulation is a method to locate radio transmitters and it works by measuring different attributes like Radial Distance, Direction of the received signal from multiple sources (Cotera et al., 2016; Curran et al., 2011).

Trilateration is a mathematical technique. This approach calculates the distance of a point from several known geometrical entities.

RToF (Return Time of Flight) measures the propagation time of signal while it is emitted from the transmitter to the receiver and returns to the transmitter (i.e, transmitter-receiver-transmitter) (Zafari et al., 2017).

PoA (Phase of Arrival) considers differences among several phases at which signals arrive to the receivers.

These techniques are being applied thanks to multiple technologies like: WiFi, Bluetooth, Zigbee, RFID (Radio Frequency Identification), UWB (Ultra Wide Band), Visible Light, Acoustic Signal, and Ultrasound.

This research work benefits from the technology of Maple High Tech company³. The technology proposed by this company applies TWR and TDoA thanks to a UWB-based technology to provide users with real-Time location of objects. Their application is capable of both indoor and outdoor localization thanks to use of GPS technology.

Chapter 3 will illustrate the application of ILS and **structure of location event logs** in more depth.

³<http://www.maplehightech.com/en/index.html>

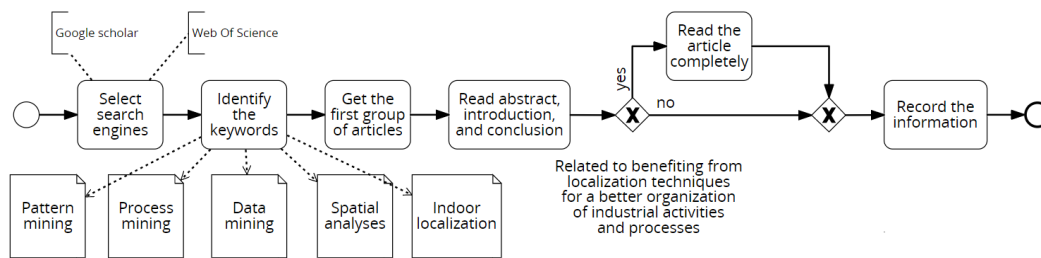


Figure 2.7 – The approach for analyzing research works related to applying localization techniques for a better organization of processes and activities in the industry of the society in general.

2.6 Data mining and process mining have met location data before:

This section presents a systematic literature and technological review in a five-year (2015-2019) period. The objective for choosing this period was to recognize the recent trends in the field of applying location data for data and process mining.

Figure 2.7 shows the approach for analyzing the literature related to the works that have applied localization techniques.

Google scholar, and Web Of Science were the two search engines used for this purpose. Several keywords were used such as *process mining*, *location data*, *indoor location/localization systems*. As a result 42 research works were recorded. We tried to ensure about the proximity of the obtained works to the purposes of this research work. Therefore, by skimming the abstract, conclusion and introduction of these works, 31 out of 42 were selected for final literature analysis.

Unfortunately, the **association of process mining and ILS did not receive lots of attentions**. However, several data mining applications were addressing **pattern and trajectory mining**. These works were selected due to the nearness of these methods to the applied algorithms for discovery of the end-to-end process-like patterns.

The objectives for this analysis were to see:

1. What is the **recent trend for using location data**?

For this, we have used two criteria:

- Research works introducing novel approaches for pre-processing, and interpreting location data.
- Research works that aimed only to extract knowledge from location data.

2. What are **the similar to process mining applications**?

For this purpose, we analyzed the literature based on:

- Research works addressing process-like and trajectory pattern discovery.
- Research works applying conformance checking to detect deviating behaviors.
- Research works using the location data for enhancing the outcome of business processes or evaluating the performance of different patterns.

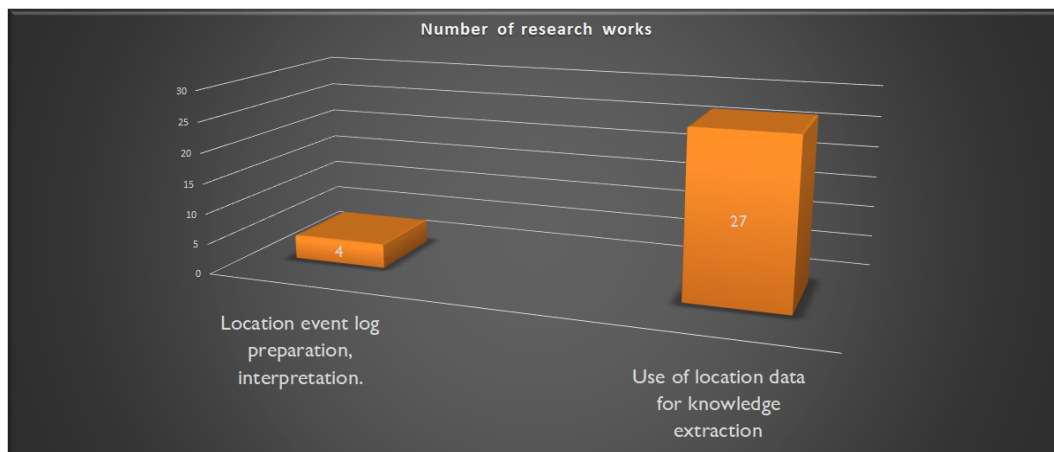


Figure 2.8 – Analysis of the literature of process mining, indoor localization systems, and data mining relevant to the approaches in which the location data were used.

In some cases researchers worked on several of the mentioned criteria. For instance the work in (Muzammal et al., 2018; Senderovich et al., 2016; Wan et al., 2017), it addresses both mining process-like patterns and how to interpret the location data.

According to the presented analysis in figure 2.8, the recent trend is mainly on **the extraction of knowledge from location data**. On the other hand **the interpretation and preparation of location data were almost completely neglected despite its importance**.

#5-A decision based on the state of the art

The processing of location data is not only a technical challenge, but it can be seen as a scientific challenge as well. Due to the high ambiguity existing in location data, it is not feasible to extract proper information from event logs. This ambiguity can be caused by the way these data are being registered.

These data are recorded by **coordination**. They are not representing activities and they only illustrate the positions of tags. Therefore, researchers need to address **how one can interpret this type of data**. Simply put, **sense-making is missing in the literature**.

Some researchers insisted on this important step as a need to provide an abstraction of what exists in the location data (Senderovich et al., 2016).

Further down in chapter 3, it will be highlighted that there is a need to link location data to the actual concepts in the monitoring environment (c.f. figure 3.1).

This is the motive for presenting the **data state** in chapter 3. Within this state the **DIAG meta-model** would link the information in location data with the actual concepts in an environment.

In addition, the location data interpretation rules would reveal how the concepts in a location data can be interpreted for process mining activities. This will be discussed also in chapter 3.

Figure 2.9 presents for which similar-to process mining application researchers applied location data analyses. By this analysis, it has been inferred that, researchers mainly fo-

	Cited research works
location event log preparation and interpretation	Araghi et al. (2018a), Muzammal et al. (2018), Senderovich et al. (2016), and Wan et al. (2017)
Use of location data for extraction of knowledge	Ertek et al. (2017), Hwang et al. (2017), Mazimpaka et al. (2016), and Szttyler et al. (2016), Bao et al. (2017), Garaeva et al. (2017), Ramos et al. (2017), Rojas et al. (2017a), Tanuja et al. (2017), and Y. Zheng (2015b), Aryal et al. (2017), Feng Ling et al. (2016), Tanuja et al. (2016), and Zhenjiang et al. (2017), Blank et al. (2016), Fernandez-Llatas et al. (2015), Lamr et al. (2016), and Y. Zheng (2015a), Jin et al. (2015), Liao et al. (2015), Martinez-Millana et al. (2019), Miclo et al. (2015), and Tang et al. (2015), Araghi et al. (2018a), Araghi et al. (2019), Araghi et al. (2018b), Dogan et al. (2019a), and Namaki Araghi et al. (2018)

Table 2.1 – Cited research works that addressed the knowledge extraction from location data and those which addressed interpretation of these data.

cused on the **pattern discovery activity**. Evaluating the performance and conformance checking were almost absent.

Table 2.2 presents the cited works that presented methods for *discovering patterns*, *conformance checking* and the *performance evaluation and enhancement*.

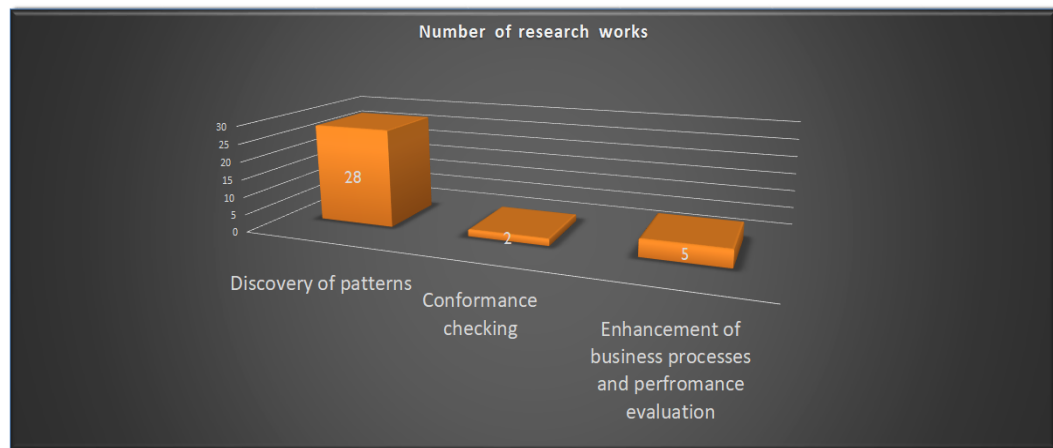


Figure 2.9 – Analysis of the literature relevant to three criteria; discovery of patterns, conformance checking, and enhancement of performance.

Figure 2.10, shows from 2015 until 2017 **almost none of the works mentioned how to evaluate the performance** of processes or trajectory patterns by using location data. This trend has changed since 2018.

The following will present briefly some of the significant information extracted from these works.

	Cited research works
Discovery of patterns	Araghi et al. (2018a), Araghi et al. (2019), Araghi et al. (2018b), Aryal et al. (2017), Bao et al. (2017), Blank et al. (2016), Dogan et al. (2019a), Ertek et al. (2017), Feng Ling et al. (2016), Fernandez-Llatas et al. (2015), Garaeva et al. (2017), Hwang et al. (2017), Jin et al. (2015), Lamr et al. (2016), Liao et al. (2015), Martinez-Millana et al. (2019), Mazimpaka et al. (2016), Miclo et al. (2015), Muzammal et al. (2018), Namaki Araghi et al. (2018), Ramos et al. (2017), Rojas et al. (2017a), Senderovich et al. (2016), Szttyler et al. (2016), Tang et al. (2015), Tanuja et al. (2016), Tanuja et al. (2017), Wan et al. (2017), Y. Zheng (2015a), Y. Zheng (2015b), and Zhenjiang et al. (2017)
Conformance checking	Szttyler et al. (2016)
Performance evaluation and enhancement	Araghi et al. (2019), Araghi et al. (2018b), Dogan et al. (2019a), Martinez-Millana et al. (2019), and Namaki Araghi et al. (2018)

Table 2.2 – Cited works addressing process mining activities.

2.6.1 Preparation and interpretation of location data

Despite the lack of attention for this notion, several robust works proposed different methods to interpret the location data prior to the knowledge discovery. In (Senderovich et al., 2016) the authors used a so-called interaction mining to transform the real-time location data into standard event logs for enabling process mining activities. The transformation of event streams was according to the defined notion of interaction. This interaction provides a knowledge layer that links the raw sensor data and process instances.

At first, they identified a reference of the existing information in a RTLS (Real-Time Location Systems) event log.

At the second step, they have defined a so-called RO⁴ log. The RO log creates a new log from the raw recorded data in RTLS event logs. In RO logs there are 4 types of information: ID of the monitored person, type of the process actor (patient, or healthcare staff), room, and time-stamps.

At the third step they defined the AD log. This log relates to the activity data and labels of activities. As can be observed, there is a need in their approach to link the activity instance in the AD log and the observed behavior in the "RO" log to understand the observed data is relevant to which activity. Thus, the authors have defined the "ROAD" solution. This solution has two steps: applying interaction mining on sensor data and

⁴Some of these acronyms are not detailed because their names have not been justified in the corresponding research works.

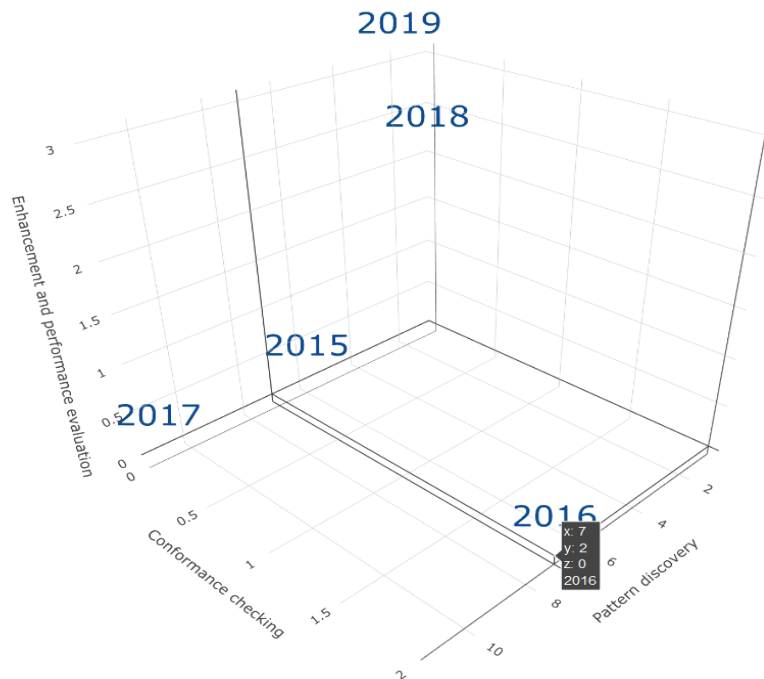


Figure 2.10 – Analyzing the literature of the application of process mining on location data during a 5 year-period. Accordingly, in 2016 the enhancement and performance evaluation did not receive enough attention.

using process knowledge to match the recorded events into the best-match activity. For this matching they proposed integer linear programming. The result would transform a RO log to an AD log.

For this solution they need the process knowledge from the historic data. And also, the user need to know what is the normative process for patients. Accordingly, the researchers have made two assumptions: A process actor can not be involved in two activities at the same time, and each activity instances corresponds to only one interaction.

In (Muzammal et al., 2018), the authors have used the uncertain sensor data and tried to transform these data into probabilistic trajectory data using pre-processing routines. Then, they have used a dynamic algorithmic approach to extract the preferable trajectories. Their approach consists of 3 main steps: *data cleaning*, *data compression* and finally *trajectory mining*. The trajectory creation step which generates togetherness pattern, common path pattern and group patterns and cyclic patterns.

The work in (Wan et al., 2017) presents a framework for deriving information about people's continuous activities from their individual GPS data. The authors have proposed a framework which includes two major techniques for processing GPS data: (i) A fuzzy classification for differentiating the activity patterns. (ii) A scale- adaptive method for refining the location of activities within outdoor and indoor environments. The clustering techniques used here helped to classify which activities are occurring. It helped also to detect the outliers and indicated in which positions an activity is being executed.

As observed, a certain amount of uncertainty exists. This is due to **relying on the classification techniques for discovering an activity**. Additionally, **the lack of a standard framework** for assigning each activity to different zones leads to the need to extract activity-like instances. This lack of standard framework relates to the necessity

of acquiring a data science oriented procedure (as mentioned previously in **decision 1**).

2.6.2 Use of location data for knowledge extraction

This issue attracted the majority of the researchers in this area.

Hwang et al. (2017) used ILS event logs as input of process mining activities to monitor the patterns of shoppers pathways in south Korea. They have applied the inductive mining algorithm to extract block-structured work-flow nets. They observed through changing the shops display the behavior of customers changes. Additionally, the sales and traffic volume within the shop was changed. However, they mentioned the causality between these concepts need further investigation.

Sztyler et al. (2016) applied process mining in association with several sensor data to monitor the patients' well-being and the individual behavior of patients. They aimed at using the process discovery activity to extract the pattern that each individual follows. They also used conformance checking thanks to certain alignment methods to detect the deviations in individual behaviors and an already defined normative model.

Another interesting work is related to mining RFID data for enhancing a schedule-based system (Ertek et al., 2017). This work is not addressing process mining activities, but it has an interesting approach for detecting the time-slots that corresponds to a certain activity.

Mazimpaka et al. (2016) presented a survey about the **trajectory data mining methods**. As expected in their work **they don't address end-to-end process-like patterns**, but they have categorized current methods and approaches to mine the trajectory patterns:

- *Characterization of moving objects*: this group was related to monitoring humans and understanding the transport modes for humans. The used mining methods were classification, frequent pattern mining and clustering.
- The second application was *Discovery of social interactions and relations*. This was applied to case studies for detecting social relationships between individual entities, and detecting the communities. Also, it was used to mine the interactions between animals as well. The discovery methods were: clustering, frequent pattern mining, and group pattern mining.
- The third application was related to the *classification of places or regions on maps*. This was applied for discovering the land usage for economic analyses, detecting hot-spots and traffic jams. The used methods were clustering, classification and flock pattern mining. Also it was applied for mining the relationships among different regions.
- The fourth application was for *detecting social events*; clustering and classification were the applied methods.
- The fifth is relevant to *trajectory-based prediction*. For this, case predictive statistical methods were used. Example of such applications was for prediction of traffic jams.
- *Trajectory-based recommendation* is another application of data mining methods on location data. This is used for trip recommendation, place recommendation, and etc. Frequent pattern mining was used in this application.

The work by Rojas et al. (2017a) addressed the use of location event logs for process mining. Their approach mainly focused on discovering of process models for emergency departments. They have used previously mentioned PALIA tool for the process discovery actions.

The authors in (Y. Zheng, 2015b) have used the location data and by using a tensor calculation method tried to present an efficient trajectory data processing model for mining the hot routes.

Tanuja et al. (2017) presented a new algorithm for monitoring the transportation data in a case study. They have applied a clustering approach. In (Ramos et al., 2017), they have developed a systems thanks to several clustering and sequence clustering methods to guide people with optimal guidance for their daily traveling paths.

Other approaches such as the one in (Garaeva et al., 2017) addressed one of the intriguing data mining methods known as *spatial data mining*. This is related to extraction of spatial relations and interesting patterns from spatial data. They addressed the challenge of co-location pattern mining. Bao et al. (2017) also addressed the co-location pattern mining and they were focused mainly on user interaction with extracted patterns. Basically, they used the user knowledge for gaining the optimal patterns.

Another approach was based on using RTLS data for trajectory pattern mining with an objective to predict the next location of a moving object (Zhenjiang et al., 2017). In their approach they collected RTPT and HT, then they introduced the RTmatch method for calculating the best matching path. This would help to predict the next location of the object.

Still most of the works are driven by data mining methods specially spatial data mining. Some approaches addressed using location data for transportation and traffic control. By this approach, analysts were able to detect the hot patterns that accidents occur (Lamr et al., 2016). Similarly in (Aryal et al., 2017) the authors proposed a shared nearest neighbor clustering method to find the dense clusters which relates to the movements of taxi cabs in New York city. Another example for monitoring behaviors of people was related to a case study in china for mining the patterns in which tourists are travelling (Feng Ling et al., 2016). In their paper they have used a hierarchical analytic method.

In (Blank et al., 2016) authors used process mining and location event logs to extract the spaces in the facility that are considered as hot zones. Their contribution is a new algorithm for detecting the ID's that are following similar process like patterns.

One of the similar research works to the one presented here is the work by Llatas et al in (Fernandez-Llatas et al., 2015). In this article they presented a process mining-based methodology for extracting information from the location data. Accordingly, they have applied the PALIA algorithm within the PALIA suite application.

As a following work in (Martinez-Millana et al., 2019), the authors have presented a paper to evaluate features of a process mining-based dashboard that can receive location data as inputs. Within their dashboard they have analyzed processes based on different methods such as clustering. Once again, it is not clear how to obtain location data and prepare it for performing process mining.

Another similar approach is presented in (Miclo et al., 2015). The authors have used Disco process mining application to apply on a set of simulated location data to extract models of processes. There is no clarification on how to extract the location data, and how to link the real-time location systems directly to a process mining application. The authors also aimed at presenting **business process simulation as a step after process**

discovery which would not be possible since they are not extracting process models with execution semantics.

Tang et al. (2015) proposed a framework for mining trajectory patterns from uncertain data that exist in cyber-physical systems . Their solution called as "LiSM" (Line-in the sand miner) tries to discover the proper patterns from untrustworthy data.

LiSM constructs a surveillance network from sensor data and measures the locations of intruder appearances based on the link information of the network.

The system recalls a cone model from the historical trajectories to track multiple intruders. At the final step, the system ensures the extraction outcomes and updates sensors' reliability scores in a feedback process.

Additionally, they proposed the LoRM algorithm which filters and refines the data to reduce the distance of computational overhead on road networks and it uses a method to detect the shortest distance to measure and detect intruders of system.

Dogan et al. (2019a) presented results of monitoring customers paths analysis in a shopping mall . They have used Bluetooth-based technologies. They have aimed to monitor customers behavior based on their **gender**. Therefore, they used the a virtual zone as the bathroom to extract the gender of the customers by whether choosing the male or females' bathrooms. Based on their results, the processes' features are different for the two groups of male and female customers. For instance, the duration of processes for male and female customers are different.

Nevertheless, in order to extract meaningful information from location event logs several main challenges exist.

Martin (2019), has raised several difficult questions from a conceptual angle with regard to location data integration in hospital information systems (HIS). These challenges are:

1. **Presence of data quality issues in ILS data**, for instance the absence of case identities, and problems in *determining the start and end of an activity*.
2. **Presence of human errors when initializing the data gathering**, for example when a staff member makes a mistake by importing the wrong ID tag for tracking a certain patient.
3. **Need for acquiring a systematic way to obtain domain knowledge**. This is relevant to understanding what activities occur in each zone of a hospital. Also, being able to address the nature of the processes, since for example, the processes in healthcare organizations are different than the ones in manufacturing plants.
4. **The integration of HIS data and ILS data needs to be performed in a semi-automated way**. There is a need to get the domain knowledge and append it with the mined processes.

 #6-A **decision based on the state of the art**

In light of these challenges, it is important to investigate different approaches for interpretation and integration of location event logs.

The first two challenges were either technological or social obstacles. They are not addressed here. This research work focused on proposing proper solutions for the third and fourth mentioned challenges. Accordingly, next chapter will present the **configuring the environment and systems** function within data state of DIAG methodology. This function would aid the data and business experts to capture and structure the domain—healthcare—knowledge. This Function uses the DIAG meta-model which supports the task of interpreting location event logs. It has also the potential of integrating the HIS (hospital information system) and ILS data.

Additionally, next chapter will cover the interpretation of location data within the context of DIAG methodology.

II

Second part: DIAG Methodology

DIAG methodology statement

Introduction

This part of the dissertation focuses on a proposed methodology for **extracting knowledge from location event logs**. This methodology is labeled as **DIAG**, and it has four different states; **Data state**, **Information state**, **Awareness**, and **Governance**.

Figure 2.11 presents this methodology. As shown, every state consists of one or several functions. These functions are defined to undertake the task of transforming *location data* from one state to another.

During this action, the level of understanding that one can conceive from these data evolves; state by state.

In the second part of the dissertation, each chapter presents these functions in the context of the mentioned states.

Defining DIAG as a methodology is supported by the definition in (Ishak et al., 2005), where the author defines a methodology as “ a systematic, theoretical analysis of the methods used to a field of study. This analysis includes the theoretical analysis of the **body of methods** and principles associated with a **branch of knowledge**. Typically, it encompasses concepts like **paradigms**, **theoretical models**, **phases** and **quantitative or qualitative techniques**.”

Background

Multiple process mining methodologies have been introduced within the literature. For instance, the authors in (Mărușter et al., 2009) introduced a methodology to better associating process mining and simulation. Their approach was focused on pre-processing the data with an objective to get better quality results.

Another reliable example is the work by Bozkaya et al. (2009). Their approach consisted of six different stages. The first stage is the *log preparation* in which the mined event log will be presented. Secondly, at the *log inspection*, the experts try to get a primary understanding about the process. Next stage is *control flow analysis*. At this stage the event log will be checked to ensure if the **descriptive model** is close to a **normative model**. The fourth stage is *performance analysis*. This stage involves the detection of bottlenecks and calculating the process cycle time. Next, in the *role analysis* stage, the

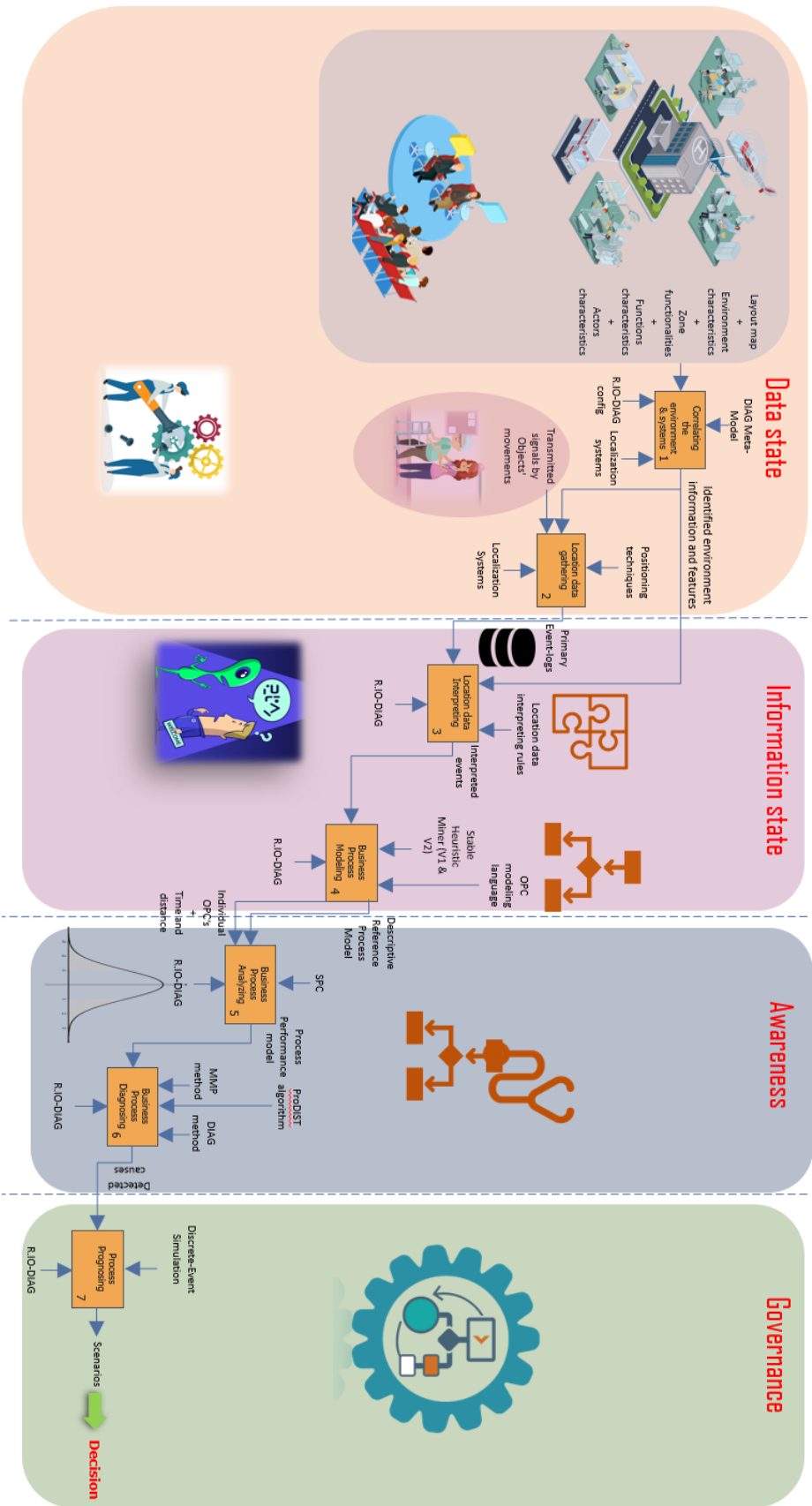


Figure 2.11 - The DIAG methodology.

user gets an organizational perspective. Finally, the last stage is the *transfer of results*. This stage delivers the outcomes of the methodology to the process owner in a way that he or she can comprehend the results.

Another proposed methodology is within the process mining manifesto by W. v. d. Aalst et al. (2012). The authors proposed five stages; *justification and planning*, *extract*, *create control-flow model and connect to event log*, *create integrated process model*, and *provide operational support*.

The authors in (M. L. v. Eck et al., 2015) suggested a six stage methodology which is called as PM^2 . These stages are *planning*, *extraction*, *data processing*, *mining and analysis*, *evaluation*, and *process improvement and support*.

Several other methodologies have been developed for the application of process mining in healthcare as well; examples are (Cho et al., 2014), (Fernandez-Llatas et al., 2015), (Rojas et al., 2017b), and (Johnson et al., 2019).

By a regard to all of the presented approaches, there is still a missing part of the puzzle. To best of our knowledge, there is **no detailed methodology for extracting knowledge from location data**. Such a methodology should start at the very beginning by **dealing with the non-refined data** until providing results for **making data-driven decisions**. This issue in chapter 2 has been raised by the “**decision 1**”.

The closest approach that has been cited in this work is the methodology presented in the work by Fernandez-Llatas et al. (2015). The authors presented an approach with seven steps to extract a declarative model of processes. Their approach has been defined in the context of a tool which directly transforms the location data into the **model-based analysis**. The first step is the *setup of the localization system*. The second one is the *data gathering*. At the third step, they have defined a phase as *semantic aggrupation* of areas which is concerned with identifying the zones and rooms in the hospital. Fourth, the *process filtering* focuses on the process mining activities. The process owner who designs the virtual zones will select the samples of the events that will be used to perform a process discovery in the later steps. In the *process discovery* phase, the tool uses different filtering options and algorithms for process discovery. After in the sixth step, two tasks will be done; *conformance checking* and *process enhancement*. They illustrated their research by the PALIA-ILS SUITE application.

However, it is not clear **how one should interpret different concepts within the location event logs**. **What should be the interpretation approach for dealing with noisy and fuzzy data within location event logs?** **How the enhancement function works?** **How one can diagnose a problem?** It should be noted that, without diagnosing actions, it is not clear which aspect of the process should be improved.

Objectives

Such gaps in the literature of process mining and ILS became the main motive behind this research work to present the **DIAG methodology** which begins by providing a supporting framework known as the *DIAG meta-model*, and it finishes by generating multiple scenarios for an optimum decision making.

Table 2.3 presents a general summary about the mentioned challenges and the proposed contributions within the DIAG methodology.

The represented chapters in this part of the dissertation are illustrating different states and involved functions within the DIAG methodology through development of a tool named as **R.IO-DIAG**.

Second part: DIAG Methodology

The scientific challenge	The contribution	The corresponding state in the methodology	The corresponding function in the state	Chapter
Devising a formal framework for associating a localization system and a process mining application.	DIAG meta-model developed in R.IO-DIAG application	Data	Configuring the environment and the systems	3
Direct interpretation of location event logs for process mining.	Location data interpretation rules	Information	Location data interpreting	3
Discovering the descriptive reference process model that describes the common behavior registered in an event log.	Stable heuristic miner V1 & Stable heuristic miner V2 algorithms	Information	Business process modeling	4
Distinguishing types of activities in a process model.	OPC process modeling language	Information	Business process modeling	3
Quantitative evaluation of the process performance.	Application of statistical process control	Awareness	Business process analyzing	5
Measuring the distance between a descriptive model and a normative model.	ProDIST algorithm	Awareness	Business process diagnosing	5
Automatic diagnosing of processes.	MMP method & DIAG method	Awareness	Business process diagnosing	5

Table 2.3 – A general overview of the scientific and technical challenges and the corresponding solutions applied by the DIAG methodology.

The next chapter is investigating the mentioned challenges and it starts by presenting the **Data state** and the *location data interpretation rules* within the **Information state**.

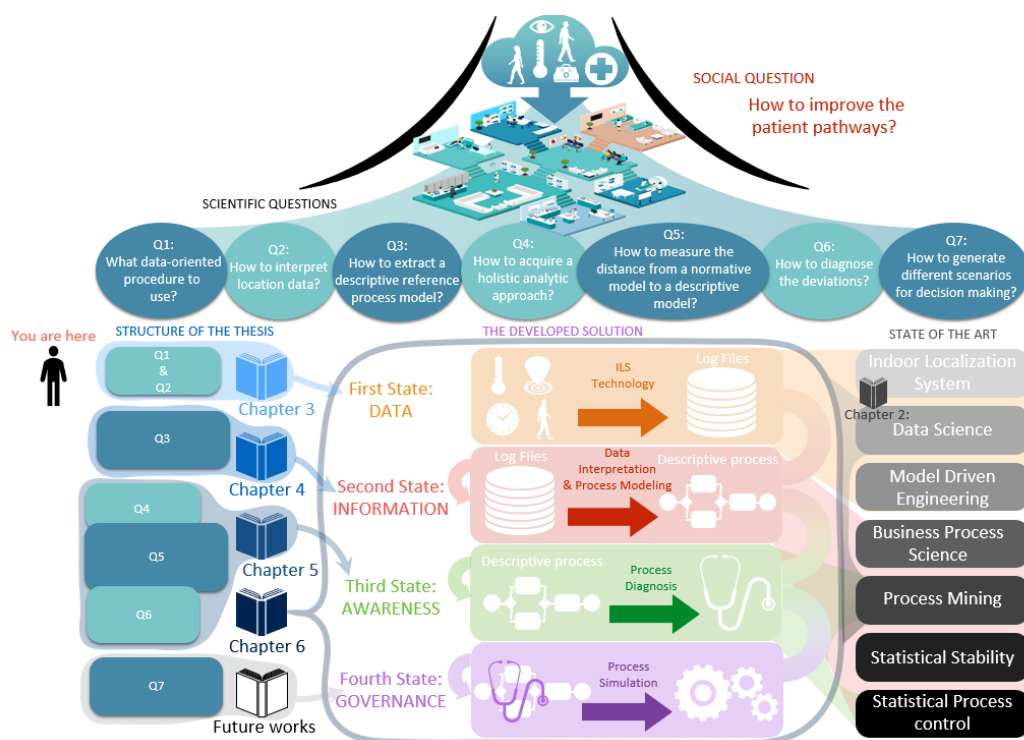


Figure 2.12 – The overall picture of the dissertation. A map for the readers.

3

From the Data to the Information state

3.1	Introduction	49
3.1.1	Why Data state is defined in the DIAG methodology?	50
3.1.2	How the connection between the location event logs and the actual concepts in a process can be constructed?	51
3.2	Function 1: Configuring the environment and systems	52
3.2.1	DIAG meta-model: a framework to support process model discovery from location data	52
3.2.1.1	The Process and Location Event-log packages	54
3.2.1.2	The Function and Healthcare Functions packages	54
3.2.1.3	The Organization, Healthcare Resources, and Objectives packages	56
3.2.1.4	DIAG meta-model (first version)	57
3.2.2	R.IO-DIAG initialization	59
3.3	Function 2: Location data gathering	65
3.3.1	Objects movements	65
3.4	Function 3: Location data interpreting function	69
3.4.1	Add start-event	71
3.4.2	Add end-event	72
3.4.3	Add task-event	73
3.4.4	Add knowledge on start-event	76
3.4.5	Add knowledge on end-event	77
3.4.6	Add knowledge on task-event	78
3.4.7	Operation Process Charts (OPC)	78
3.5	Recap	82

“It’s easy to have faith in yourself and have discipline when you’re a winner, when you’re number one. What you’ve got to have is faith and discipline when you’re not yet a winner.”

Vince Lombardi

3.1 Introduction

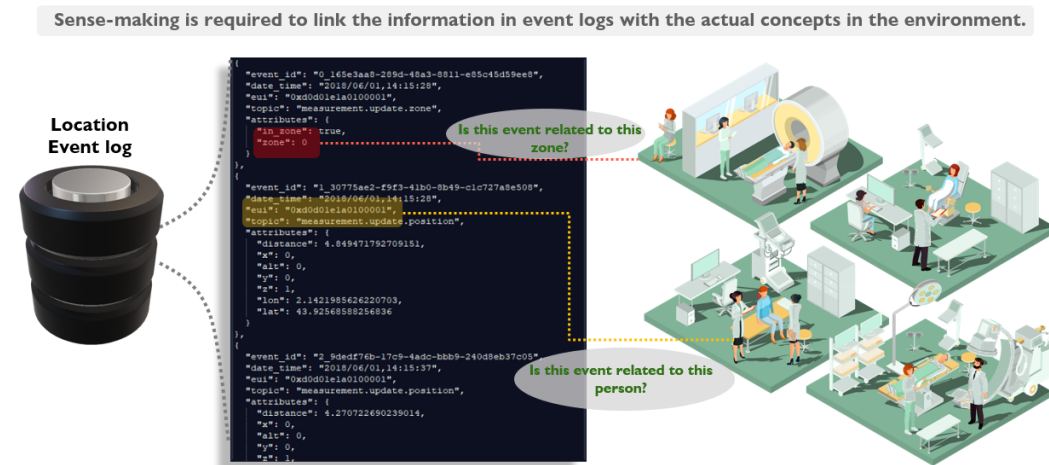


Figure 3.1 – Illustrating the need to provide a sense-making procedure for the recorded information in location event logs.

Imagine this scenario “a patient has arrived in a hospital to visit a certain physician. Assuming that the required personal information are registered and a location tag is given to the patient for tracking his or her process. The objective is to ensure about the safety of the patient and the performance of the healthcare process to meet the patient’s expectations.”

Within this scenario several concepts are presented such as: a patient, a hospital, personal information, a tag, a process and its performance.

All of these concepts contain useful information for visualizing patients’ processes. In order to automatically track and analyze a patient’s process, one should consider the key concepts and the relationships among them.

These concepts can derive adequate information to monitor patients’ processes. This is an important issue, because from the human perception, it is easy to comprehend—in this scenario—what is the process, what are the corresponding activities, whom we are monitoring and etc. However, to automatize the task of patient monitoring and analyzing healthcare processes, **every concept and its relevant relations should be identified** for the systems.

In view of this statement, this chapter presents the prerequisite actions for configuring the systems prior to receiving the location data as the input. For the purpose of presenting this chapter, three questions are defined:

- Why **data state** is defined in the DIAG methodology?
- How to link the information in **location event logs** with the **real concepts** in a hospital?

3.1.1 Why Data state is defined in the DIAG methodology?

The main reason for defining data state is the need to provide a **sense-making** approach for location data interpretation.

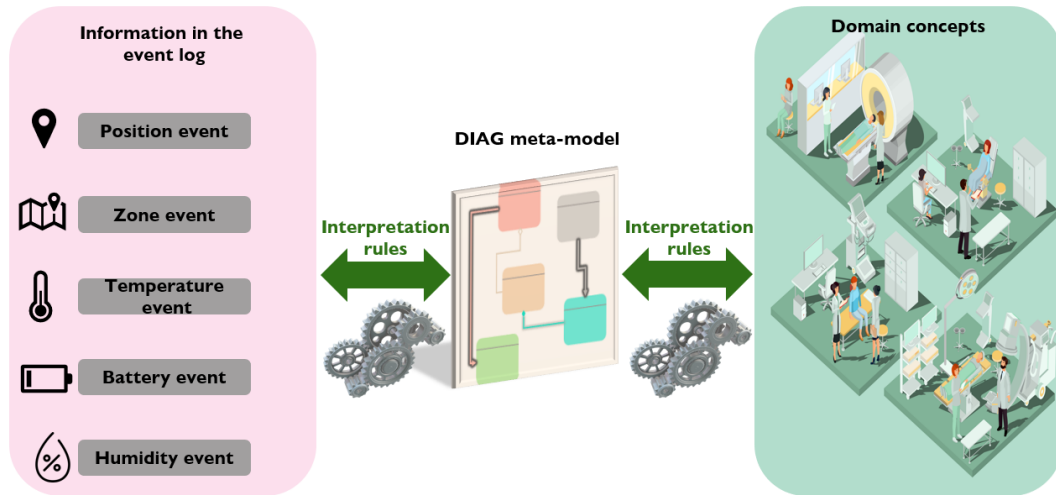


Figure 3.2 – Representing the approach for detecting the existing concepts in the location event logs.

Assume that a location event log similar to the one in figure 3.1 is given for further analyses (in this research work the extracted event logs have JSON format). By focusing in such event logs, one could observe that the given data is vague. This means it is not clear which recorded information belongs to the patients, which belongs to the healthcare staff, and what is the corresponding activity. In that, there exist many other blurred information that should be detected. In fact, there is a considerable gap to link the actual concepts in patient—pathways—processes with the registered data in the location event logs.

Figure 3.1 embodies the motive for defining the data state in the DIAG methodology. This is related to the need for defining a **sense-making** approach.

3.1.2 How the connection between the location event logs and the actual concepts in a process can be constructed?

In order to provide a sense-making approach, this research work proposed to apply a model-driven engineering approach. Such an approach reveals all of the required concepts for monitoring patient pathways and it demonstrates the relationships among them.

As a result of this approach, the **DIAG meta-model** and **location data interpretation rules** are presented as contributions of this chapter. This meta-model is evolving through the development of this research work. Yet, it mainly became the primary stone of the foundation of this research work. A detailed publication also covers this solution (Araghi et al., 2018a).

In addition, a series of interpretation rules are required to interpret the extracted data with the added knowledge coming from the the meta-model(c.f. figure 3.2).

DIAG meta-model is used within the *data state* to support the interpretation of location event logs. In addition, it has been applied for the **business process diagnosing** of DIAG methodology as well; this will be discussed in chapter 5.

As shown in figure 3.3, through the remainder of this chapter, section 2 and 3 will present two main functions in the *Data state*. Section 4 presents the first function of the *Information state*. Within this section the interpretation rules are presented. Finally, the last section concludes chapter 3.

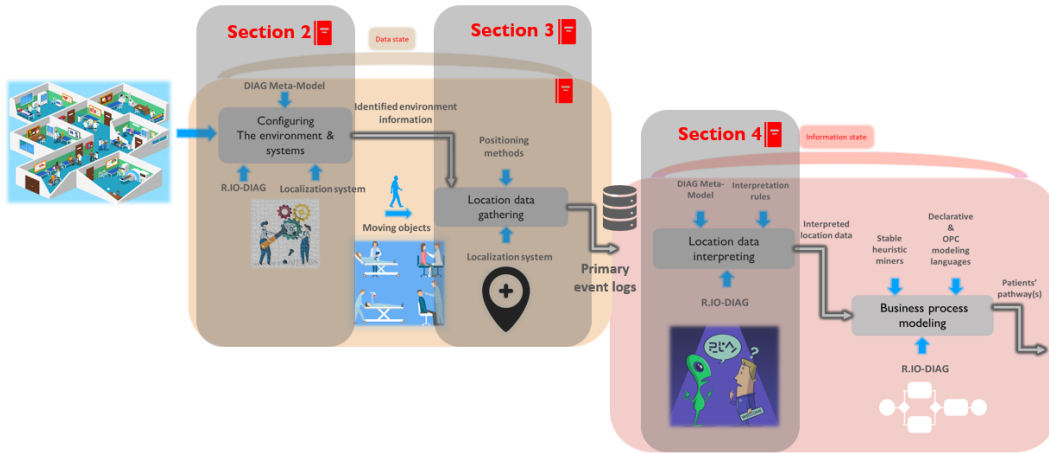


Figure 3.3 – Structure of chapter 3.

3.2 Function 1: Configuring the environment and systems

As shown in figure 3.4, this function employs the DIAG meta-model for constructing a link between two systems. These systems are a localization system, and R.IO-DIAG application which is in charge of extracting meaningful information and knowledge from location event logs.

As **inputs**, this function receives the relevant domain information such as the map of the environment, the existing zones, functions, and corresponding types of process actors. For instance, in case of a hospital, it would receive the map of the consultation departments, the involved functions and services in each department, and the attributes of patients and the staff.

As **outputs**, it would provide a primary knowledge about processes and organization types, what are the corresponding activities, and actors of the processes. This function would heavily support the interpretation of the connections between location event logs and the existing concepts in each process scenario.

3.2.1 DIAG meta-model: a framework to support process model discovery from location data

This meta-model consists of several packages. Before presenting the meta-model itself, these packages will be described. Then, an example illustrates the defined classes and their relationships.

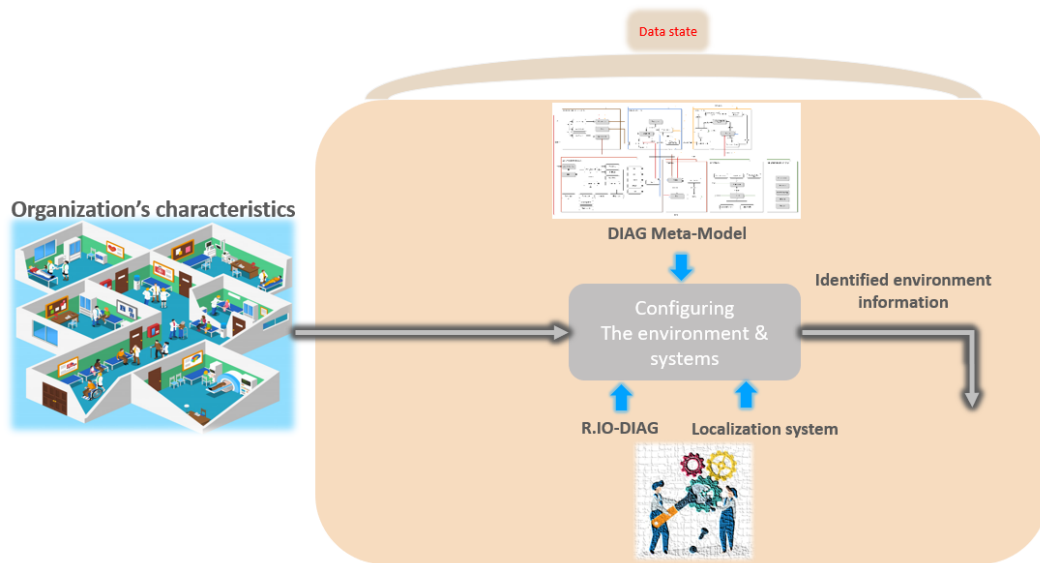


Figure 3.4 – The first function of data state, configuring the environments and systems

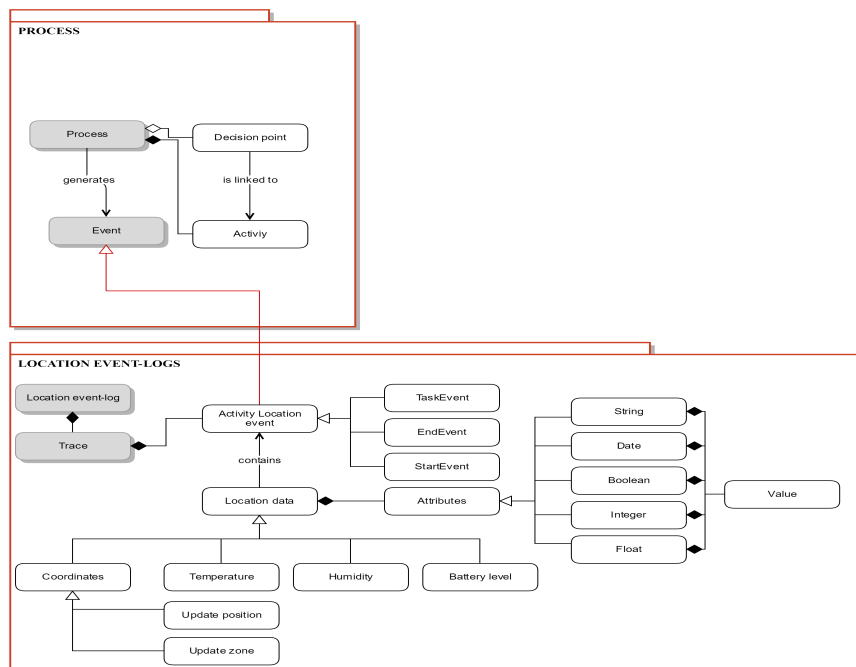


Figure 3.5 – Process package of DIAG meta-model

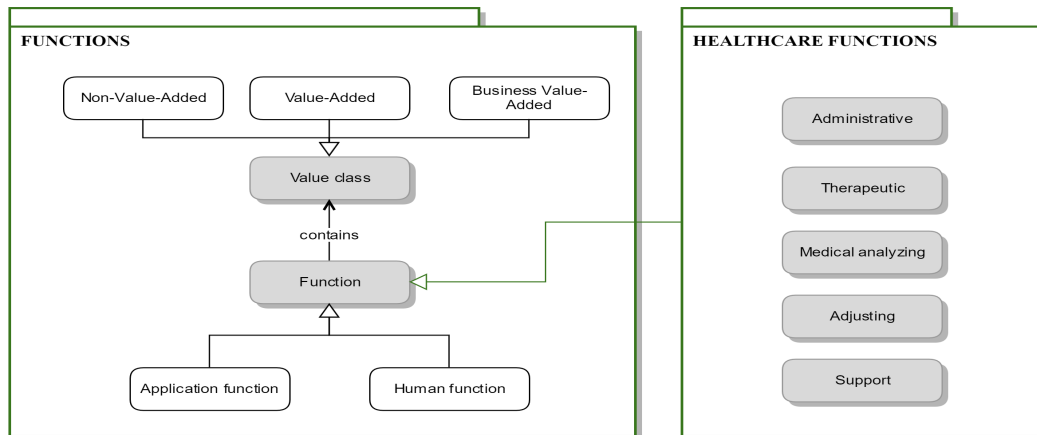


Figure 3.6 – Function and healthcare function packages within DIAG meta-model

3.2.1.1 The Process and Location Event-log packages

These packages (c.f. figure 3.5) simply present the classes that are considered for process modeling from location data. According to the definition provided in (Dumas et al., 2018), a business process consists of *decision points*, *process actors* (like patients or hospital staff), and *activities*.

The class of *process actors* will be expressed inside the **Healthcare Resources package**. Each process execution leads to generation of *events*. These events are either corresponding to a start and end of a process, or they relate to the execution of a task.

The generated location event logs by the indoor localization systems will be used as an input of a process mining application to automatically discover patient pathways. However, the registered information in a location event log is vague and they are not sufficiently suitable for performing process mining activities (Araghi et al., 2018a).

As an example, in a traditional event log which is used for process discovery, each event corresponds to an activity. In contrary, by looking at figure 3.5 one can observe that in a location event log, each event can be related to several information such as; coordination of the tagged objects, temperature, humidity, and the battery level of tags.

Accordingly, each event log is structured by one or several traces. These traces are represented by a series of events. Eventually, these events are relevant to several information in the location data. It is a necessity to interpret and extract the events that are expressing the process activities.

As a result, there must be a prerequisite action to understand which event corresponds to an activity. Section 4 presents the series of [interpretation rules](#) for preparing the location event logs as an input of process mining activities. To the best of our knowledge, such a formal structure for interpreting location data to perform process mining activities has not been addressed in the literature.

3.2.1.2 The Function and Healthcare Functions packages

These two packages are related to the functions that lead to the execution of activities. Figure 3.7 presents these packages. Within the **Function package**, the *human* and *application functions* are considered as the sub-classes. Each function contains a *value class*.

This value class can be either *value-added (VA)*, *business value-added*, or *non-value-added (NVA)*.

1. **VA (Value-Added)**: these functions relates to the activities that generate value from the both patient and the organization's perspectives (e.g visit of a physician).
2. **NVA (Non-Value-Added)**: this class represents activities that have no promising value for the healthcare process, such as waiting to receive a certain treatment.
3. **BVA (Business Value-Added)**: this class is addressing mainly administrative activities. Generally, each activity that is generating value only for the healthcare organization is sorted within this class.

Distinguishing the value class of each function can aid for a better evaluation of the process performance. Evaluating the performance of business processes based on the value class of activities will be investigated in chapter 5. An introduction of the definition of these value classes were presented in chapter 2 as well.

The **Healthcare Functions package** describes all the functions that are possible to be monitored within a medical or non-medical process.

- **Administrative functions (Non-Medical)**: these functions are related to different types of tasks for executing bureaucratic activities in processes. These functions have the Business-Value-Added (BVA) attribute, which means they do not add any value on the process performance level from patients' perspectives. However, these functions are required for hospitals to satisfy their needs. Some instances of these functions are: billing, admission, managing human resources, managing healthcare procedures and policies, and etc.
- **Therapeutic functions**: this class is related to the functions which provide health care treatment to patients. It could be inferred that these functions directly impact *beneficiary class* (patients are categorized in this class). Instances of this class are: physical therapy, speech and language pathology, respiratory treatments, psychology, social services, athlete medicine, nursing, and dietary/supplement services. Evidently, such functions are Value-Added (VA).
- **Support functions**: activities such as logistics in hospital, and maintenance of systems and instruments are categorized in this class. These functions are executed by the *healthcare staff* class and are considered as BVA.
- **Analyzing and diagnosing functions**: in general these functions are related to the medical tests and exams that could help physicians to propose prognostic actions. Instances of these functions are scanning (imaging), medical tests which are take in place in laboratories, emergency medicine, cardiology, and etc. These functions are categorized as **VA**.
- **Adjusting functions**: this research work defines this class to consider functions which are being executed to calibrate the process by its constraints such as lack of a resource in hospital. Instances of this class are functions like waiting, patient transports, and etc.

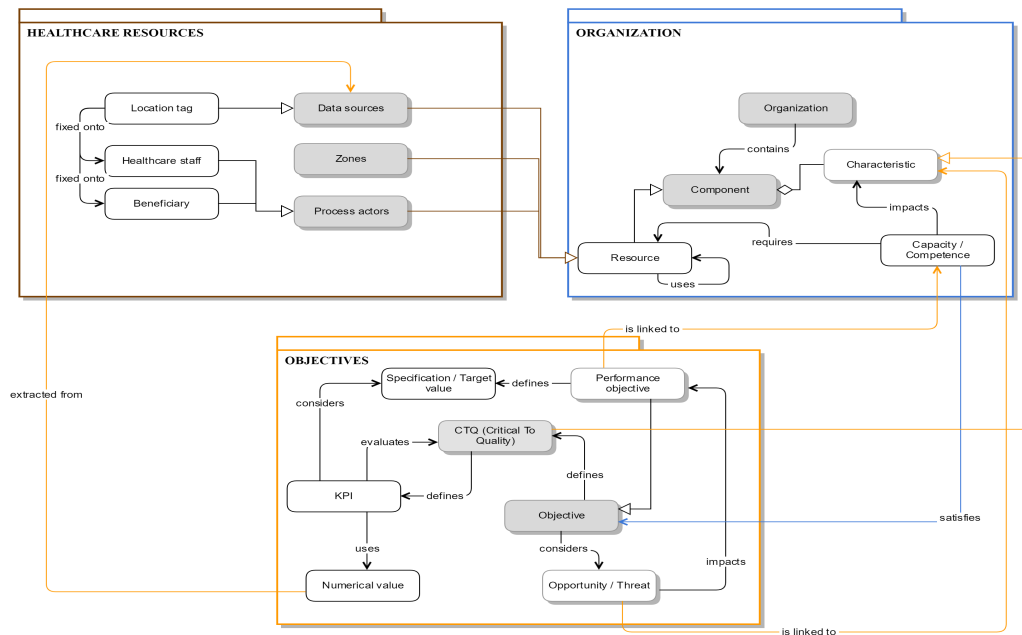


Figure 3.7 – Three of main packages in DIAG meta-model; Organization, Resources, and Objectives.

3.2.1.3 The Organization, Healthcare Resources, and Objectives packages

These are the three main packages in the DIAG meta-model. Within the **organization package**; note that, a main class as *organization* is defined in this package. An instance of such class can be a hospital. Each organization contains several *components*. These components are composed by several *characteristics* which are mostly influenced by the *capacity* and the *competence* of the organization. This capacity is aligned with the *resources* that an organization can acquire.

For instance, in the case of a healthcare organization (based on the presented case study in this research work) several resources are defined in the **Healthcare Resources package**.

Data sources class has *location tag* as a sub-class. Each location tag is given to a *process actor* such as the *healthcare staff* or patients (the *beneficiary* class). As a sub-class of the data source class, the EMR (Electronic Medical Record) can be defined too. Moreover, *zones* in a hospital are seen as the resources of such organization.

The **Objective package**, consists of several important concepts. First, consider the *CTQ (critical to quality)* class. Each product or a service that is the result of a defined process should contain certain quality characteristics. These quality characteristics or CTQs indicates how a client perceives such a product or service.

An important use of CTQ in this research work

The concept of CTQ is one of the main pillars in quality engineering (Montgomery, 2019). CTQ class has three different types. Physical CTQ such as the length,

weight, voltage. Sensory CTQ, like taste, appearance, color. Time-oriented CTQ, such as reliability, durability, and serviceability of an organization.

The CTQ class is a sub-class of the organization *characteristics* and it defines the key performance indicators (*KPI*) which uses *numerical values* to evaluate the quality characteristics. The defined *specification targets* by the *performance objectives* helps to set a desired value for the evaluation of the quality characteristics.

For instance, if one wants to evaluate the “quickness of a process execution”, he or she should consider the concept of quickness as a time-oriented CTQ. Therefore, a KPI such as the “cycle time” of a process can be deliberated. This KPI uses certain numerical values such as “operation time” to evaluate the quickness of the process.

In essence, a target value should be considered as the goal which represents quickness of the process. For example, if the cycle-time is less than a target value of 20 minutes, the process is considered as a quick process.

This will be recalled in chapter 5 for the quantitative performance analyses of processes.

3.2.1.4 DIAG meta-model (first version)

The first version of DIAG meta-model is presented in figure 3.8. This version was devised without the required classes for supporting business process diagnosing function. Later, in chapter 5, the second version of DIAG meta-model will be presented with the support for performing an **automatic business process diagnosis**.

In figure 3.8, the meta-model shows the relationships among different packages. For instance, the *value class* is being used for *KPI's*. These KPI's also use *numerical values* that are being extracted from the *data sources*.

The described packages and their relations can be expressed by the **following example**. Consider a scenario in which a hospital wants to ensure about the satisfaction of its patients. Thus, the *organization class* here is a hospital which has a fixed **objective**. Corresponding to the defined objective the hospital considers a CTQ which is the serviceability of the processes. In this order, a KPI is studied as the **process cycle efficiency** (Namaki Araghi et al., 2018).

Process cycle efficiency is expressed as the total duration of the value-added activities divided by the total cycle time of a process (Namaki Araghi et al., 2018).

If patients spend less time doing *non-value-added* activities; thus, they will be more satisfied with the competence of the hospital. Consequently, the *numerical value* here is time.

Knowing these, the hospital decides to monitor patient pathways by tracking patients movements in the hospital. Therefore, the *process actor* here is the patient which is defined in the *beneficiary class*. The patient uses certain *resources* in the hospital such as the *zones*.

Therefore, each time a patient enters a zone, he or she does certain **functions** to run the process *activities*. While the process is running, it generates *events*. These events are considered as the *data sources* for measuring patient satisfaction levels.

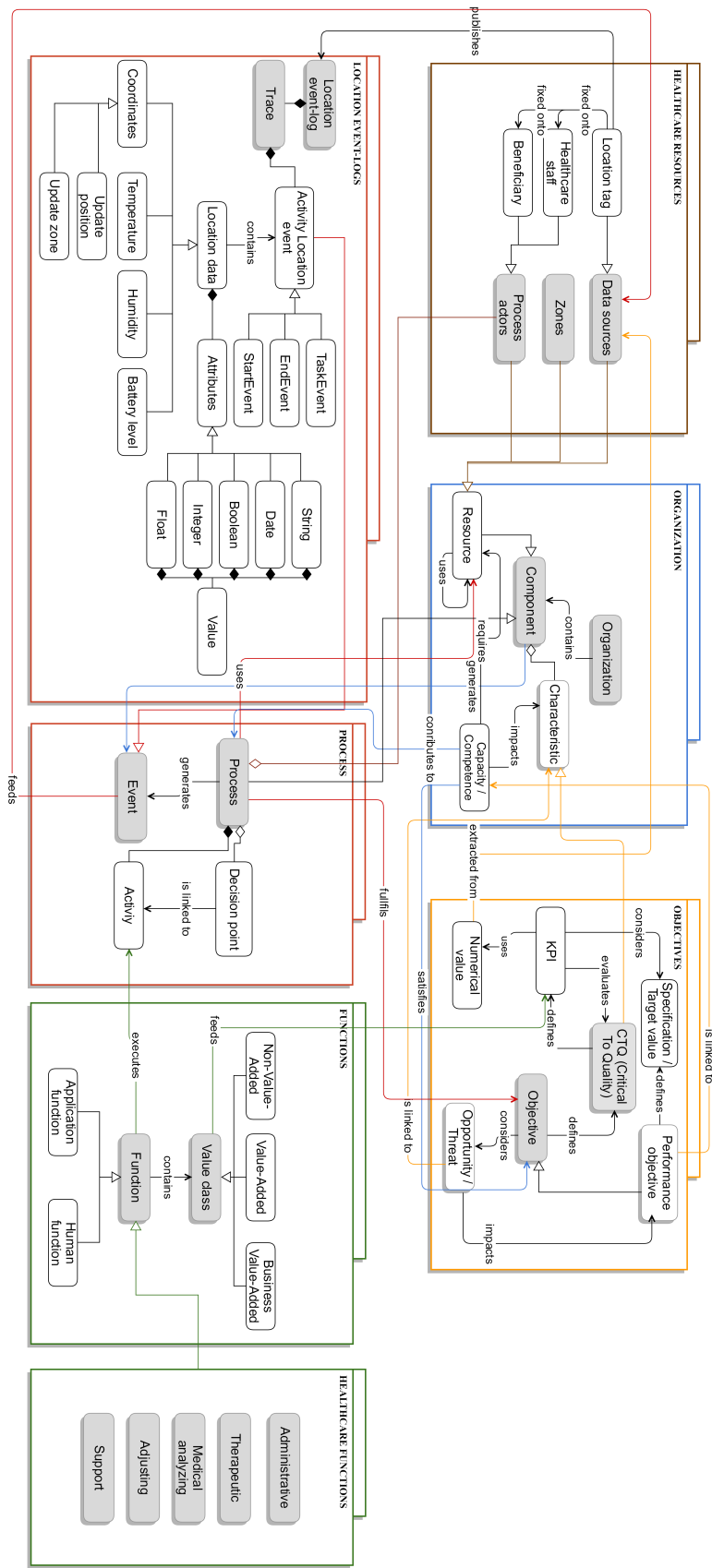


Figure 3.8 – DIAG meta-model first version without development of the support for the business process diagnostic.

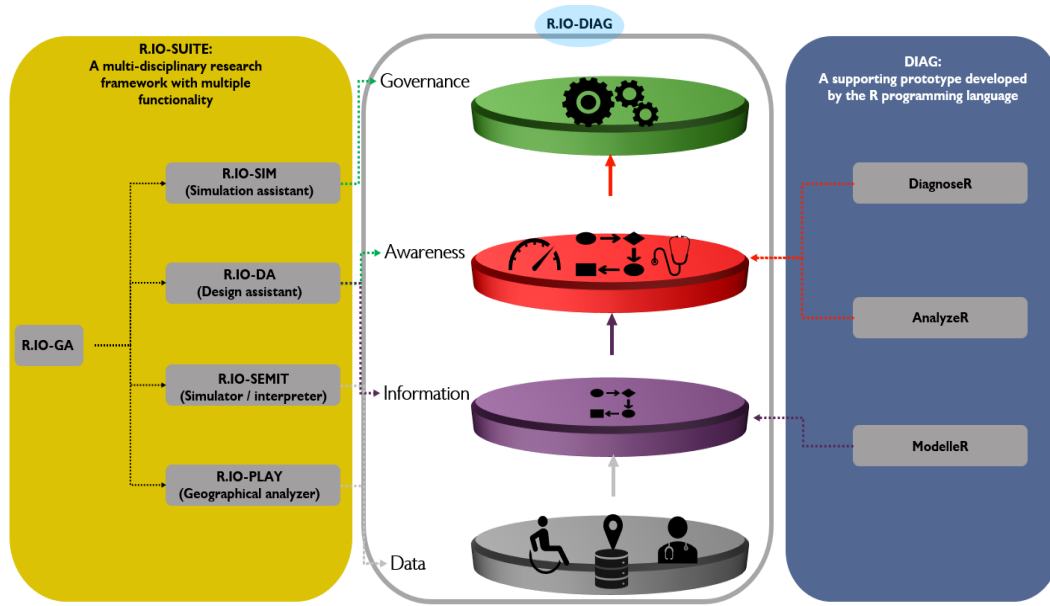


Figure 3.9 – R.IO-DIAG framework.

There are also certain elements that can improve the chances of having better results, or they can threaten the performance of the process. For instance, a bad signalization in the facility could become a *potential environmental cause* for patients to be lost.

Another example can be the lack of adequate amount of resources in the hospital which can lead to more waiting time in the process. These elements are dedicated to the *opportunity / threats* class.

As expressed here, all the mentioned concepts in this example can be adopted by the DIAG meta-model (c.f. figure 3.8). Now, it is required to realize this meta-model in terms of an application. This application must be able to endorse the **modeling, analyzing, and diagnosing** of patient pathways. This would prove the applicability of the DIAG meta-model.

To do so, **R.IO-DIAG application¹** is developed. The next section would demonstrate the primary functions of this application. Note that, this application and its functions endorse the **DIAG methodology**.

3.2.2 R.IO-DIAG initialization

A proof of concept application is developed to support the realization of DIAG methodology and the involved methods within it. This application is labeled as R.IO-DIAG.

As illustrated in figure 3.9, R.IO-DIAG is being developed within R.IO-SUITE framework. R.IO-SUITE is a mature and multi-disciplinary framework which copes with several research applications such as healthcare, supply chain, and crisis management. R.IO-SUITE framework is the result of years of research and practice by the members of Industrial Engineering Center of IMT Mines Albi (Bénaben et al., 2016), (Benaben et al., 2008).

¹<https://research-gi.mines-albi.fr/display/RIOSUITE/R-IOSuite+Home>

To support the realization of this research work in R.IO-SUITE, a proof of concept application known as **DIAG** is developed by the **R** programming language. This application verifies the authenticity of the scientific solutions prior to the implementation in R.IO-SUITE.

DIAG, supports three modules. The first one is the **ModeleR**. This module is developed for realization of two novel process discovery algorithms, stable heuristic miner(V1 and V2). Chapter 4 will investigate these algorithms.

The **AnalyzeR** module applies statistical process control techniques to measure the performance and quality of processes. First section of chapter 5 will address these methods.

The **DiagnoseR** module, is destined to apply two new methods for automatic business process diagnosing. Also, it is capable of measuring the distance between different business process models. This distance presents the differences between the structures of two process models. These methods will be further discussed in chapter 5.

R.IO-DIAG has several functionality thanks to the support by R.IO-SUITE and DIAG:

1. Location data interpretation: for the sense-making approach.
2. Business process discovery: for extracting process-like patterns from patients movements.
3. Business process analyzing: to evaluate the quality characteristics of the hospital.
4. Business process diagnosing: to detect the cause of deviations in patients' pathways.
5. Business process simulation: to propose new solutions for removing the deviation causes and improving processes.

Each functionality is being supported by a specific module within R.IO-SUITE.

- **R.IO-GA** this application orchestrates all the other applications for a better integration.
- **R.IO-PLAY** it supports the geographical representation of the data or business scenarios.
- **R.IO-DA** this application offers several benefits. It is possible to realize the link between actual concepts in an environment such as a hospital with the recorded information in the location data. Thanks to this application, one can define the organization and its resources, the process actors and other defined concepts in DIAG meta-model (c.f. figure 3.10, 3.11, and 3.12). It also, supports the extraction of process models.
- **R.IO-SEMIT** is in charge of simulating different types of data and scenarios. This application has a major role in this research work since it replays the added location data and it integrates the defined knowledge in the R.IO-DA application with the recorded information in the event logs.
- **R.IO-SIM** supports the statistical analyses of processes and also it can be in charge of business process simulations.

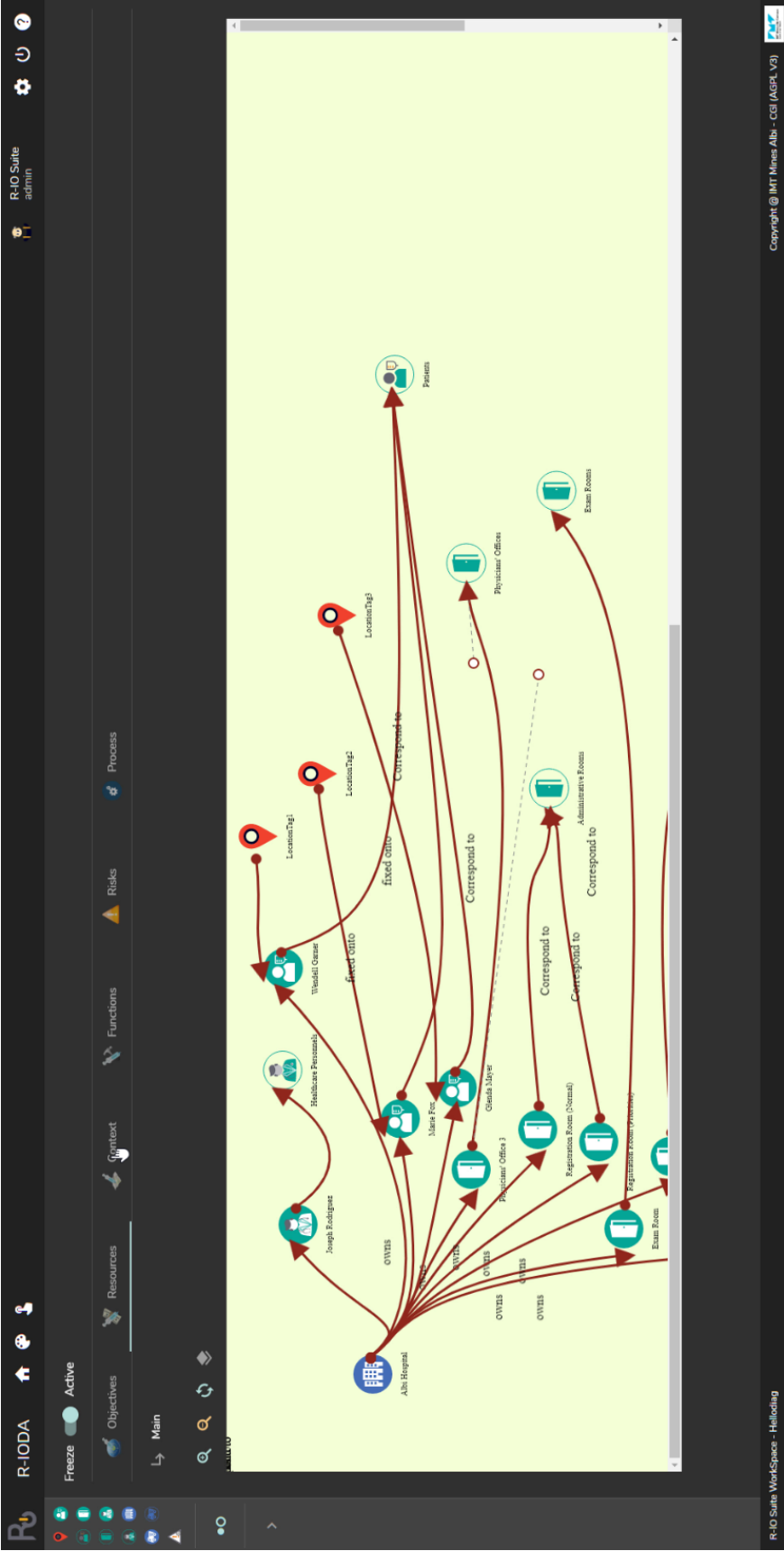


Figure 3.10 – A screen-shot of R-IO-DA which represents the modeling of the organization and its resources.

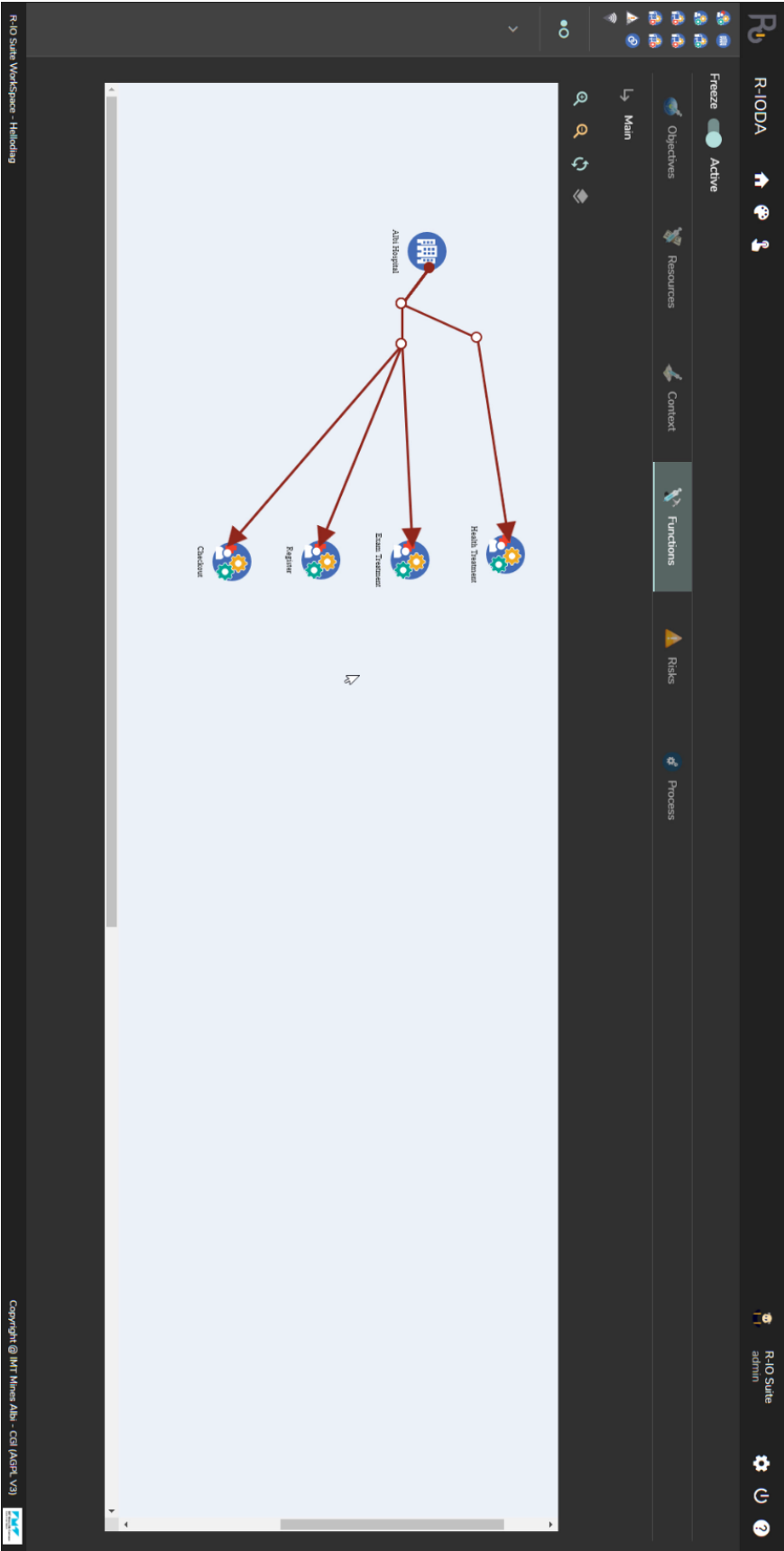


Figure 3.11 – A screen-shot of R.I/O-DA which represents the modeling of the functions.

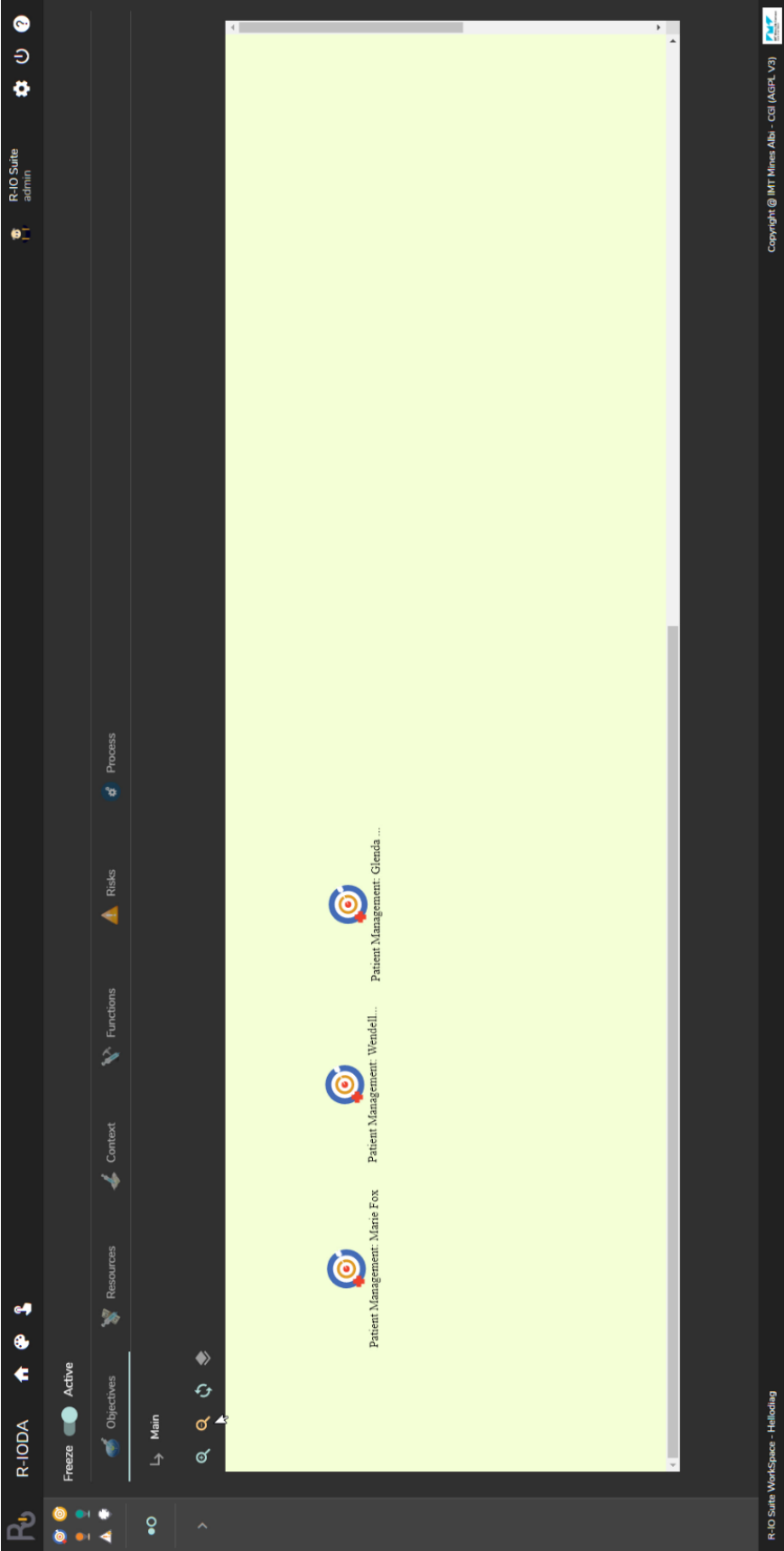


Figure 3.12 – A screen-shot of R.IO-DA which represents the modeling of the objectives which is related to each case profile.

In this chapter the main focus is on the organization modeling thanks to the R.IO-DA module.

To realize the **configuring the environment and the systems** function, at first, the organization and its resources that are involved for patient pathways will be modeled. A simple example illustrates this statement in figure 3.10.

In this example Albi hospital is considered as the organization which has three patients. Each patient is equipped by a location tag. As a result after gathering the data one can load the event log withing the location tags.

Several zones are defined for this example. This helps to identify each "zone-id" in the event log is related to which actual zone in the hospital.

Now, the functions in the organization should be modeled in order to extract what are the types of activities that can be executed in these zones. As an example in figure 3.11 four functions are defined for patient pathways. Consider "health treatment" function, by modeling this function and assigning the zones for executing this function, each time a patient enters the assigned zone, a health treatment activity will appear in the process model.

The objective of a patient's pathway can be modeled here as shown in figure 3.12. This helps to dedicate a normative scenario to the patient's pathway according to his or her profile. This profile can be related to the reason (disease) that patients are registered in the information system.

Note that, until here the location data are not received; this is just a preparation for the next functions which will be introduced in the following sections.

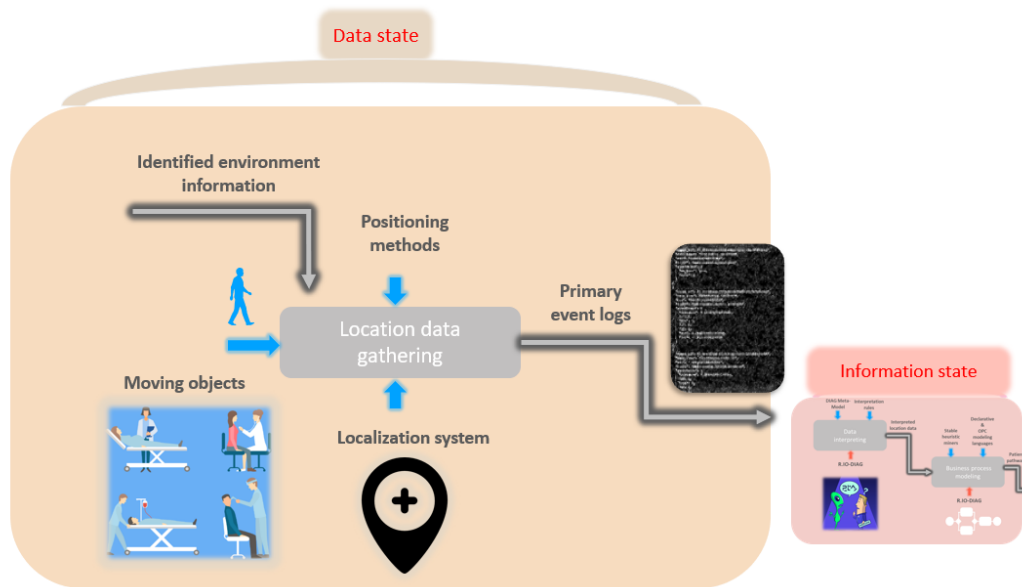


Figure 3.13 – The second function of data state, location data gathering.

3.3 Function 2: Location data gathering

As shown in figure 3.13, the location data gathering function uses the output of the previous function as a-priori knowledge. As a result, the extracted information by this function would be more comprehensible. This improvement is caused by the creation of links between the actual concepts in the studied environment and the existing information in the location event logs.

3.3.1 Objects movements

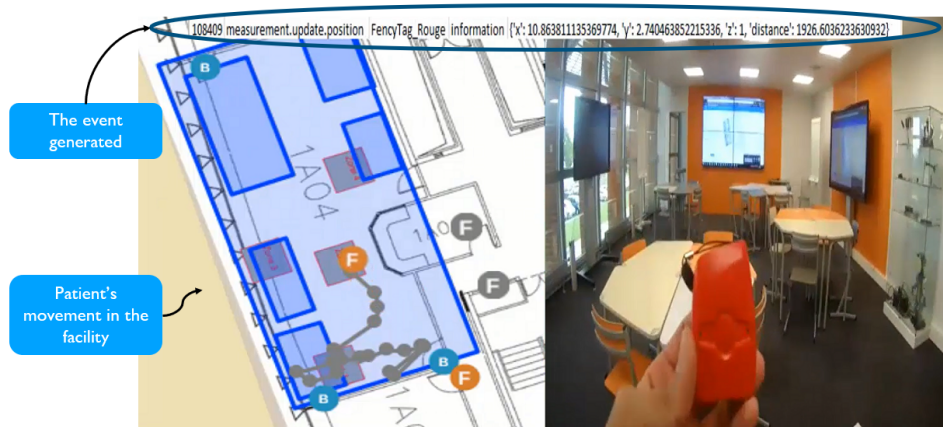


Figure 3.14 – An illustration to present how movements of objects are tracked by the localization system.

The data gathering starts by defining the virtual zones within the localization system and in parallel within R.IO-DIAG application.

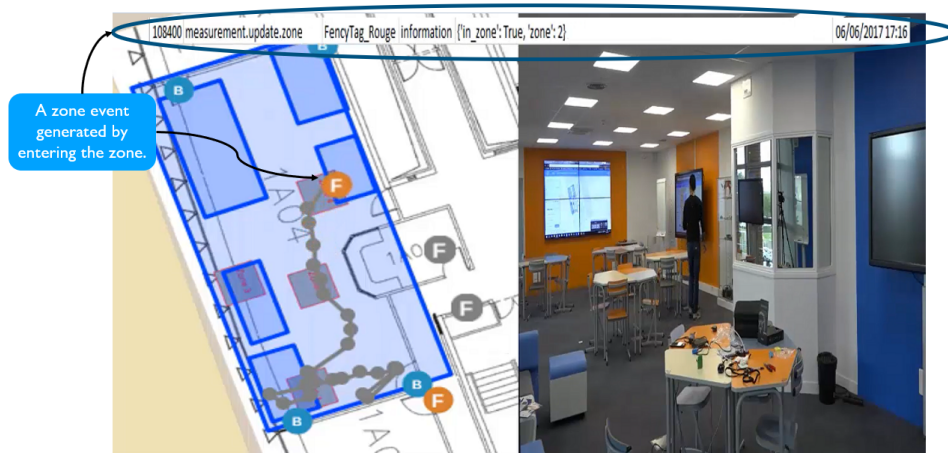


Figure 3.15 – Each entry to a zone generates a “update.zone” event.

Each virtual zone represents an area in which certain activities can take place. Therefore, inside the localization system the zone ID will be defined. For instance, zone 1 in the localization system will be the “waiting area”. Note that, without modeling these functions in the R.IO-DIAG application (c.f. figure 3.11), it will not be possible to extract the type of activity which is registered in this zone.

Next action will be equipping each process actor with a real-time location tag. As the object moves around the area, the location tag starts to emit signals and an event will be registered per second.

As presented in figure 3.5, the location data registered in events can be related to the humidity, temperature, tag battery level, and coordination of the tag. However, not all of these data are required for discovering process-like patterns.

Figure 3.14 shows an image of an experiment within a smart room. A person is equipped with a tag while moving around in an area in which four RTLS (Real-Time Location System)² antennas are installed. The tags are labeled on the map with “F” and each antenna is labeled with “B”.

As shown in figure 3.14, several virtual zones are defined for the experiment. While the person moves around these zones the events are being registered. These events are labeled as “update.position”.

If the person enters or exits a zone, an event will be recorded as “update.zone”. An example of “update.zone” events is presented in figure 3.15.

If the tag stops moving for a couple of seconds the system will start to update the information about the humidity (“update.humidity”), temperature (“update.temperature”), and the tag battery level (“update.battery.level”).

In the next two pages an example shows how location data are registered.

The first event with the *event_id* : 1, 8314 indicates that a tagged objects entered a *zone* with an *id* = 0.

From the very first information, it is clear that this type of data is not suitable for efficient process mining activities. For instance, it is not clear what represents *zone0*, *eui*, and how to interpret different attributes in each event.

²RTLS is categorized as a sub-class of indoor localization technologies.


Extraction of proper information for process mining activities such as automatic discovery of business process models is dependent on interpretation of the **coordination data**. However, now thanks to the DIAG meta-model, it is possible to link these data with actual concepts. Therefore, it is necessary to interpret this information with the added knowledge.

Basically, in the literature of process mining application on location data, it is not clear how to apply process discovery algorithms on top of these multi-level event logs. Thereby, this research work proposed a novel approach which addresses this issue. This approach starts by the development of the DIAG meta-model and the location data interpretation rules.

The primary results of this approach is presented in (Araghi et al., 2018a).

These interpretation rules create a bridge to transfer primary location event logs into *information state* to apply process discovery algorithms.

Next section (c.f. figure 3.16) will present one function of the information state and the journey to convert the non-refined location event logs into business process models.

 **An example of the primary location event logs without any knowledge about the name of the zones, executable functions within each zone, and other non-interpreted data of a process:**

```
public class ApiJSON {
    [
        {
            "event_id": 1,8314
            "topic": "measurement.update.zone",
            "attributes": {
                "in_zone": true,
                "zone": 0
            },
            "date_time": "2017/06/06,17:14:21",
            "eui": "0xd0d01e1a01000005"
        },
        {
            "event_id": 108320,
            "topic": "measurement.update.position",
            "attributes": {
                "x": 6.077144509583998,
                "y": 5.041517340297722,
                "z": 1,
                "distance": 1916.53
            },
            "date_time": "2017/06/06,17:14:22",
            "eui": "0xd0d01e1a01000005"
        },
        {
            "event_id": 124332,
            "topic": "measurement.update.position",
```

```

    "attributes": {
      "x": 6.770617485351836,
      "y": 4.570135581857626
      "z": 1,
      "distance": 1917.4282523195945
    },
    "date_time": "2017/06/06,17:14:24",
    "eui": "0xd0d01e1a01000029"
  },
  {
    "event_id": 108333
    "topic": "measurement.update.zone",
    "attributes": {
      "in_zone": true,
      "zone": 1
    },
  },
  {
    "event_id": 108337,
    "topic": "measurement.update.position",
    "attributes": {
      "x": 7.0354728558519435,
      "y": 5.05198545983953,
      "z": 1,
      "distance": 1917.97809563563
    },
    "date_time": "2017/06/06,17:14:25",
    "eui": "0xd0d01e1a01000029"
  },
  {
    "event_id": 108343,
    "topic": "measurement.update.position",
    "attributes": {
      "x": 6.959970240602145,
      "y": 5.080710963451253,
      "z": 1,
      "distance": 1918.0588780561638
    },
    "date_time": "2017/06/06,18:14:33",
    "eui": "0xd0d01e1a01000069"
  },
  {
    "event_id": 108353,
    "topic": "measurement.update.position",
    "attributes": {
      "x": 6.846516125195448,
      "y": 4.7694732447635335,
      "z": 1,
      "distance": 1918.549660156371
    },
  },

```

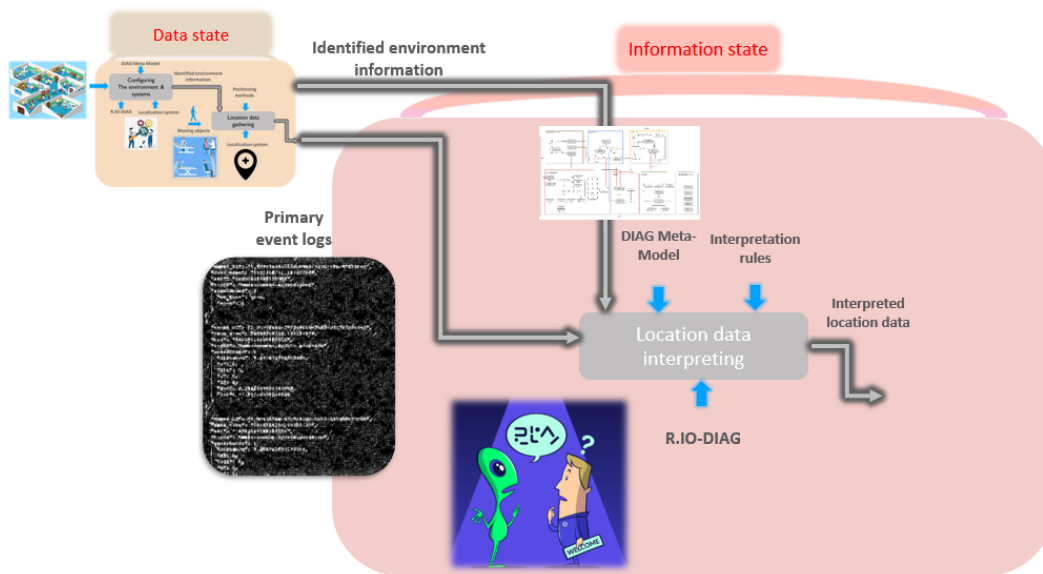


Figure 3.16 – Data interpreting function.

3.4 Function 3: Location data interpreting function

The primary location event logs will be uploaded to each defined "location tag" class in the "resource" tab of the application (c.f. figure 3.10).

As shown previously the primary location data are received as streams of events. Thus, two sets of complex event processing (CEP) rules are designed to carry on the task of location data interpretation.

The idea to extract activities from location data was to focus on the information within the events that are related to the "coordination data". These data are either related to a zone, or a changed position (c.f. figure 3.5)

The motive for defining two different sets was due to the fact that each set is in charge of different actions.

Figure 3.17 shows how these two sets of CEP rules are working to detect-and-add more **information** to the extracted location data. They mainly focus on the "update.zone" events.

The enrichment and adjustment of the primary data by the **domain knowledge** is the reason for positioning this function within the **information state**.

1. **CEP rules-group 1:** this set is constructed by three rules. They are in charge of transforming the very first visualization of data as "nodes" with out any semantics. These graphs will be sent to a *graph database management system*. For example, outputs of this step will be series of nodes labeled as "An activity in the Zone #".
2. **CEP rules-group 2:** after detecting the nodes, second group of CEP rules will start retracting the registered knowledge within the application. As mentioned this action is possible by the modeled information within the "resource" tab of the application (c.f. figure 3.10).

3.4.1 Add start-event

This algorithm (c.f algorithm 1) is used in the first group of CEP rules for extracting the "start-event" of a process from the location event logs. It receives the primary event log and considers the events relevant to the zones (*update.zone*) within the hospital facility. These events have properties such as '**id**', '**topic**', '**date_time**', '**eui**'. "eui" is the localization id which is related to a tag for a certain process actor.

The other attribute of *update.zone* is the '*in-zone*' attribute which could have two values; '**FALSE**' (out of the zone) or '**TRUE**' (in the zone).

Therefore, main functionality of the first group of CEP rules is dependent on this attribute in order to discover the activities.

After finding the in-zone attribute, it will look for the start event corresponding to the zone with an id equal to '0'. The zone with the id of '0' is representing the *general process zone*. For example a hospital could be the zone '0'. As soon as the process actor is in the general process zone (*zone.0*) the process starts.

Therefore, algorithm selects this '*zone_event*' as the **start-event** if the '*in_zone*' attribute equals to 'TRUE'. Finally, it records all the corresponding attributes of the event such as the time-stamp start of the process, and value class of the activity.

As mentioned in chapter 2, the OPC modeling language is used here. Thereby, a specific start-event representation will be used (OPC modeling language will be discussed in detail within the last section of this chapter).

Algorithm 1 Discovering the start activity

```

1: procedure AddStartEvent
2:   Input < EventLog >= Structure
3:     < Name >: string
4:     < Id >: integer
5:     < Tags >: integer
6:
7:   Filter(EventLog, Topic == "measurement.update.zone")
8:
9:   < zone_event >= Structure;
10:    < Topic >= string
11:                                     ▷ measurement.update.zone = Topic
12:    < EventID >= integer
13:                                     ▷ j=1,2,3,...,m
14:    < eui >= integer
15:    < DateTime >= string
16:    < InZone >= boolean
17:    < Zone.ID >= integer;                                     ▷ i= 0,1,2,3,...,n
18:   for zone_event do                                     ▷ Receiving stream of location events
19:     if (Topic == 'measurement.update.zone', AND Zone.ID == 0,
20:     AND _in_zone == 'TRUE') then;
21:       Select zone_event ;                                     ▷ Selecting the event as the start activity ;
22:                                     ▷ StartEvent representation;
23:       Define zone_event as 'StartActivity';
24:       Register StartActivity_properties ;
25:     end if
26:   end for
27: end procedure

```

3.4.2 Add end-event

This algorithm receives the primary location event log as an input like the previous algorithm. Each *zone_event* has several properties that algorithm will use them to identify the last event of a process. As it has been shown in "algorithm 2", the algorithm will search for all the events that have the topic equal to "*measurement.update.zone*".

This indicate that we are investigating the events related to the zones (not update position, temperature, or battery). simultaneously, it identifies the event related to the zone '0' which is the general zone of process. Later on, if the *in_zone* attribute equals to "**FALSE**" it would consider the event as an "exit" from the general zone of the process. Therefore, it will register all the event's properties as the **end-event** of the process.

Algorithm 2 Discovering the end activity

```

1: procedure AddEndEvent
2:   Input < EventLog >= Structure
3:     < Name >: string
4:     < Id >: integer
5:     < Tags >: integer
6:
7:   Filter(EventLog, Topic == "measurement.update.zone")
8:
9:   < zone_event >= Structure;
10:     < Topic >= string
11:     < EventID >= integer
12:                                     ▷ j=1,2,3,...,m
13:     < eui >= integer
14:     < DateTime >= string
15:     < InZone >= boolean
16:     < Zone.ID >= integer;                                     ▷ i= 0,1,2,3,...,n
17:   for zone_event do                                     ▷ Receiving stream of location events
18:     if (Topic == 'measurement.update.zone', AND Zone.ID == 0,
19:     AND in_zone == 'FALSE') then;
20:       Select zone_event ;                                     ▷ Selecting the event as the start activity ;
21:                                     ▷ EndEvent representation;
22:       Define zone_event as 'EndActivity';
23:       Register EndActivity_properties ;
24:     end if
25:   end for
26: end procedure

```

3.4.3 Add task-event

Algorithm 3 add-task-event

```

1: procedure AddTaskEvent ▷ Discovering the Human Task
2:   Input < EventLog >= Structure
3:     < Name >: string
4:     < Id >: integer
5:     < Tags >: integer
6:
7:   Filter(EventLog, Topic == "measurement.update.zone")
8:
9:   < zone_event >= Structure;
10:    < Topic >= string
11:    < EventID >= integer
12:
13:    < eui >= integer ▷ j=1,2,3,...,m
14:    < DateTime >= string
15:    < InZone >= boolean
16:    < Zone.ID >= integer; ▷ i= 0,1,2,3,...,n
17:   for zone_event do
18:     if (z1 = zone_event [z1.topic== 'measurement.update.zone', AND z1.id > 0 ]
        AND (z2 = zone_event[z2.topic == 'measurement.update.zone' AND z2.id > 0 AND
        z1.id==z2.id AND z1.eui == z2.eui]))
19:     then
20:       Select "z1.event_id" as 'HumanTask' ;
21:       Define 'name' ("z1.event_id" as "Activity.Inside.Zone");
22:       Register 'z1.event.propoerties' as 'HumanTask.propoerties';
23:     end if
24:     if ((z1.id == z2.id) AND( z1.eui==z2.eui )AND
        (z1.zone.attribute_in_zone=='TRUE') AND (z2.zone.attribute_in_zone=='FALSE')
        AND (z1.date_time, z2.date_time ))
25:
26:       OR ▷ In case of existing noises
27:       (z2.zone.attribute_in_zone == 'TRUE') AND z1.zone.attribute_in_zone=='FALSE')
        AND (z2.date_time, z1.date_time)))
28:     then
29:       Register zj.date_time;
30:     end if
31:   end for
32: end procedure

```

The presented algorithm in this section (c.f. algorithm 3) describes how to discover the activities from the location event logs. Assuming human functions are executing the process activities within defined zones of the facility, one can infer that an activity (a task) starts when a process actor enters a certain zone and it finishes when the process actor leaves that zone. Algorithm 3 expresses in more detail the rule for detecting such activities.

At first glance, algorithm 3 works like the previous rules (algorithms 1 and 2), the algorithm is looking for the events that are related to the zones (*measurement.update.zone*).

However, here the algorithm is seeking zones that are within the general process zones or zones with the id greater than '0' (If the $z_j.id > 0$).

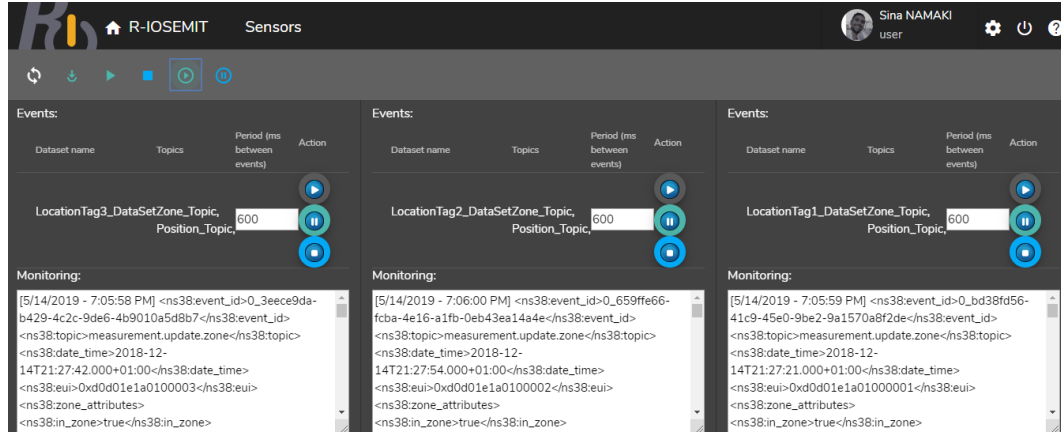


Figure 3.18 – A screen shot of the application loading the event logs.

After identifying the zone in z_1_event we look for the next event that share the same zone id ($z_j.id$), and additionally, we ensure that the second event also is related to the same patient (for example: $z_3.eui == z_6.eui$). This helps to ensure that the activity has been finished and the algorithm is not considering an **incomplete data** in the event log. This is an important characteristic to provide a refined set of data sets for avoiding the *incompleteness* notion in event logs.

After extracting the starting point of an activity, the algorithm randomly defines a node and name it as "**Activity.Inside.Zone #**" and all the defined properties of the node such as time will be recorded.

Note that a process actor may go severally into a zone and goes out during his/her process. In order to avoid misinterpreting the activity sequences, the algorithm uses the *date_time* concept (at line 14 of the algorithm 3).

At the final step, duration of the activity is registered thanks to the extracted start and end-event.

Consequently, by using the first set of CEP rules, three types of nodes will be extracted; **start**, **end**, and **task events** (c.f. figure 3.19).

Later on, these nodes will be retracted by the "data interpreter" in order to add more knowledge on them. The following set of algorithms help to interpret which activity is being executed in a certain zone.

Note that this a-priori knowledge is provided thanks to the implementation of the DIAG meta-model (c.f. figure 3.11 and 3.8).

Figure 3.18 presents how the event logs are being charged within the application.

Figure 3.19 shows the extracted nodes before adjusting them with the defined knowledge.

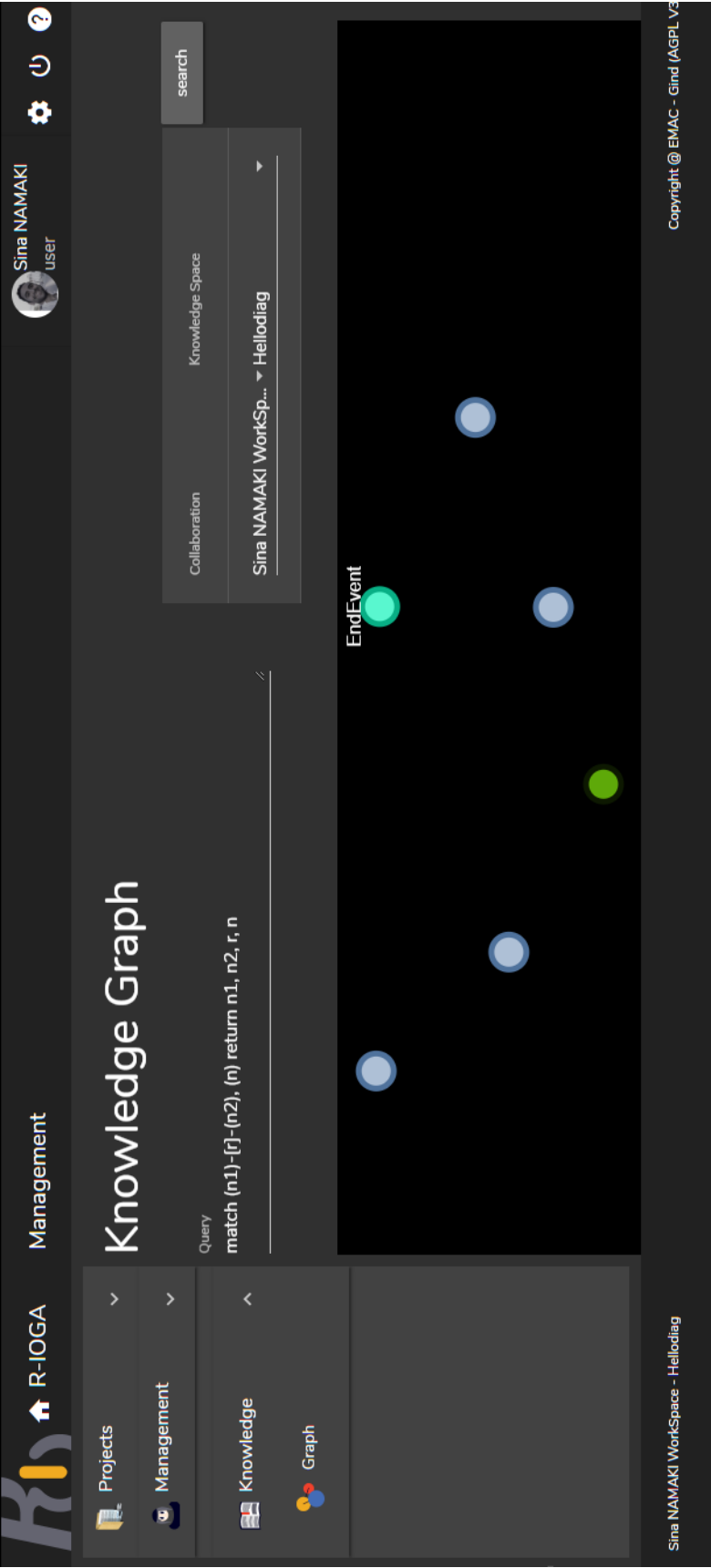


Figure 3.19 – A screen shot of the application extracting primary nodes without any added knowledge.

3.4.4 Add knowledge on start-event

This algorithm is applied to understand, how each graph is related to a certain process and to adjust the start event representation to the grammar of OPC modeling language.

To do so, at first, the '*StartEvent*' is received as an output of algorithm 1(add-start-event). Then, the corresponding *eui* of the tag will be detected. Note that each tag is assigned to a particular process actor; thus, it helps to identify the process actor.

By identifying the process actor, the corresponding graph will be dedicate to the process actor with the relevant *eui*.

Next in line 7, the node adopts the OPC grammar. Finally, the node attributes are registered.

Algorithm 4 add-knowledge-on-StartEvent

```
1: procedure Add-Knowledge-On-StartEvent
2:   Input 'StartEvent';
3:   for all StartEvent do
4:     if 'StartEvent.eui' == 'beneficiary.eui';
5:   then
6:     Select 'beneficiary';
7:     'StartEvent.node.type' == 'StartEvent.OPC';
8:     'Process.name' == "beneficiary.name";
9:     Register 'node.property'
10:    end if
11:  end for
12: end procedure
```

3.4.5 Add knowledge on end-event

Similarly to the previous algorithm 4, this algorithm receives the result of algorithm 2 then tries to match the corresponding process actor with the *eui* attribute.

Next, it selects the process actor (**beneficiary**) and goes for transformation of the basic graph by OPC representations. By registering attributes of such events, the algorithm is able to calculate the duration and the taken distance of the processes.

Algorithm 5 add-knowledge-on-EndEvent

```
1: procedure Add-Knowledge-On-EndEvent
2:   Input 'EndEvent';
3:   for all EndEvent do
4:     if 'EndEvent.eui' == 'beneficiary.eui';
5:   then
6:     Select 'beneficiary';
7:   end if
8:
9:     'EndEvent.node.type' == 'EndEvent.OPC';
10:    Register 'node.property'
11:  end for
12: end procedure
```

3.4.6 Add knowledge on task-event

Last but not least, the *Add knowledge on task-event* algorithm (6) will help to extract the activity types.

Recall that in DIAG meta-model several human function classes are defined. These classes are defined as **Therapeutic**, **Administrative**, **Support**, **Medical analyzing**, and **Adjusting functions**. Each of these classes has one of the mentioned value classes (VA, BVA, NVA). Algorithm 6 exhibits how the system extracts the types of information related to the human functions.

Algorithm 6 add-knowledge-on-Task

```

1: procedure Add-Knowledge-On-Task
2:   Input 'Activity.Inside.Zone'
3:   for Activity.Inside.Zone do
4:     if 'z.id' == 'j' ;
5:   then
6:     return 'zj.properties(zj.category,zj.eui)';
7:
8:     if 'zj.category' == 'c' then
9:       return 'c.HumanFunctionCalss';
10:      Register 'node.properties'
11:      'node.property.name' == 'HumanFunction.value'
12:    end if
13:  end if
14: end for
15: end procedure

```

At first, It receives all of the extracted tasks by the algorithm 3. Next, it uses the "id" of an activity zone, and it returns the corresponding information.

For example, if an activity is executed in *zone1(z.1)*; the algorithm will find three elements corresponding to *z.1*. It finds the *eui* of the process actor and the category of the zone (*z_j.category*). Then, it returns the corresponding human function which has been defined for that certain zone.

Then, it replaces the name of the activity with the proper function and registers all the necessary information such as the duration of the activity and etc.

Finally, it replaces the basic representation of the graph by an OPC modeling language element.

The result of these rules will be a set of individual processes which are extracted from event logs. Figure 3.20 shows an example of the explained procedures. Note that a merged representation of the ensemble of all processes will be addressed in the next chapter, where the automatic process discovery algorithms are represented.

The OPC modeling language will be further discussed in below.

3.4.7 Operation Process Charts (OPC)

The birth of **operation process chart (OPC)** is aligned with world war 2 period, when the industrial solutions of U.S army aided the allied troops for improving logistics of goods.

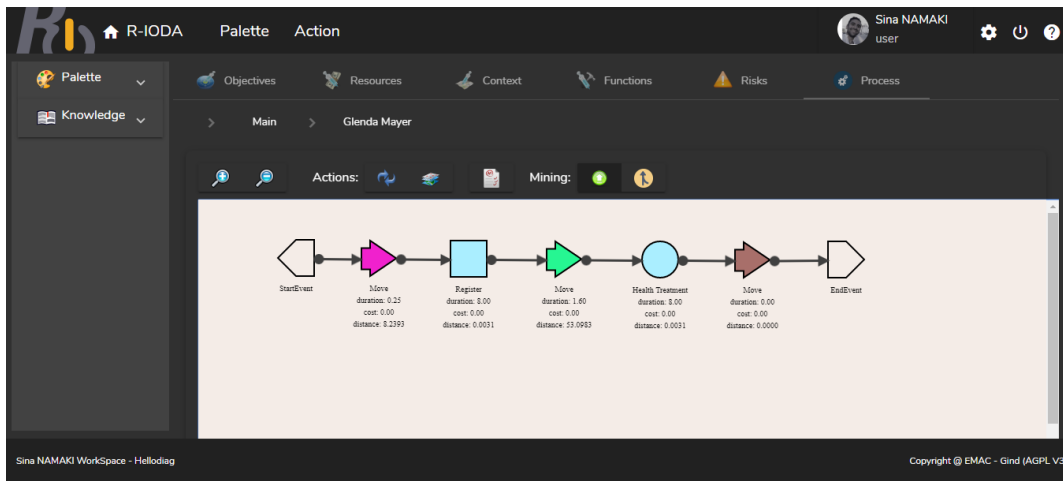


Figure 3.20 – A process model representing the activities of an individual case by OPC modeling language.

Modeling operations and tasks by OPC permits to visualize different types of activities within a process model. This solution became more popular during late 80's when Deming proposed it within the DMAIC (define, measure, analyze, improve, control) approach (Montgomery, 2019). Until now, this solution was mainly proposed for the manufacturing industries and it allows for distinguishing **non-value-added (NVA)**, **business-value-added (BVA)**, and **value-added (VA)** activities (Montgomery, 2019). To best of our knowledge, previous to this research work, the application of this solution for automatic business process discovery, and additionally healthcare process modeling has not been addressed yet (Araghi et al., 2018a).

Figure 3.21 presents the OPC modeling language elements and its adaptation for representing healthcare activities within medical and non-medical processes.

Figure 3.22 shows an example of applying OPC modeling language for 3 cases in R.IO-DIAG application. Thanks to the previously mentioned location data interpretation rules and the support of R.IO-DA plugin, the application is capable to model patients' pathways by distinguishing different types of each activity.

Comparatively to the existing **declarative** or **procedural modeling languages** in the literature, this particular aspect of OPC is unique and useful for analyzing the **process cycle efficiency**.

Each modeling language method can provide advantages and disadvantages. For instance, mathematicians used **formal modeling languages** to address problems by methods like Markov chains, queuing networks. Computer scientists use Petri-nets, transition systems to model processes. However, these languages have *unambiguous semantics for further analysis*.

On the other hand, the **conceptual methods** such as BPMN, OPC, and EPC are used in practice since users have difficulties using **formal languages**.

It should be noted that, such languages like OPC are useful solutions for **the execution of patient—pathways—processes** (an example of process executions: simulating business processes).

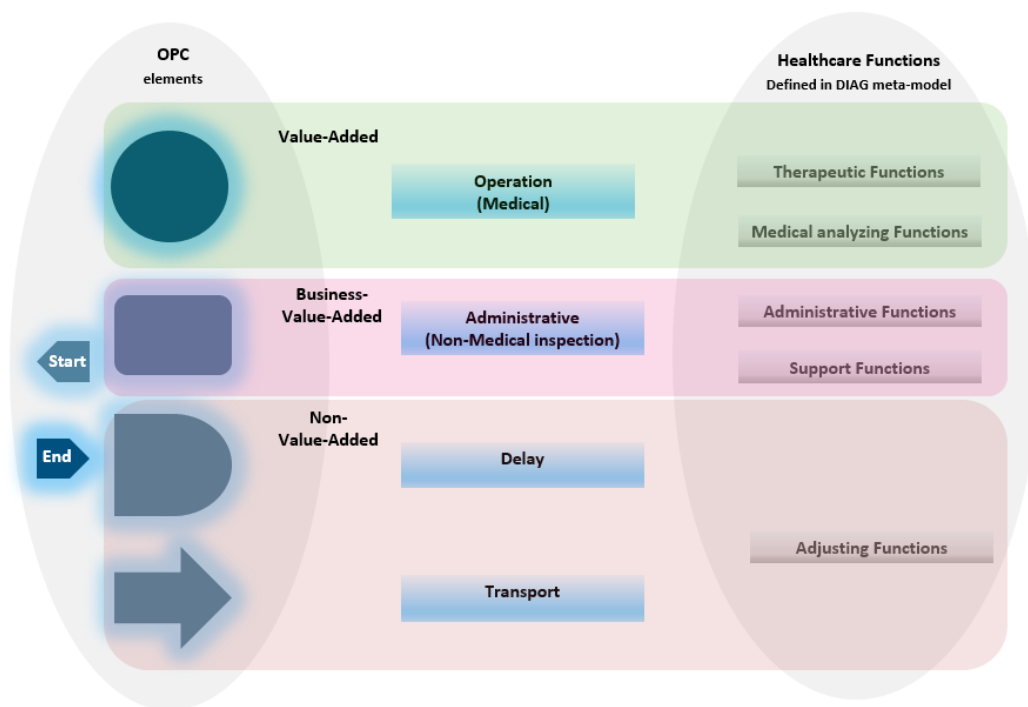


Figure 3.21 – OPC modeling language elements and their relations with the health-care functions package and value-class concept in DIAG meta-model.

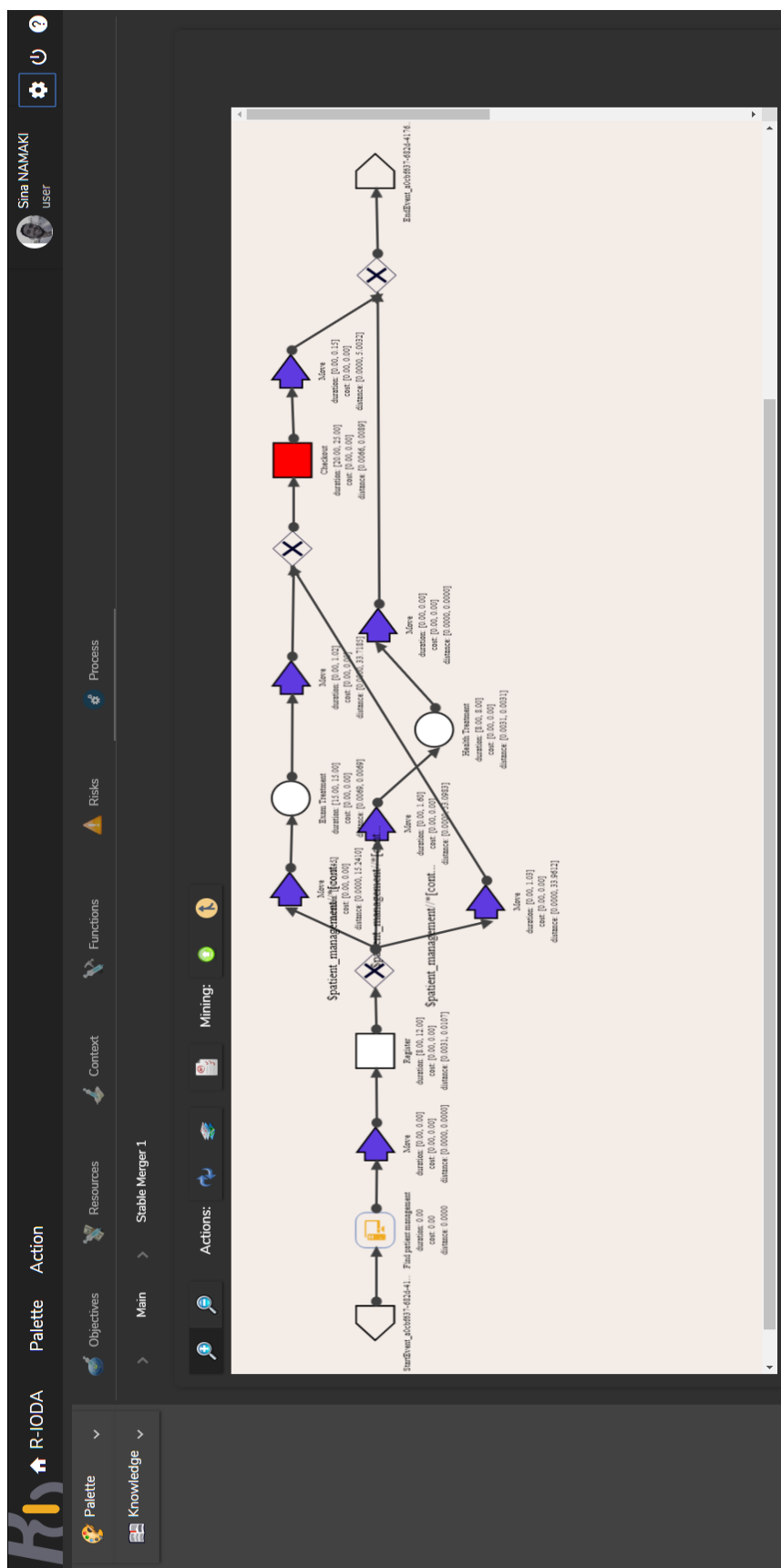


Figure 3.22 – An example of patient pathways represented in R.IO-DIAG application by OPC modeling language.

3.5 Recap

As expressed by Benaben et al. (2019), organizations need to take three actions in order to acquire a smart data-driven decision making approach.

First, they should *gather* the data. Second, they should *interpret* these data. Finally, they are able to *exploit* the meaningful information for making the optimum decisions (Benaben et al., 2019).

This issue has been addressed here by the DIAG methodology and is embodied by the proof of concept application, R.IO-DIAG (c.f. figure 3.9). Accordingly, to link the

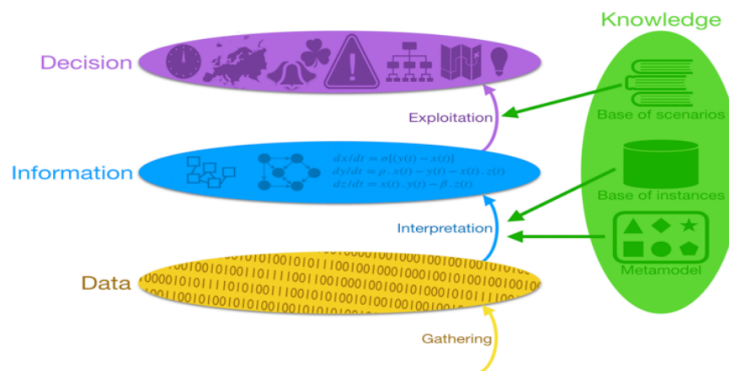


Figure 3.23 – An exemplary framework to present the relationship among data, information, and knowledge, (Benaben et al., 2019).

location data with the data-driven decisions one needs to first gather and interpret these data. To interpret certain raw data, it is necessary to add **a-priori knowledge** about the environment which is being examined. Therefore, the motive for this chapter was: “**How to provide a-priori knowledge for gathering and interpreting the location event logs**”.

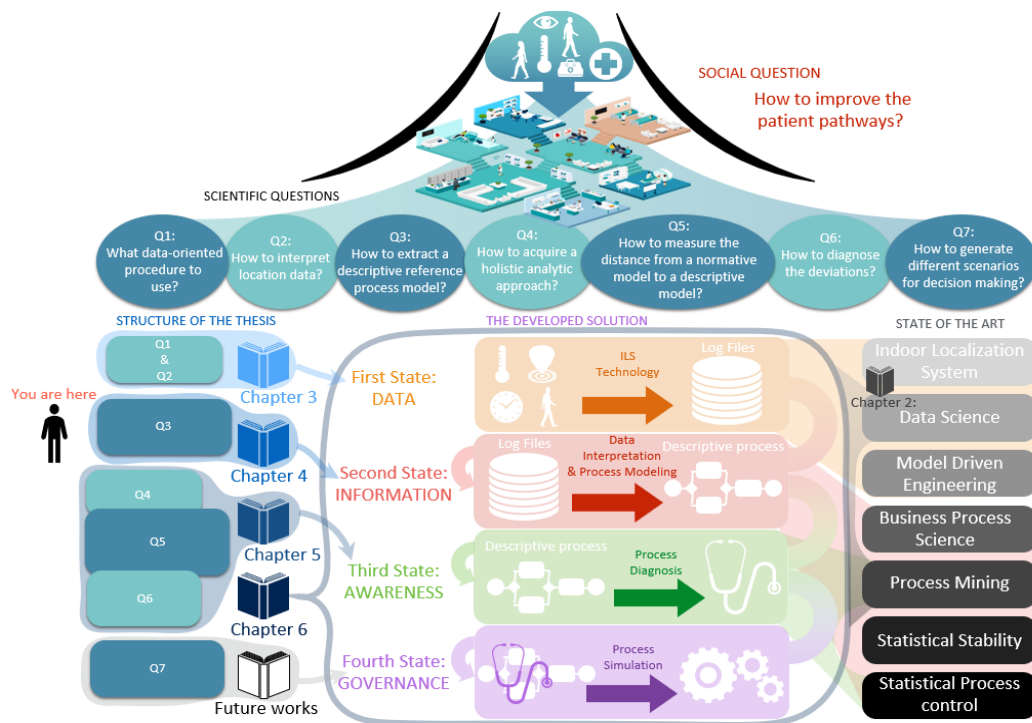
As presented in the second chapter (state of the art), the interpretation of location data is almost neglected. Most research works went straightly from data to the extraction of knowledge, without mentioning the interpretation procedures.

Thus, this research work suggested a model-driven engineering approach as the closing-gap solution. As a consequence of using this approach, the **DIAG meta-model** has been presented.

The DIAG meta-model links the recorded and blurry data in location event logs with the actual elements of business processes in an organization. This meta-model endorses the task of gathering and interpreting location data. However, it is not sufficient individually. It is necessary to accompany this framework with certain interpretation rules.

Consequently, at the fourth section, the “**location data interpretation rules**” are presented. These rules take us from the **data state** into the **information state**. Within the data state, the primary data were not sufficiently expressive. It is indeed within the **information state** that one can infer what is the AS-IS situation.

Next chapter will investigate in more depth the information state.



4

The Business process modeling function within the Information state

4.1	Introduction	87
4.2	Background	88
4.2.1	Automatic business process discovery-Existing approaches	88
4.2.1.1	Characteristics of process discovery algorithms	89
4.2.2	The classic heuristic miner	91
4.2.3	Statistical stability	92
4.3	Stable Heuristic Miner V1	94
4.3.1	An imaginary hypothesis	94
4.3.2	Preliminaries	94
4.3.2.1	Assumptions	95
4.3.2.2	Definitions and the sequence of functions in the algorithm	96
4.4	Stable Heuristic Miner V2	105
4.4.1	Preliminaries	105
4.4.2	Order of calculation	108
4.4.3	Comparing the results of the algorithms	114
4.5	Recap	118

"I love it when people discount me... Cowards and Champions have the same fears. The difference is how they attack them."

Mat Fraser



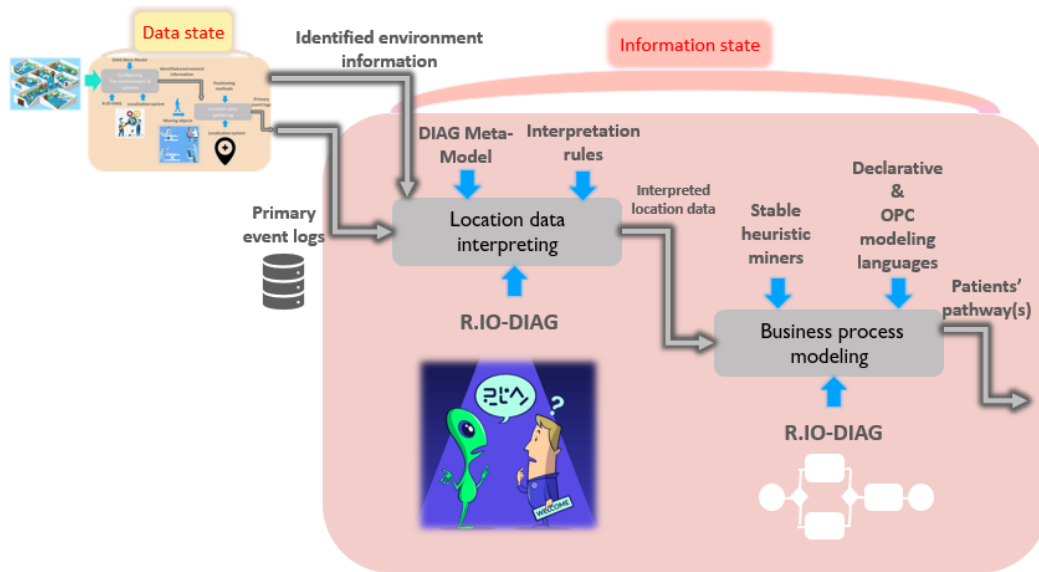


Figure 4.1 – A schema to present the Information state and the integrated activities within.

4.1 Introduction

According to DIAG methodology, after interpreting primary event-logs, it is a necessary action to extract a set of information regarding the AS-IS situation which is hidden in the location event logs.

To carry on with this objective, the **Business process modeling function** is defined within the **Information state** (c.f. figure 4.1). This function will be fully presented in this chapter.

As has been highlighted on the overall image of the dissertation, this chapter focuses on a scientific question which is:

- How to extract a **descriptive reference process model** from event logs?

Such a model expresses the **common recorded behavior** in an event log. This **common behavior**, as mentioned in chapter 1, relates to the activities and edges that are **normally** present for each patient pathway.

In line with this question, two process discovery algorithms are designed within the **Information state** of **DIAG methodology**. In essence, this chapter focuses on highlighting these two algorithms:

- **Stable Heuristic Miner V1**
- **Stable Heuristic Miner V2**

The following section presents why existing methods are not sufficiently relevant for extracting the descriptive reference process model. The third section explains in detail the scientific approach of the **stable heuristic miner V1**. The fourth section covers the second version of this algorithm, **stable heuristic miner V2**. Finally, the last section concludes this chapter and the developed work within the Information state.

4.2 Background

4.2.1 Automatic business process discovery-Existing approaches

As well as using location event logs, some researchers in this area aim to extract a **so-called descriptive reference model** which represents the **major behavior pattern** recorded in event logs.

Existing discovery methods in the literature for the extraction of descriptive reference models are: **model-based** and **clustering-based** approaches, which mainly address the *deviation detection* from the behavior recorded in an event log (Li, G. et al., 2017). The authors also introduced a new approach for discovering the reference model which could be regarded as a **profiling-based approach**.

Model-based approaches simply aim to discover a model as the descriptive reference model which represents the major pattern of behavior recorded in an event log. They then filter the cases that do not fit the model and classify them as deviations.

Examples of these approaches are the two works of Bezerra et al. (2013), where they propose the iterative and sampling algorithms for finding frequent cases and detecting variations. They identified the "dynamic threshold algorithm" for anomaly detection of traces in event logs.

In (Jalali et al., 2010), the authors suggest a method consisting of five steps to select the most compatible model. Their method has five steps: *scoping*, *process discovery*, *filtering of fitting models*, *model selection*, and *splitting of the log*. Notably, the authors use genetic algorithms to discover an appropriate behavior from a refined event log, then the algorithm will classify the cases that are not seen as deviations.

In summary, these model-based approaches extract a fitting-model as a reference model, and later on, by applying conformance checking techniques they try to categorize cases that do not fit the model as deviations (Li, G. et al., 2017).

Clustering-based approaches are seen as more suitable than model-based ones for use in unstructured environments, such as hospitals. However, these methods aim to detect clusters of behaviors rather than detecting the deviations that are causing the instability in the processes. Consequently, they can be time-consuming (Li, G. et al., 2017).

These approaches have been applied in several research works, such as (Ghionna et al., 2008), where the authors tried to associate a discovery algorithm with a cluster-based anomaly-detection procedure to deal with categorical data. This approach required a threshold to be set *manually* in order to filter the infrequent cases. As a result, determining the level of information that should be presented in the model is highly dependent on the experience of the user.

Rebuge et al. (2012), also applied a sequence-clustering method as a pre-processing step to handle the variability within large event logs. Their approach represents a set of robust methods for distinguishing regular behaviors. They also evaluated their approach by a case study in an emergency department.

The model-based and clustering-based methods have been challenged by Li, G. et al. (2017). The authors have indicated that these methods are either slow or inaccurate when dealing with complex event logs and unstructured processes that may contain *many activities*. Therefore, they proposed a novel **profiling-based** approach which

creates a "profile" of cases that are representative of the majority of normal behaviors in the event log.

Their approach has several steps: first, they sample all the cases in the event log. Then, based on a defined norm function they gather normal cases and identify them as the mainstream cases. Once they have found the mainstream, they compute the similarities between the mainstream and other cases.

By creating the concept of a "profile" they classify cases with mutual features. Then, their method quantifies the similarities based on the profile and identifies the normal cases and deviating cases. By **adjusting manually** the "norm function", one could increase (or decrease) the probability for sampling of the normal cases.

By performing an experiment, they proved that their approach is faster than previous cluster-based methods, and more accurate than model-based methods.

In spite of this, the adjustment of the norm function requires an expert who has complete knowledge of the process. This could be viewed as a *disadvantage*, since the algorithm itself should be capable of showing a stable and comprehensible amount of information from event logs.

On the other hand, the application of process mining on location event logs in healthcare is emerging rapidly. For instance, in (Kamel Boulos et al., 2012) the authors provide an overview of the application of real-time location data in healthcare. The authors highlight the advantages of using location data for monitoring patients' movements and increasing the safety of the patients and staff.

Fernandez-Llatas et al. (2015) covered a methodology with seven steps to extract a declarative model of processes. Their approach is defined in the form of a tool which directly transforms the location data into the model-based analysis.

The first step is the set-up of the localization system. The second step is data gathering. In the third step, they defined a phase as the semantic grouping of areas identifying the zones and rooms in the hospital. Fourthly, the process filtering focuses on the process mining activities. The process owner, who designs the virtual zones, will select the samples of the events that will be used to perform a process discovery in the later steps. In the process discovery phase, the tool uses different filtering options and algorithms for process discovery. The sixth step, will cover the conformance checking and process enhancement activities.

These works provided robust contributions regarding the application of indoor localization and process discovery activity in healthcare. However, they do not cover the scientific question mentioned in the above; **how to find a model that expresses the descriptive reference process**.

Therefore, this chapter presents the necessary characteristics that a process discovery algorithm should have in order to come close to discovering a so-called descriptive reference process model.

4.2.1.1 Characteristics of process discovery algorithms

Every process discovery algorithm has four main properties: **representational bias**, **ability to deal with noise**, **completeness notion**, and the **approach used** (W. M. P. v. d. Aalst, 2016).

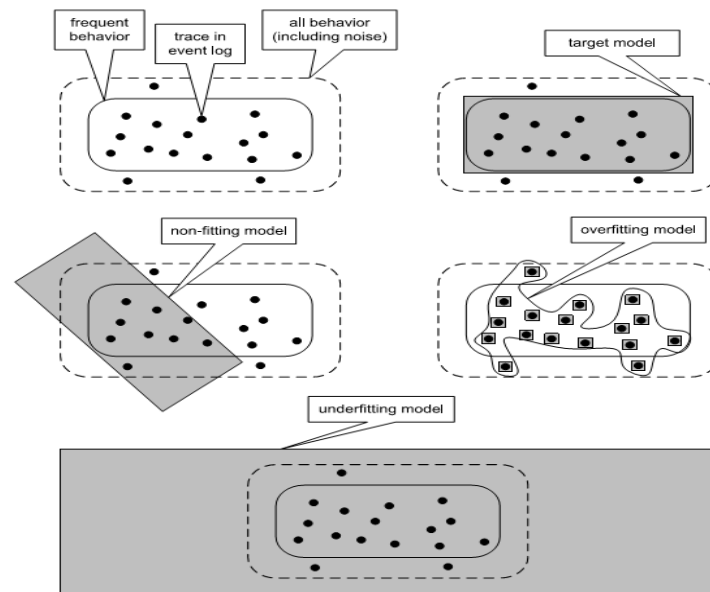


Figure 4.2 – A view on how a process discovery algorithm should address the noise problems cited in (W. M. P. v. d. Aalst, 2016).

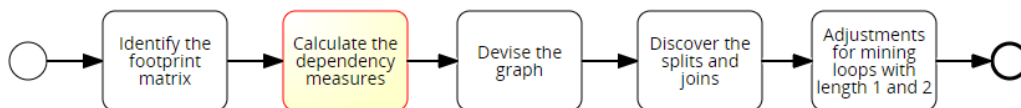


Figure 4.3 – Basic steps in heuristic mining algorithms (W. M. P. v. d. Aalst, 2016).

One of the most important features is the ability to deal with noises, and the lion's share of research works regarding process discovery do indeed address this issue. In (W. M. P. v. d. Aalst, 2016), the author expresses how this challenge should be perceived.

As shown in figure 4.2, Van der Aalst painted a picture of how one should understand the difficulties for dealing with noises (W. M. P. v. d. Aalst, 2016).

Accordingly, the author defines a "target or an ideal model" as a process model which allows for the **frequent behaviors** seen in an event log. In the same publication, the author mentions the uncertainty involved in defining normal behavior. He states that normal behavior can be defined simply as a model which allows for 80 % of the existing behaviors in the event log. However, it seems somewhat arbitrary to decide what percentage of recorded behaviors should be expressed in a mined process model.

As shown in figure 4.2, a "**non-fitting**" model will not be suitable, as it is not able to capture enough volume of information to express the main behavior patterns. Also, any process discovery technique should avoid **over-fitting** too. This means that an ideal model should not capture all of the behaviors in the event log.

Additionally, the **under-fitting models** allow for the generation of behaviors that do not exist in the event log. Given these constraints, **it is not clear how to find a trade-off among these notions.**

Under these circumstances, the author indicates that there is **no exact definition of the target model** and it is not clear how to respect all the criteria mentioned.

The methods presented here address this issue, and they are inspired by the well-known family of heuristic mining algorithms (W. M. P. v. d. Aalst, 2016), (A. J. M. M. Weijters et al., 2011), (T. Weijters et al., 2004), (De Cnudde et al., 2014).

The motive behind considering these algorithms is their ability to deal with health-care processes (Rojas et al., 2016). However, these algorithms are faced with many **structural problems**, such as **dealing with variations**, defining ways to **filter noisy behaviors**, and **random selection of thresholds values**.

Next section will illustrate the classic heuristic miner algorithm proposed in (A. Weijters et al., 2006).

4.2.2 The classic heuristic miner

In (W. M. P. v. d. Aalst, 2016), (A. J. M. M. Weijters et al., 2003; A. J. M. M. Weijters et al., 2011; A. Weijters et al., 2006), (De Cnudde et al., 2014), authors have presented heuristic mining algorithms by a series of steps to capture the behavior according to an event log (c.f. figure 4.3).

The traditional heuristic mining algorithms contain five main steps in order to extract a process model which represents the behavior of the log.

These steps are: (i) *identifying the footprint matrix*, (ii) *calculating the dependency measures*, (iii) *devising the graph*, (iv) *discovering the splits and joins*, and (v) *adjustments for mining loops with length of 1 and 2*.

This research work presents **an alteration at the second step**, which is one of the major steps in heuristic mining algorithms (c.f figure 4.6). The last 2 steps are not relevant to this research work, and thus will not be discussed in detail here.

In (A. J. M. M. Weijters et al., 2011), the authors define the dependency graph as the result of the first 3 steps. This term is identified as follows:

Description of the "Dependency Graph"

$$Dependency\ Graph = \{(a, b) \mid (a \in E \wedge b \in a\Box) \vee (b \in E \wedge a \in \Box b)\} \quad (4.1)$$

Here ' E ' is defined as a limited set of activities. For this set of activities several events are recorded. ' $\Box b$ ' stands for the activities that come before ' b '. ' $a\Box$ ' denotes the activities that come after ' a '. Hence, in a dependency graph each activity can have *input – output* activities which are presented as a dependency relation (a, b) . In order to devise the dependency graph, the number of times that an activity is directly followed by another one is presented in the format of a footprint matrix. Previously, heuristic mining algorithms aimed to define values among relationships known as "dependency measures". These algorithms (A. J. M. M. Weijters et al., 2003) (A. J. M. M. Weijters et al., 2011) would present different volumes of information in the process model by *manually adjusting several thresholds* within a range of -1 and 1. These values are calculated by this formula:

Formula for calculating the "dependency measure" value

$$Dependency\ Measure : a \Rightarrow_w b = \frac{|a >_w b| - |b >_w a|}{|a >_w b| + |b >_w a| + 1} \quad (4.2)$$

where ' w ' presents the event log with ' n ' number of activities and $|a >_w b|$ the number of times activity ' a ' is followed by ' b '.

The dependency measure value helps to determine the **relationship between two certain activities**. In the current application of these algorithms, experts use multiple thresholds.

Since these thresholds are determined in an arbitrary way, **the validity of the mined process model is dependent on the experience of its users**. This is a major structural challenge for heuristic mining algorithms (De Cnudde et al., 2014).

This research work addresses this issue by making an alteration to the previous approaches. At the second step, instead of applying the dependency measures calculation (equation 4.2), another method has been used to discover a state that represents the statistically stable behavior of the event log. The objective of this alteration is to be able to **discover a more structured process model** and to give the algorithm greater ability to **deal with complex processes and noise**. In addition, it removes the previous **challenge of counter-intuitively choosing the value of thresholds**.

Accordingly, this research work presents the concept of **statistical stability** and suggests two algorithms for discovering the descriptive reference process model. These algorithms have been inspired by the definition provided by Gorban in (Gorban, 2017) where he declares that **in order to represent the main behavior of a system with emergent properties such as a hospital, statistical stability needs to be found between the frequencies**.

4.2.3 Statistical stability

This phenomenon can be illustrated by an example in nature. Consider a flock of birds in the sky, as in figure 4.4a. Their flying motions can be conceived as shapes.

These shapes represent the behaviors of different groups of birds which have emergent properties. The differences in these shapes are due to the different properties of the groups. Such behaviors (shapes) can be revealed by the **statistical stability phenomenon** (Gorban, 2017).

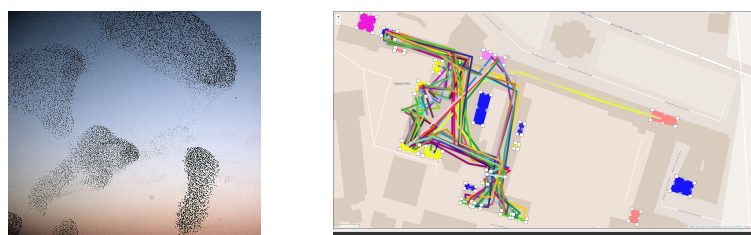
In each of these groups there are some existing deviating behaviors which at first are not seen by the eye of an observer while looking at them from distance, because these behaviors do not represent the **stable behavior** of their movements.

However, detecting these deviations is feasible. For instance, by looking closely at the upper flocks one can see these individual cases which are not following the **common behavior** of the other birds. Such an observation for detecting common and deviating behaviors is manifested by the **statistical stability phenomenon** (Gorban, 2017).

Inspired by this observation, this research work proposes two novel algorithms based on the statistical stability phenomenon to discover the stable behavior of patients from their movements in hospitals, like the example in figure 4.4b. This would help to remove the deviating behaviors which do not represent the common behavior of a group of patients and thus capture a descriptive reference process model.

But how one can relate the statistical stability to the patients' pathways?

To answer such question, this research work relies on the definition of **systems with emergent properties** provided in (Gorban, 2017). These systems consist of entities that share a *mutual objective and they create a complex whole*. This means the entities which do not share the same objective cannot represent the behavior of the system.



(a) An example of several flocks of birds. (b) A view of patients movements in a health-care facility.

Figure 4.4 – An illustration of two different systems with emergent properties.

These systems were addressed by the scientists of antiquity such as Aristotle, who stated that the *whole is greater than the sum of the parts* (Gorban, 2017). They also stated that emergent properties are manifested by monitoring the behavior of all groups of people in an environment and not only a particular case.

The statistical stability of frequencies is an important property for analyzing the common behavior of a system with emergent properties. This phenomenon is manifested not only by considering the frequency of mass events, but also by the **stability of the averages, the variances, and the standard deviations of the samples**, and this is a feature that can be inherent only in the **collection of events**. (Gorban, 2017).

As has been mentioned previously, the movements of patients within a hospital can be seen as a behavior of a system with emergent properties. Therefore, analyzing the behavior of patients' pathways by the notion of statistical stability is not just a suitable approach but it is also a necessity.

Existing in the literature	Methods & approaches	Cited research works	Contribution of this article
Similar methods for capturing complex behaviors	Heuristic mining algorithms	<ul style="list-style-type: none"> Weijters and van der Aalst 2003 Weijters et al 2011 De Cnudde et al 2014 Van der Aalst 2016 	<div style="border: 1px solid black; border-radius: 10px; padding: 10px; text-align: center;"> Stable Heuristic Miner (An algorithm for capturing the stable behavior of patients from complex location event logs, which represents the common pathways that patients take for executing their activities) </div>
Existing approaches for discovering a reference model	<ul style="list-style-type: none"> Model-Based approaches Clustering-Based approaches Profiling-Based approaches (a relatively recent approach) 	<ul style="list-style-type: none"> Ghionna et al 2008 Jalali et al 2010 Rebuge et al 2012-Bezerra et al 2013 Li and van der Aalst 2017 	
Inspiring methods for discovering a descriptive reference process model (they were not applied before)	Statistical stability	<ul style="list-style-type: none"> Montgomery 2007 Gorban 2017 	

Table 4.1 – The approach of the chapter in covering the state of the art.

To summarize this section, table 4.1 illustrates how this research work addressed the literature for answering the second scientific challenge.

After highlighting the motive for using statistical stability to monitor patients' behaviors, the next two sections introduce the steps of two algorithms that embraces the notion of statistical stability in the search for the descriptive reference process model.



Don't get confused about the statistical stability!

Well, concretely, it is not really clear how to demonstrate the statistical stability phenomenon (Gorban, 2014). Yet, let's discuss it once more.

Consider a classic weight balance, what is the objective of it? It shows the distribution of weight on two sides. If we put some unmeasured weights on the both sides of the balance, the plates will go up and down until they stop at a **stable state**.

The classic weight balance is the mean that explores stability by considering weight of two plates.

Here, the **stable heuristic miner algorithms** are the tools that evaluate the statistical stability by considering the frequencies among activities and edges in a process.

4.3 Stable Heuristic Miner V1

This algorithm targets the previously mentioned scientific questions regarding the *extraction of the descriptive reference process model*. While attempting to fill the gaps in the literature, it also addresses one of the structural problems in heuristic mining algorithms which relates to *the counter-intuitive way in which the thresholds are determined*.

4.3.1 An imaginary hypothesis

As illustrated in figure 4.5, this research work hypothesizes that an event log contains a stable amount of information that can be represented as common behaviors. Therefore, according to figure 4.5, it is feasible to imagine the desired common behaviors as the load on one side of a lever and on the other side the behaviors that the algorithm has to extract from an event log to reach the same weight as the desired common behaviors. With this in mind, the algorithm presented in this section would help to find the statistical stable state between the two sides of the lever, thus keeping it in balance.

4.3.2 Preliminaries

In order to apply the concept of statistical stability on relative frequencies, first there is a need to find a way to extract the relationships between events and the frequencies at which the sequences of events are recorded.

It should be noted that, here the relative frequencies are related to the number of times an activity is being followed and related to another one.

Accordingly, in figure 4.6, the three primary steps of heuristic mining algorithms are considered. However, an alteration is made by removing the manually configurable thresholds and replacing them by an action for statistically determining the thresholds from data. This new modification would automatically detect – from relative frequencies in an event log – the activities that represent the statistically stable behavior while removing any unstable behaviors.

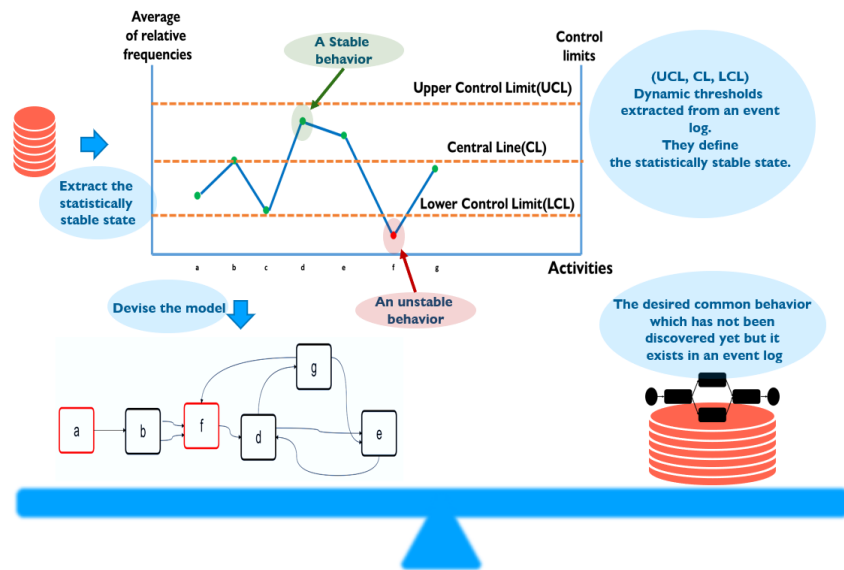


Figure 4.5 – An illustration of the hypothesis, which is to find a stable state between observed behaviors and modeled behaviors.

One of the methods used to demonstrate statistical stability is the creation of **Shewhart control charts** (Montgomery, 2019). These control charts must be devised mathematically and discovered from the event log. Eventually, they should indicate the thresholds of the statistical stable state.

Generally, control charts contain a **center line** that represents the average value of a measured characteristic, corresponding to the *in-control* state. Two other thresholds are called **Upper Control Limit (UCL)**, and **Lower Control Limit (LCL)**.

These limits are calculated by considering the standard deviations and averages of the samples. These two limits (UCL, LCL) are the borders of a statistically stable state. As long as the graphed data points fall between these two thresholds the outcomes of the process are *in-control*. If a data point falls outside these limits, it will be considered as a variation of the process outcomes, and the process will no longer be considered stable. A simple illustration of these charts is shown in figure 4.5.

The following will explore in depth the application of this notion for discovering a stable process model. At first, several assumptions and definitions have been considered.

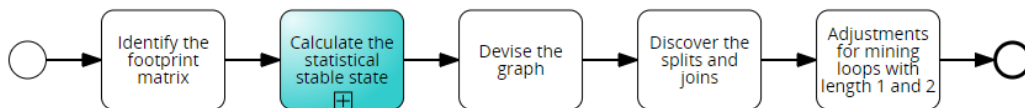


Figure 4.6 – The sequence of actions for applying stable heuristic miner.

4.3.2.1 Assumptions

Inspired by Shewhart control charts (Montgomery, 2019), three main assumptions have been made to find the statistically stable behaviors of processes from an event log.

1. The first assumption is the normality of the data distribution regarding the relative frequencies among registered activities in the event log. Most of the statistical methods for investigating the data are highly dependent on the distribution of data. If the data is not normally distributed, then certain adjustments in method should be considered so as to adapt the method to the distribution function of the data. This normality assumption is made due to the degree of freedom in which that data distribution could change. This assumption is authentic and justifiable by the Central Limit Theorem (Bárány et al., 2007) and the statistical process control paradigm (Montgomery, 2019).
2. The second assumption is the one-sided-stability assumption. This means considering and presenting the activities that have an average of their frequency greater than UCL. Usually, by applying the new method, activities with a high level of variations will be outside of UCL, which is correct mathematically. These activities are considered as not coming within the stable behavior of the log. However, such activities could provide information regarding the behaviors that show higher levels of variations and are thus the cause of the instability in the whole behavior. So as not to ignore these activities, they are displayed as “hot zones”, color-coded red. The reason behind this is that it is extremely rare to extract a pattern from an event log that shows all the activities illustrating a stable behavior. Therefore, the deviating activities that are causing instability in a model by a higher level of variations are presented in the process model.
3. In the footprint matrix the last activity will have ‘0’ values, since it will not be followed by any other activity. Here, in order to not avoid the ending zone of the process, the last activity will be considered as an “end activity” with one ‘observation’.

4.3.2.2 Definitions and the sequence of functions in the algorithm

This research work uses the definitions for event log, traces, and events given in (W. M. P. v. d. Aalst, 2016). Accordingly, the relations among these concepts are presented in chapter 3 by the DIAG meta-model within the **location event logs package**.

To illustrate the basic concepts of the stable heuristic miner *V1* algorithm, the following running example is used.

Consider the event log below:

$$L = [< a, b, c, d, e, l, m >^{12}, < a, b, f, d, e, l, m >^2, < a, b, c, d, g, e, l, m >, \\ < a, b, c, d, g, h, e, l, m >^5, < a, c, b, c, d, l, m >^6, < a, c, f, c, d, l, m >^3, \\ < a, c, b, i, c, e, c, d, g, l, m >^5, < a, c, b, c, f, c, d, l, m >^6, \\ < a, c, b, c, i, c, h, c, d, l, m >^4, < a, c, b, c, i, f, c, d, g, l, m >^6, < a, b, c, j, l, m >^6, \\ < a, c, b, k, e, d, l, m >^4]$$

Each group represents a trace. Within each of them, there are events corresponding to the activities. For example, the first trace $< a, b, c, d, e, l, m >^{12}$ shows that 12 cases have followed the same sequence of activities.

According to figure 4.6, the first action is to “**identify the footprint matrix**”. As a result, the first definition here is:

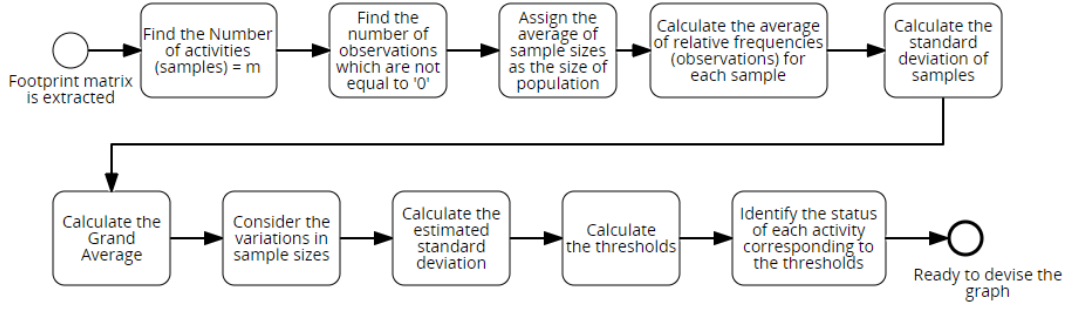


Figure 4.7 – The sequence of applying calculations for the stable heuristic miner algorithm.

Definition 1 Population (S):

The footprint matrix here is considered as the population in which all of the relative frequencies are presented. This matrix represents the relations among different activities within an event log.

The matrix below shows an example of the above-mentioned definition, according to (L). For example, this matrix shows that activity 'a' is followed 26 times by activity 'b'.

$$S = \begin{pmatrix} a & b & c & d & e & f & g & h & i & j & k & l & m \\ a & 0 & 26 & 34 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ b & 0 & 0 & 46 & 0 & 0 & 2 & 0 & 0 & 5 & 0 & 4 & 0 \\ c & 0 & 31 & 0 & 48 & 5 & 9 & 0 & 4 & 10 & 6 & 0 & 0 \\ d & 0 & 0 & 0 & 0 & 14 & 0 & 17 & 0 & 0 & 0 & 23 & 0 \\ e & 0 & 0 & 5 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 20 & 0 \\ f & 0 & 0 & 15 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ g & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 5 & 0 & 0 & 11 & 0 \\ h & 0 & 0 & 4 & 0 & 5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ i & 0 & 0 & 9 & 0 & 0 & 6 & 0 & 0 & 0 & 0 & 0 & 0 \\ j & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 6 & 0 \\ k & 0 & 0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ l & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 60 \\ m & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Now, the second step in figure 4.6 is to "calculate the statistical stable state". Therefore, to calculate this state several actions are needed. Figure 4.7 shows this sequence. In accordance with figure 4.7, the next definition corresponds to finding the number of samples.

Definition 2 Sample (s):

Each row in the population presents a vector that shows the relationships of an activity with other activities. Therefore, the corresponding vectors for the activities in the event log are considered as the samples of the population.

As a consequence, in the example above there are 13 samples (vectors related to each activity = s , which relates to $i = a$ to m). This number is identified by ' m ' (here: $m = 13$).

Definition 3 Observations (x_{ij}):

The values within the population (footprint matrix) are considered as the observations.

Each observation is identified by ' x_{ij} ', where ' i ' stands for the rows in the footprint matrix and ' j ' represents the columns.

These observations present the relative frequencies among existing samples. The total number of observations within the population is identified by ' N '. For the example shown here, $N = 30 + 1$. The '+1' is an adjustment that has been made based on the third assumption, that the "end activity" is not considered as a null sample.

Each sample has a size that corresponds to the number of observations within each vector. Thus, in this case the sample size is equal to the number of dependency values for each activity that are greater than '0'. For example, the sample size for activity ' a ' is equal to 2 ($n_{s_a} = 2$).

It is very important to note that here the sizes of the samples are not necessarily similar. Therefore, two further concepts need to be defined in order to consider this challenge. These definitions are the *grand average*, and the *estimated standard deviation*. These two indicators will determine the statistical stability within the population (Montgomery, 2019).

Definition 4 Grand average ($\bar{\bar{x}}$):

Since the size of samples is a changing variable, $\bar{\bar{x}}$ has been defined to express the average of relative frequencies (x_{ij}) within the whole population. Equation 4.3 shows how the grand average would be calculated. Note that m is the number of samples. \bar{x}_i stands for the average of the ' i th' sample.

$$\bar{\bar{x}} = \frac{\sum_{i=1}^m n_i \bar{x}_i}{\sum_{i=1}^m n_i} \quad (4.3)$$

For the example above, the grand average of the entire population is equal to 14.22.

Definition 5 Estimated Standard deviation ($\hat{\sigma}$):

Similarly to the grand average, the estimated standard deviation $\hat{\sigma}$ has been defined by equation 4.6 in order to understand how the behavior within the population deviates from one sample to another while the sample sizes differ.

Two prior steps are needed before calculating the estimated standard deviation ($\hat{\sigma}$). First, the standard deviation of each sample is calculated (σ_i). Then, a factor is introduced as C_{4n} (Montgomery, 2019). This factor is dependent on the size of each sample. The main reason for using C_{4n} is that the sizes of samples are not the same. Therefore, this factor will help to extract **an unbiased** $\hat{\sigma}$.

Equation 4.4 shows the basic method for calculating the standard deviation of each sample. Note that x_{ij} is an observation within a sample; n_i is the size of the sample, and \bar{x}_i is the average of observations for the i th sample.

$$\sigma_i = \sqrt{\frac{(1)}{n_i - 1} \sum_{j=1}^m (x_{ij} - \bar{x}_i)^2} \quad (4.4)$$

In order to measure the C_{4n} factor for each sample, equation 4.5 will be used.

$$C_{4n} = \frac{4(n - 1)}{4n - 3} \quad (4.5)$$

For the mentioned example, the values of C_{4n_i} and σ_i are presented in table 4.2.

Activities	(\bar{x}_i)	(σ)	C_{4n}
a	$\bar{x}_1 = \frac{26+34}{2} = 30$	5.65	0.80
b	$\bar{x}_2 = \frac{46+2+5+4}{4} = 14.25$	21.2	0.92
c	$\bar{x}_3 = \frac{31+48+5+9+4+10+6}{7} = 16.14$	16.82	0.96
d	$\bar{x}_4 = \frac{14+17+23}{3} = 18$	4.5	0.88
e	$\bar{x}_5 = \frac{5+4+20}{3} = 9.6$	8.9	0.88
f	$\bar{x}_6 = \frac{15+2}{2} = 8.5$	9.19	0.80
g	$\bar{x}_7 = \frac{1+5+11}{3} = 5.6$	5.03	0.88
h	$\bar{x}_8 = \frac{4+5}{2} = 4.5$	0.7	0.80
i	$\bar{x}_9 = \frac{9+6}{2} = 7.5$	2.12	0.80
j	$\bar{x}_{10} = \frac{6}{1} = 6$	0	0
k	$\bar{x}_{11} = \frac{4}{1} = 4$	0	0
l	$\bar{x}_{12} = \frac{60}{1} = 60$	0	0
m	$\bar{x}_{13} = \frac{0}{1} = 0$	0	0
Grand Average $(\bar{\bar{x}}) = 14.22$			

Table 4.2 – The calculation of \bar{x} , C_{4n} , and σ for each sample and the grand average of the population.

Now by considering these values, the equation 4.6 is able to calculate the estimated standard deviation of the population:

$$\hat{\sigma} = \frac{1}{m} \sum_{i=1}^m \frac{\sigma_i}{C_{4n_i}} \quad (4.6)$$

For the example above, the value of $\hat{\sigma}$ is equal to 6.42.

$$\begin{aligned} \hat{\sigma} = \frac{1}{13} \times & \left[\frac{5.65}{0.79} + \frac{21.20}{0.92} + \frac{16.82}{0.95} + \frac{4.5}{0.88} + \frac{8.96}{0.79} + \frac{9.19}{0.79} + \frac{5.03}{0.88} + \frac{0.70}{0.79} + \frac{2.12}{0.797} \right. \\ & \left. + 0 + 0 + 0 + 0 \right] = 6.42 \end{aligned} \quad (4.7)$$

After acquiring these metrics, the algorithm can construct the control limits (thresholds) required to extract the stable behavior.

Definition 6 Central Line (CL):

As has been shown in equation 4.8, 'CL' represents the most stable behavior, or in other words the core behaviors within the log. The activities whose average of recorded observations is close to CL are normally present in most traces of the log, and they are thus the core activities.

$$CL = \bar{\bar{x}} \quad (4.8)$$

In the example above, CL is equal to 14.22.

In the literature on the subject of determining the stable state, a certain distance from the CL is allowed (Montgomery, 2019). Previously, a distance of 3σ was used for samples with unique sizes. In this research work, since the sample sizes are changing variables, the distance is defined by considering the population estimated standard deviation and two defined constants ($A_{3\bar{n}}C_{4\bar{n}}\hat{\sigma}$).

This distance helps to define the two other limits or borders of the stable state.

Definition 7 Lower Control Limit (LCL):

This threshold filters the behaviors that do not represent the main and stable behavior of the log. The activities with an average (\bar{x}_i) lower than LCL will not be shown in the process model.

Therefore, the algorithm will consider the activities that have a stronger presence in the behavior within the event log and it will remove the deviations. Equation 4.9 will use the previous definitions to determine this threshold.

$$LCL = \bar{\bar{x}} - (A_{3\bar{n}} \times C_{4\bar{n}} \hat{\sigma}) \quad (4.9)$$

Note that $A_{3\bar{n}}$ is a customary constant for considering a previously defined distance from CL (Montgomery, 2019). It can be calculated by equation 4.10.

$$A_{3\bar{n}} = \frac{\bar{n}}{C_{4\bar{n}} \sqrt{\bar{n}}} \quad (4.10)$$

Since the thresholds apply to the whole population, the formula would consider the average of all the sample sizes for calculating the $A_{3\bar{n}}$ and $C_{4\bar{n}}$ factors. Therefore: $\bar{n} = \text{Average of sample sizes} = \frac{N}{m}$

Which is equal to the total number of observations divided by the total number of samples.

Corresponding to the mentioned example, the value of m is equal to 13. Also, the number of observations has been indicated: $N = 31$. Therefore, $\bar{n} = \frac{31}{13} \approx 3$. And, $A_{3\bar{n}} = 1.94$.

Definition 8 Upper Control Limit (UCL):

The value of UCL sets the bar for activities with the maximum amount of variations with regard to the whole population.

$$UCL = \bar{\bar{x}} + (A_{3\bar{n}} \times C_{4\bar{n}} \hat{\sigma}) \quad (4.11)$$

As a result, activities with \bar{x} greater than UCL are considered here as the zones where their behavior causes the process to be unstable. This could lead to bottlenecks at these activities while executing the process. Such activities would generate behaviors that do not normally correspond to the behavior of the whole population.

Thereby, in the example above these thresholds are equal to:

- $UCL = 14.22 + (1.94) \times (0.88) \times (6.42) \approx 26$
- $CL = 14.22$
- $LCL = 14.22 - (1.94) \times (0.88) \times (6.42) \approx 4$

Definition 9 Statistically stable state:

Finally, the conditions are then defined in equation 4.12 for determining which activities express a stable behavior in accordance with the **whole recorded information** in an event log.

$$LCL < \bar{x}_i < UCL \quad (4.12)$$

Normally, if all of the recorded activities in an event log express a stable behavior, no activity will be removed. This could imply that the process is running smoothly. But, if a variation exists in the behaviors, it will be detected by means of the two thresholds (UCL, LCL).

Definition 10 determines which activities will be considered within the modeled common behaviors of the event log.

Definition 10 Descriptive reference process model (\mathcal{P}):

The descriptive reference process model or the common behaviors will contain activities that respect the followings conditions:

Description of the descriptive reference process model by the first version of stable heuristic miner algorithm

$$[\forall \mathcal{A} \exists s] \wedge [\forall s \subseteq \mathcal{S} \exists \bar{x}_s] \therefore (\mathcal{A} \in \mathcal{P}) \rightarrow [LCL < \bar{x}_s < UCL] \cup [UCL \leq \bar{x}_s] \quad (4.13)$$

Definition 10 states that for each activity (\mathcal{A}), a vector of relative frequencies with other activities exists. This vector is defined as a sample of the population (footprint matrix). And, for each sample there exists a \bar{x}_s which represents the average of relative

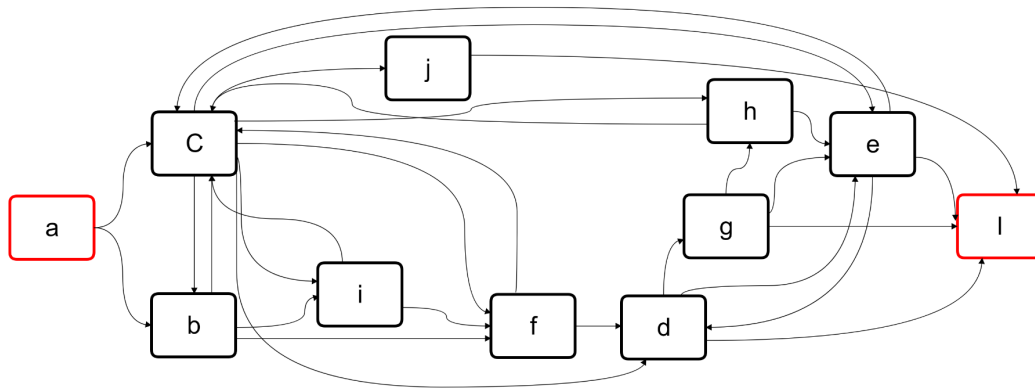


Figure 4.8 – An illustrative example of the algorithm outcome. The descriptive reference process model representing stable behavior of the example event log (L). Red activities correspond to high variation in behaviors. Two activities are removed with lower significance level for the general behavior.

frequencies. Therefore, the corresponding activity to the sample (s) will be represented in the descriptive reference process model (\mathcal{P}) if the average of its relative frequencies is between the two thresholds (considered as stable behavior) or if it is greater than the UCL value (considered as the hot zones).

Regarding the above example, table 4.3 presents which activities will be considered as the entities of the stable behavior. Accordingly, figure 4.8 illustrate the descriptive reference process model for the example log (L).

It can be seen that from 13 presented activities of the event log, 2 activities are filtered (m and k) and only 2 activities (a , and l) impose high instability in the behavior.

The steps of the stable heuristic miner algorithm could be realized by the set of algorithms presented in ("*extract the thresholds*" and "*identify the status of activities*") that show how to devise the descriptive reference process model from the footprint matrix.

One of the ambiguities is how the applicability of discovering algorithms can be ensured. This question requires *evaluating the outcome of process discovery algorithms*. In order to evaluate the applicability of the presented algorithm, a case study is presented in chapter 6.

The next section will present the second version of this algorithm which addresses statistical stability for both **activities** and **edges**.

Activities	Lower than LCL	Higher than UCL (Hot zones)	Stable state $LCL < \bar{x} < UCL$
a	No	Yes	No
b	No	No	Yes
c	No	No	Yes
d	No	No	Yes
e	No	No	Yes
f	No	No	Yes
g	No	No	Yes
h	No	No	Yes
i	No	No	Yes
j	No	No	Yes
k	Yes	No	No
l	No	Yes	No
m	Yes	No	No

Table 4.3 – Presenting the behavior of the activities in the event log.

Algorithm 7 Stable Heuristic Miner V1

```

1: procedure Extract the thresholds
2:   Input Footprint.Matrix
3:   Input Activity.Set
4:   Output UCL, LCL, CL, Sample.Attr
5:    $m = \text{length}(\text{Activity.Set})$  ▷ get number of activities
6:   ▷ total number of observations
7:    $\text{Total.Observations} = \text{sum}(\text{rowSums}(\text{Footprint.Matrix} \neq 0))$ 
8:    $\bar{n} = (\text{Total.Observations}/m)$  ▷ average size of the population
9:   ▷ get the size and mean of each activity
10:  for  $i$  in  $\text{Footprint.Matrix}[1:n]$  do
11:     $\text{Sample.Attr} = \text{data.frame}(\text{size}=\text{rowSums}(\text{Footprint.Matrix}[i,j] \neq 0),$ 
12:     $\bar{x}_i = \text{rowMeans}(\text{Footprint.Matrix}[i,j] \neq 0),$ 
13:     $\sigma = \text{rowStandardDeviations}(\text{Footprint.Matrix}[i,j] \neq 0);$ 
14:     $\text{mutate}(\text{Sample.Attr}, C_{4n})$  ▷ equation 4.5
15:  end for
16:   $\text{Sample.Attr}[\text{size}, \sigma, C_{4n}, \bar{x}_i]$  ▷ structure of Sample.Attr
17:  for  $i$  in  $\text{Sample.Attr}[\bar{x}_i]$  do ▷ equation 4.3
18:     $\bar{\bar{x}} = (\text{sum}(\text{Sample.Attr}[\bar{x}_i]) / m)$ 
19:  end for
20:  for  $i$  in  $\text{Sample.Attr}[1:n]$  do
21:     $\text{mutate}(\text{Sample.Attr}, \sigma_i / C_{4n_i})$ 
22:  end for
23:  for  $i$  in  $\text{Sample.Attr}[1:n]$  do
24:     $\hat{\sigma} = ((1/m) * \text{sum}(\sigma_i / C_{4n_i}))$  ▷ equation 4.6
25:  end for
26:   $A_{3\bar{n}} = 3 / C_{4\bar{n}} * \text{sqrt}(\bar{n})$ 
27:   $CL = \bar{\bar{x}}$  ▷ devise the thresholds
28:   $UCL = \bar{\bar{x}} + (A_{3\bar{n}} * C_{4\bar{n}} * \hat{\sigma})$ 
29:   $LCL = \bar{\bar{x}} - (A_{3\bar{n}} * C_{4\bar{n}} * \hat{\sigma})$ 
30: end procedure

```

Algorithm 8 Stable Heuristic Miner V1

```
1: procedure Identify the status of activities
2:   Input  $UCL, LCL, CL$ 
3:   Input  $Sample.Attr$ 
4:   Input  $Activity.Set$ 
5:   Output  $Graph$ 
6:    $\triangleright$  considering the average of relative frequencies for each sample
7:   for  $i$  in  $Sample.Attr[\bar{x}_i]$  do
8:      $Unstable.Activities = \bar{x}_i \leq LCL;$ 
9:      $Stable.Activities = LCL < \bar{x}_i < UCL;$ 
10:     $Hot.Zones = UCL \leq \bar{x}_i$ 
11:   end for
12:    $\triangleright$  Select the nodes
13:    $Stable.Nodes = match(Stable.Activities, Activity.Set)$ 
14:    $Hot.nodes = match(Hot.Zones, Activity.Set, Color.Attr = "red")$ 
15:    $All.Nodes = combine(Stable.Nodes, Hot.Nodes)$ 
16:    $\triangleright$  Select the edges
17:    $edges = match(Footprint.Matrix, All.Nodes)$ 
18:    $devise.graph(All.Nodes, edges)$ 
19: end procedure
```

4.4 Stable Heuristic Miner V2

The previous version of the algorithm applied statistical stability method only by considering **activities** as “samples”. However, the stability was not evaluated for the behavior of edges.

Henceforth, in the second version, the approach is improved. Instead of only considering activities, **the focus is on both each edge’s behavior and ensemble of activities as well.**

For instance, instead of only detecting hot zones, stable activities, or unstable activities; the second method will additionally discover **hot edges, stable edges, and unstable edges** too.



An important addition

Consequently, a two-way approach is used.

- First, the stable activities are discovered thanks to the first version of the algorithm.
- Second, the statistical stability is evaluated for *each edge* individually. Therefore, **each edge is considered as a sample with the size of 1.**

4.4.1 Preliminaries

It might seem an odd approach to select samples with the size of 1. However, in statistics, there are constraints that force to select such samples (Montgomery, 2019).

This novel method helps to be more accurate about **detecting the individual deviating behaviors.**

For achieving this, the **individual control charts** method (Montgomery, 2019) became the inspiration for the stable heuristic miner V2 algorithm. This algorithm discovers the thresholds for determining the statistically stable state of **edges**.

In essence, the previous assumptions and the objective for detecting a stable state are still authentic for this new version as well.

However, some of the definitions for measuring the stable state, hot edges and unstable edges are completely different.

Again, the same running example (*L*) is used to explain this algorithm.

To be precise, the definitions for this algorithm starts by the *Population* which is identical to the previous definition 1. Therefore, to avoid over-explanation, it will not be re-introduced.

In previous definitions, each row in the footprint matrix (*population*) represented a *sample*. In the second version, the algorithm in parallel to mining activities behavior, it considers **each value** of the footprint matrix as a **sample** of the population.

Definition 11 Sample-edge (s_e):

*Each connection or an edge between two activities with a value greater than ‘0’ is considered as a **sample**. All the samples have a unique size of 1.*

Accordingly, in the example above there are '30 + 1' samples (edges).

Definition 12 Observation-edge (x_{ij}):

Each edge or sample within the population has a value, this value is relative to the Observation value.

As an example, in the previous footprint matrix, the value of $a \rightarrow b$ is equal to 26. Therefore, $a \rightarrow b$ is a **sample** of the population (footprint matrix), and the value of this sample is an **observation**.

The value of each observation will be used to build upon the extraction of a statistically stable state. Needless to mention, there are 31 edges in this example and accordingly '31' observation values (one value is added in accordance to the assumption 3).



Important to consider

As mentioned before in the definition of the statistical stability phenomenon, it is a necessity to understand the behavior of the collection of mass events in order to capture stable and unstable behaviors.

To do so, one must monitor the variations, and the average value of the analyzed metric.

Since, the **sample size here is unique** and equals to '1', a new variable is considered as the **Moving Range (MR)** to monitor the variations among samples.

Definition 13 Moving Range (MR):

MR represents the difference from one observation(x_{ij}) to another.

For instance, the $x_{ab} = 26$ and $x_{ac} = 34$. Therefore, the MR from x_{ab} to x_{ac} is equal to '8'.

The value of MR helps to consider the variations in samples behaviors.

Definition 14 Average behavior (\bar{x}):

As it can be expected, this value represents the average of all the samples(edges) values.

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n} \quad (4.14)$$

Definition 15 Average of Moving Range (\bar{MR}):

While the value for each edge changes from one to another, the \bar{MR} gives a value to represent the average of variations.

$$\bar{MR} = \frac{MR_1 + MR_2 + MR_3 + \dots + MR_n}{n} \quad (4.15)$$



Important to consider

After calculating these values, it is possible to construct the thresholds for evaluating the statistical stability among edges. It must be noted that, the formation of these thresholds is following the same class of methods; control charts. But, the type and calculation of these methods are completely different.

Definition 16 Central Line-edges ($CL.edges$):

This central threshold determines the most statistically stable edges (samples). More the value of an edge is closer to this line, the more expected for this edge to be seen in future behaviors.

As shown in equation 4.16, the most stable behaviors are close to the average behavior.

$$CL.edges = \bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n} \quad (4.16)$$

Definition 17 Upper Control Limit-edges ($UCL.edges$):

This threshold determines the limit to conceive edges behavior as stable. If an edge value passes this threshold, it will be considered as an edge with high variation ratio in its behavior.

Equation 4.17 presents the mathematical model of UCL . Note that, this formula is not the same as the previous one for stable heuristic minerV1.

A customary constant as ' d_2 ' is defined here. The value for ' d_2 ' is defined through series of calculus operations and it has led to certain constant values. d_2 values are depending to the number of monitored samples in a population. Additionally, these values are presented in most of statistics handbooks (Montgomery, 2019).

$$UCL.edges = \bar{x} + 3\left(\frac{\bar{MR}}{d_2}\right) \quad (4.17)$$

Definition 18 Lower Control Limit-edges ($LCL.edges$):

The LCL value ascertains the unstable behaviors. The edges with values less than LCL are not expressed within the discovered process model. This threshold helps to remove the so-called "dirt-roads" from a process model automatically.

Equation 4.18 refers to the mathematical model for extracting LCL value.

$$LCL.edges = \bar{x} - 3\left(\frac{\bar{MR}}{d_2}\right) \quad (4.18)$$

Definition 19 Descriptive reference process model V2 (P)

The updated definition of the descriptive reference process model or the common behaviors will contain **activities** (A) and **edges** (E) that respect the followings conditions:

Description of the descriptive reference process model by the second version of stable heuristic miner algorithm

$$\begin{aligned}
 & [[\forall \mathcal{A} \exists s] \wedge [\forall s \subseteq \mathcal{S} \exists \bar{x}_s]] \wedge [[\forall \mathcal{E} \exists s_e] \wedge [\forall s_e \subseteq \mathcal{S} \exists x_{s_e}]] \\
 & \therefore \\
 & ((\mathcal{A}, \mathcal{E}) \in \mathcal{P}) \rightarrow [LCL < \bar{x}_s < UCL] \cup [UCL \leq \bar{x}_s] \\
 & \wedge \\
 & [LCL.edge < x_{s_e} < UCL.edge] \cup [UCL.edge \leq x_{s_e}]
 \end{aligned} \tag{4.19}$$

This definition expresses the conditions for the sets of **activities** (\mathcal{A}) and **edges** (\mathcal{E}) to be included within the descriptive reference process model (\mathcal{P}).

As shown, for all activities (\mathcal{A}) there exists a sample which represents the behavior of each activity. A value as \bar{x}_s shows the average behavior of each activity.

Additionally, the behavior of all the edges (\mathcal{E}) is presented by a sample which has a value of x_{s_e} . Therefore, a set of an activity and its edges is within the descriptive reference process model (\mathcal{P}), if the average of the activity behavior (\bar{x}_s) is between the UCL and LCL . If \bar{x}_s is greater than UCL then the activity is considered as an activity with high instability.

Simultaneously, edges (and the linked activities) are within \mathcal{P} if their values (x_{s_e}) falls between the two thresholds, $UCL.edge$ and $LCL.edge$. If x_{s_e} is greater than $UCL.edge$, then this edge is considered as an edge with high instability as well.

Note that the new evaluation of $UCL.edges$, $LCL.edges$, and $CL.edges$ are in parallel with the previous calculation of the first version of the algorithm. Basically, this branch of calculations is added to improve a previous weakness in the algorithm.

Figure 4.9 shows the steps for the new version of the algorithm which is demonstrated by the following example. This figure represents how the new branch of calculations are executed for extracting the statistical stability for edges as well as for the activities.

4.4.2 Order of calculation

After devising the footprint matrix out of the extracted event log (c.f. figure 4.6), certain sequence of calculations will be taken to discover the thresholds of the stable state. This sequence is presented in figure 4.9.

BPMN is used to illustrate the association between the previous version of the algorithm (*upper lane*: statistical stability miner for the activities) and the new branch of calculations (*lower lane*: statistical stability miner for edges).

The upper lane in figure 4.9 is presented within the context of stable heuristic miner V1. Therefore, the following expresses the result of applying the new calculations (lower lane).

The first task is to count the number of edges with the value greater than 0. Previously, this value is mentioned, 31. Then, all the edges and their values will be gathered in a data table (c.f. table 4.4). **These values will be recorded based on the order in which activities are recorded in the footprint matrix.** For instance, if primarily activity ‘a’

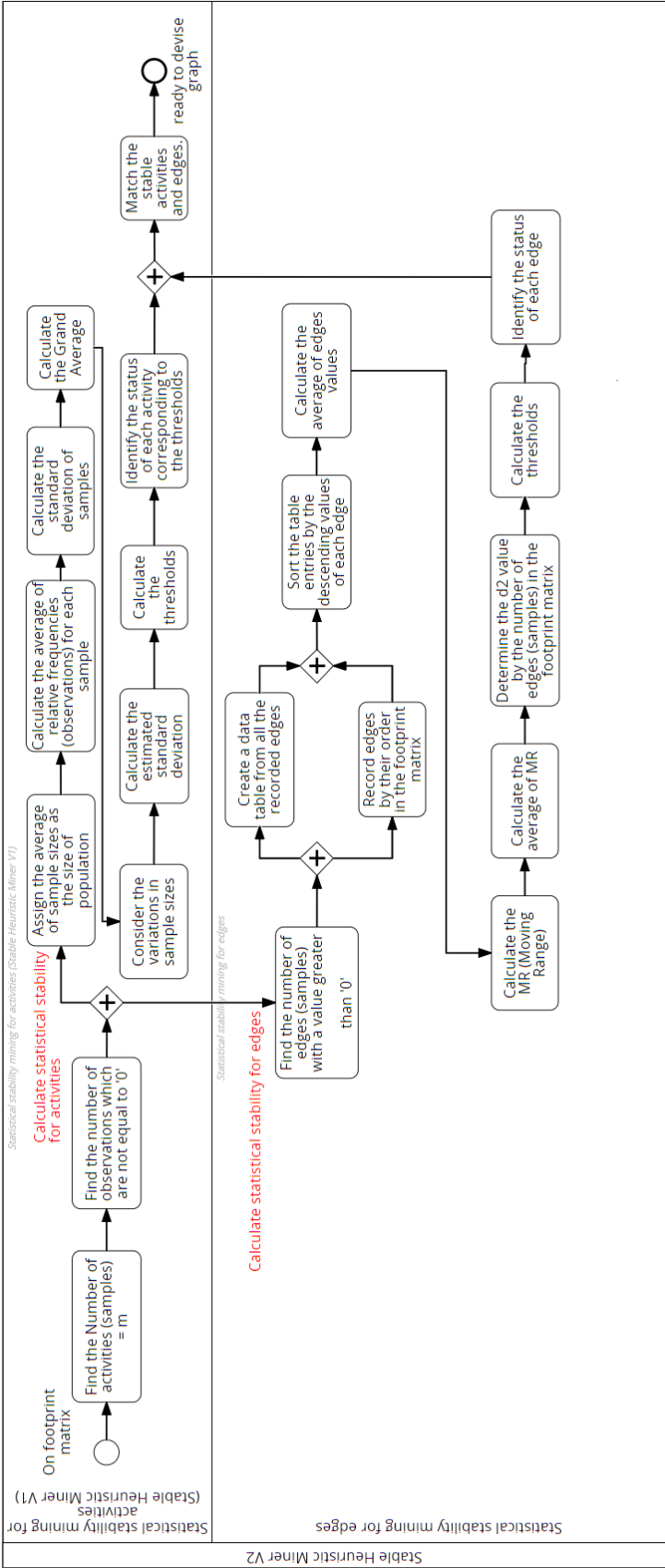


Figure 4.9 – Order of tasks to extract the stable state by the stable heuristic miner V2.

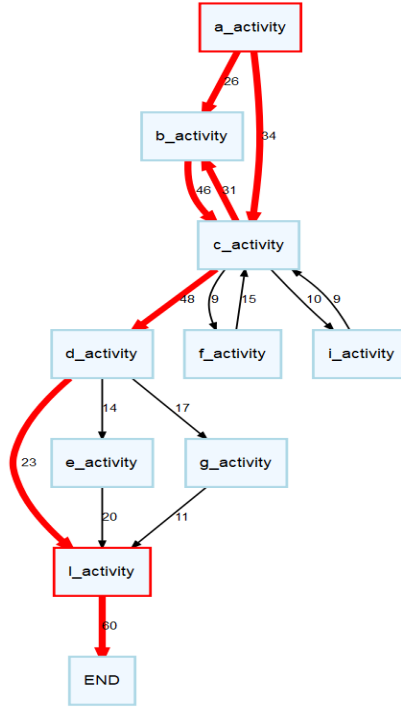


Figure 4.10 – The result of applying stable heuristic miner V2. Statistical stability is applied for both edges and activities behaviors.

is recorded in the footprint matrix, then, all the edges from ‘a’ to other activities will be recorded at first.

After this step, for each activity, edges will be sorted by a **descending order**. As an example, after considering activity ‘a’, firstly, the edge ‘a → c’ with a value of 34 will be recorded, then ‘a → b’ will be recorded with a value of 26.

After sorting the edges, the average behavior \bar{x} will be calculated. Then, MR values or the changes in the values of edges will be calculated. This will lead to calculating the average of moving ranges (\bar{MR}).

Accordingly, the average value of variations is $\bar{MR} = 10.51$. The average value for the recorded behavior of edges is $\bar{x} = 14.225$.

The value for d_2 will be extracted from the d_2 table. For 31 samples in the population, d_2 is equal to ‘4.113’.

Finally, the state of each edge will be determined according to the previous formulas 4.16, 4.17, and 4.18.

- **UCL** = $14.225 + (3)\left(\frac{10.51}{4.113}\right) \approx 22$
- **CL** = **14.225**
- **LCL** = $14.225 - (3)\left(\frac{10.51}{4.113}\right) \approx 7$

Table 4.4 presents the results of performing this new sequence of calculations for the previously mentioned event $\log(L)$.

Order	FirstActivity	SecondActivity	x (edge value)	MR
1	a	c	34	0
2	a	b	26	8
3	b	c	46	20
4	b	i	5	41
5	b	k	4	1
6	b	f	1	3
7	c	d	52	51
8	c	b	31	21
9	c	i	10	21
10	c	f	6	4
11	c	j	6	0
12	c	e	5	1
13	c	h	4	1
14	d	l	23	19
15	d	g	17	6
16	d	e	14	3
17	e	l	20	6
18	e	c	5	15
19	f	c	15	10
20	f	d	2	13
21	g	l	11	9
22	g	h	5	6
23	g	e	1	4
24	h	e	5	4
25	h	c	4	1
26	i	c	9	5
27	i	f	6	3
28	j	l	6	0
29	k	e	4	2
30	l	m	60	56
$\bar{x} = 14.566$			$MR = 11.133$	

Table 4.4 – The table presenting the result of performing the sequence of calculation for the stable edges miner algorithm. In the “edge value” column, the red cells are related to the hot edges, the violet ones are representing “dirt roads”, and the other cells represent the stable edges.

After performing these calculations, it is feasible to extract the state of each edge by the value of thresholds and devise the process model.

Figure 4.10 presents the final result of applying stable heuristic miner V2.

As shown in this model, out of 13 activities only 10 are considered within the process model which have substantial behavior. From 31 primary registered edges, 15 edges are present in the final model and 7 of these edges are expressing high variations comparing to the total population. This selection of this set of activities and edges (\mathcal{A}, \mathcal{E}) are due to the new definition of the descriptive reference process model (\mathcal{P}).

The general form of this algorithm is presented in the following; algorithm 9 and 10. Note that the first two algorithms for presenting stable heuristic miner V1 are used as well to determine the state of activities.

Algorithm 9 Stable Heuristic Miner V2

```

1: procedure Extract the thresholds
2:   Input Footprint.Matrix
3:   Input Edges.Set
4:   Output UCL, LCL, CL
5:    $m = \text{length}(\text{Edges.Set})$  ▷ get number of edges
6:   ▷ total number of observations
7:   in Footprint.Matrix
8:     gather (columns, values, row);
9:     arrange (row, -val); ▷ sorting in a descending order
10:    filter (val !=0);
11:    create data.table1
12:    in data.table1
13:      mutate( $MR = \text{absolute}(x_i - x_{i-1})$ );
14:      mean( $x_i$ )
15:
16:     $CL = \bar{x}$  ▷ devise the thresholds
17:
18:     $UCL = \bar{x} - 3(\frac{MR}{d_2})$ 
19:
20:     $LCL = \bar{x} + 3(\frac{MR}{d_2})$ 
21: end procedure

```

Algorithm 10 Stable Heuristic Miner V2

```

1: procedure Identify the status of activities and edges
2:   Input  $UCL, LCL, CL$ 
3:   Input  $Edges.Set$ 
4:   Input  $Activity.Set$ 
5:   Input  $data.table1$ 
6:   Output  $Graph$ 
7:    $\triangleright$  considering the average of relative frequencies for each sample
8:   for  $i$  in  $data.table1[x_i]$  do
9:      $Nodes.Stable.Edges = unique(subset(data.table1, LCL < x_i < UCL))$ 
10:     $Nodes.Unstable.Edges = unique(subset(data.table1, x_i \leq LCL));$ 
11:     $Node.Hot.Edges = unique(subset(data.table1, UCL \leq x_i))$ 
12:   end for
13:    $All.Nodes = match(combine(Nodes.Stable.Edges, Nodes.Unstable.Edges),$ 
     $Activity.Set);$ 
14:    $\triangleright$  Select the nodes
15:    $Stable.edges = match(Nodes.Stable.Edges, data.table1[FirstActivity, SecondActiv-$ 
     $ity]);$ 
16:    $Hot.edges = match(Node.Hot.Edges, data.table1[FirstActivity, SecondActivity],$ 
     $Color.Attr = "red");$ 
17:    $Process.edges = combine(Stable.edges, Hot.edges)$ 
18:    $\triangleright$  Select the edges;
19:    $devise.graph(All.Nodes, Process.edges)$ 
20: end procedure

```

4.4.3 Comparing the results of the algorithms

Figure 4.11, 4.12, 4.13, 4.14, and 4.15 present the results of classic heuristic miner for the mentioned event log. As it is in nature of this method, the thresholds for determining the level of dependency measures for activities in the event log is set in an arbitrary manner which is considered as a disadvantage of this algorithm (De Cnudde et al., 2014).

In essence, the new application of **statistical stability** ensures about detection of a stable amount of information from an event log. This task will be carried out by certain **reactive thresholds** which are determined by statistical stability methods.

Therefore, by using the stable heuristic miners, thresholds are not arbitrary selected anymore.

As a result, activities and edges with insignificant behaviors will be removed. A common path will present the stable behavior in an event log, and the most variant behaviors will be detected as well.

Figure 4.16 and 4.17 present a comparison for the outcomes of two proposed algorithms; stable heuristic miner V1 and stable heuristic miner V2.

As it has been shown in figure 4.16, the first version of stable heuristic miner determines automatically the statistical stability thresholds **only for activities**. This led to filtering certain insignificant behaviors in the log.

Eventually, the latest update on the algorithm led to the automatically calculation of thresholds for **both activities and edges behaviors**. Figure 4.17 shows the result of this algorithm for mining a stable amount of information from the event log.

The descriptive reference process model extracted by this algorithm is more informative in comparison with the first version of the classic heuristic miner. This is due to automatic detection of stable and unstable behaviors.

In the end, within the current literature of process mining, one can not decide that a certain algorithm is the optimal algorithm (W. M. P. v. d. Aalst, 2013b), (Augusto et al., 2019a). This is due to the absence of an overall evaluation procedure (De Cnudde et al., 2014).

However, here as shown in figure 4.5, the quest was to find an endorsing method to **automatically** extract the **stable amount of information** from event logs **without the need to manually (and arbitrary) modify a filtering value**.

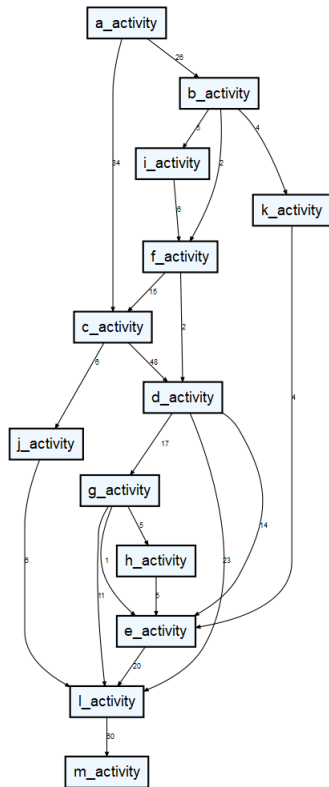


Figure 4.11 – The process model extracted by the classic heuristic miner approach with a manual thresholds set at **20%**. Accordingly, the model presents activities that have a dependency measure higher than 20%. Only 67% of the recorded behaviors respect this threshold.

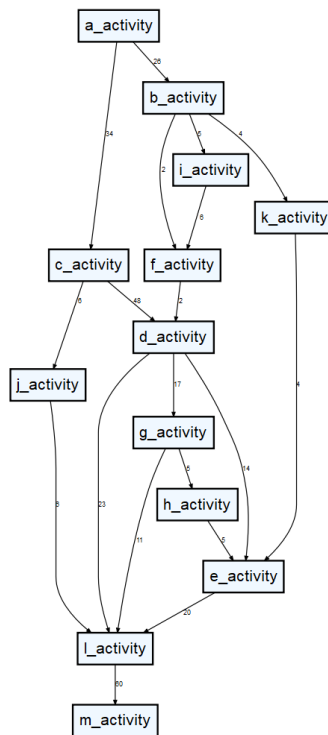


Figure 4.12 – The process model extracted by the classic heuristic miner approach with a manual thresholds set at **50%**. Accordingly, the model presents activities that have a dependency measure higher than 50%. Only 61% of the recorded behaviors respect this threshold.

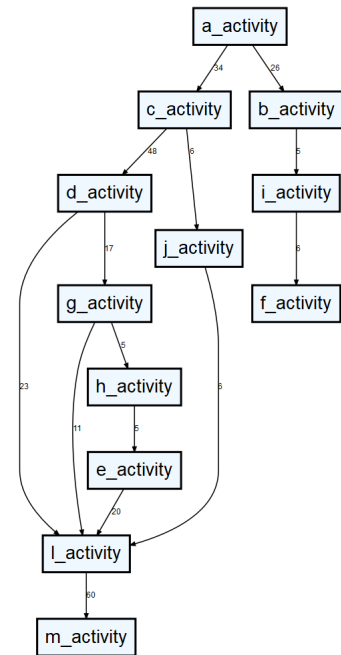


Figure 4.13 – The process model extracted by the classic heuristic miner approach with a manual thresholds set at **80%**. Accordingly, the model presents activities that have a dependency measure higher than 80%. Only 45% of the recorded behaviors respect this threshold.

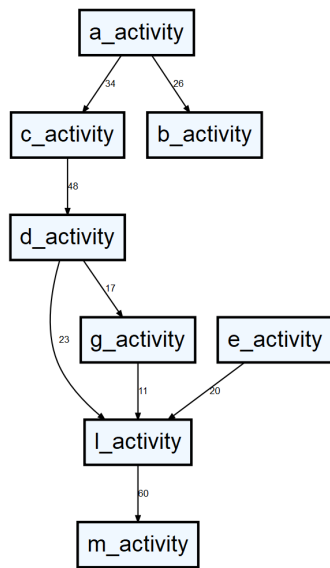


Figure 4.14 - The process model extracted by the classic heuristic miner approach with a manual thresholds set at **90%**. Accordingly, the model presents activities that have a dependency measure higher than 90%. Only 25% of the recorded behaviors respect this threshold.

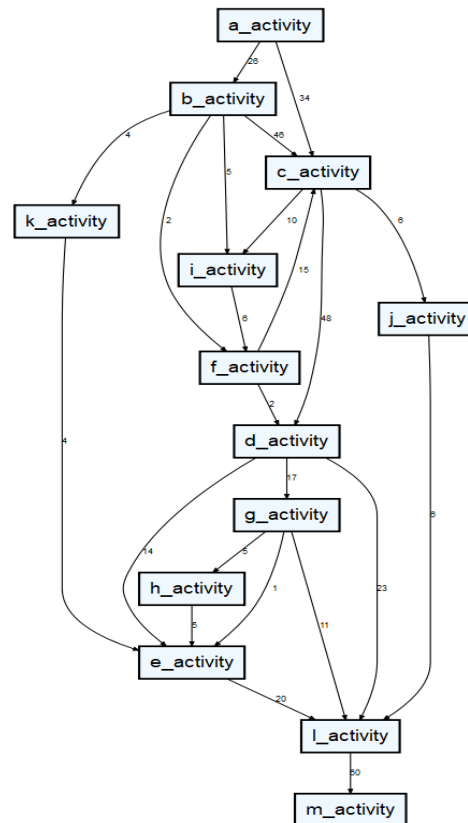


Figure 4.15 - The process model extracted by the classic heuristic miner approach with manual configuration of thresholds. The value for threshold is **set at 0**, therefore, the model shows all the registered behaviors.

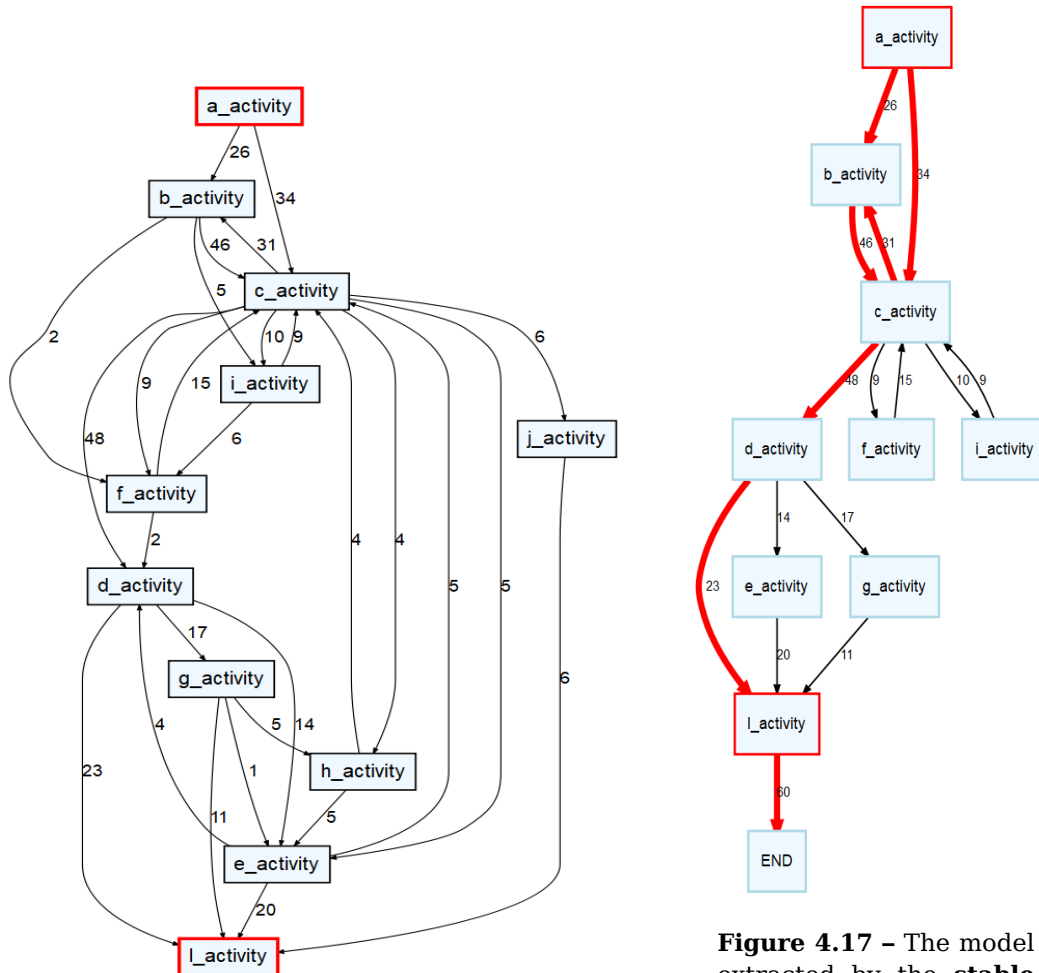


Figure 4.16 – The process model extracted by **stable heuristic miner V1**, automatic detection of thresholds for activities. Red activities have shown high level of variations in their behavior according to the algorithm.

Figure 4.17 – The model extracted by the **stable heuristic miner V2**, automatic detection of thresholds for both activities and edges. Red edges and activities have shown high level of variations in their behavior according to the algorithm.

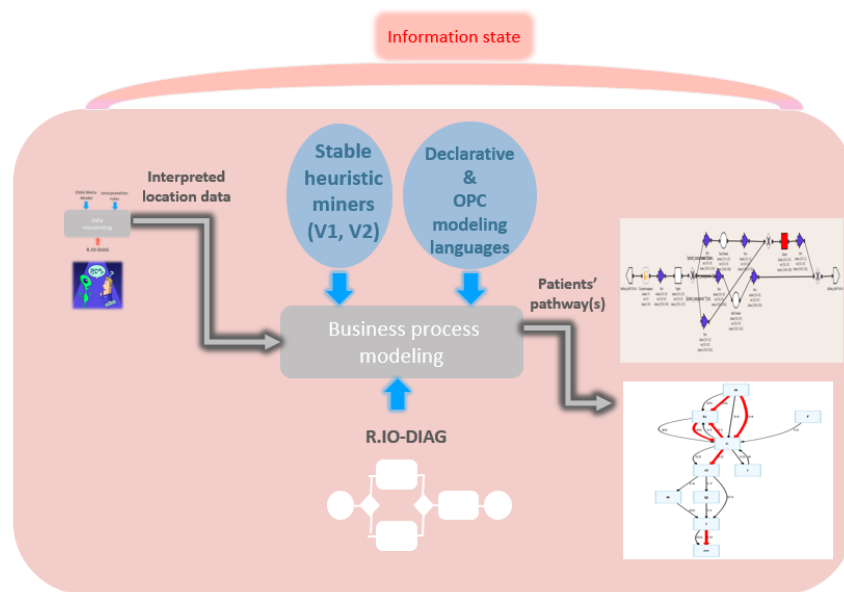


Figure 4.18 – A general summary of what has been presented in this chapter.

4.5 Recap

During the very first hands on experiences for extracting process-like patterns from location data, it has been seen that it is unclear what constitute a set of usual activities for patients to be involved in.

Therefore, using automatic process discovery algorithms were the primary clue. As a consequence of a deep research in the literature of process mining in healthcare, **heuristic mining algorithms** seem to produce reliable results from healthcare data (Rojas et al., 2016). Henceforth, this family of algorithms have been used to extract a **common process-like pattern** for illustrating patient's activities.

Later, an obstacle emerged in selecting a model as a **reference**. These algorithms use manually configurable thresholds for extracting different levels of information from an event log. Previously, this has been mentioned in the literature as an unsolved issue for these algorithms (De Cnudde et al., 2014). Therefore, **it is unreliable to decide on the value of these thresholds for discovering common behaviors**.

As it has been defined in this chapter, a **descriptive reference process model** should present the **stable behaviors** of cases in an environment.

To address such an issue, this chapter presented the statistical stability phenomenon to evaluate the stability within all of the existing relationships among activities and edges.

As a result, two versions of **stable heuristic miner algorithm (V1, V2)** embrace the statistical stability phenomenon and help to improve the mentioned disadvantage for many process discovery algorithms.

The first version of the algorithm only focused on the stability among activities in an event log. In that, not considering the statistical stability of edges became a limitation of this algorithm. Therefore, the second version solved this issue and provided a more holistic approach for extracting the descriptive reference process model.

Thanks to the presented methods, the **reactive thresholds** of stable heuristic miner algorithms automatically detects unstable behaviors in an event log and based on their importance, decides to include them within the **descriptive reference process model** or not.

Within chapter 6, a case study demonstrates the result of classic approach of heuristic miner algorithm in comparison with the new stable heuristic miner algorithms.

To sum up, as shown in figure 4.18, this chapter focused mainly on **how to extract meaningful informative process-like patterns from location event logs**.

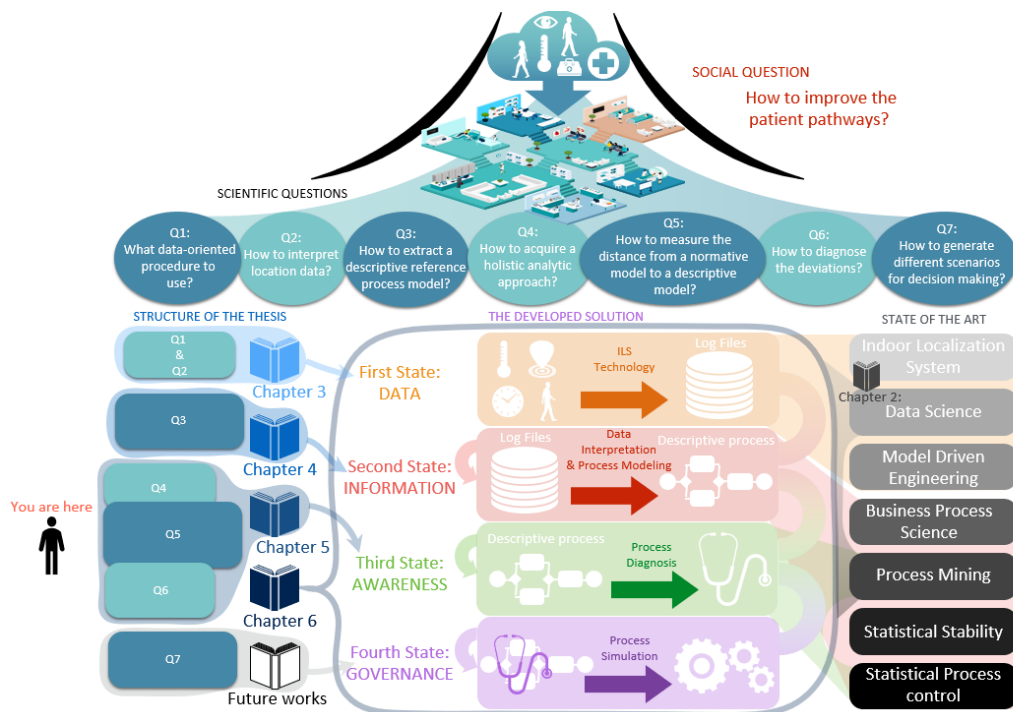
With this objective, **business process modeling** function has been presented within **information state of the DIAG methodology**.

This function is realized in R.IO-DIAG application by three main contributions:

1. Stable Heuristic Miner V1 algorithm
2. Stable Heuristic Miner V2 algorithm

In the end of the information state, one is capable to observe the current status of process and how patients are executing their processes. However, one can not be **aware** about the performance and quality level of the processes. Hence, next chapter will focus mainly on another problem which is:

The ambiguity on how to measure **the performance and quality ratios of a process** within BPM literature for healthcare, and for process mining applications in general. Moreover, the concept of **automatic business process diagnostic** will be investigated.



5

Awareness: analyzing and diagnosing patient pathways

5.1	Introduction	122
5.2	Business Process Analyzing Function	125
5.2.1	Pre-processing the information	125
5.2.1.1	Extracting a data table from the observed behaviors	126
5.2.1.2	Sort the data by ID of cases	126
5.2.1.3	Add the duration of activities	126
5.2.1.4	Calculate the total duration and distances of processes	127
5.2.1.5	Generate the second data table for sampling the duration of processes	127
5.2.1.6	Generate the third data table for sampling the distances of processes	127
5.2.2	Statistical process control (SPC) application to support the enhancement activity of process mining	129
5.2.2.1	Control charts	129
5.2.2.2	Process Capability Ratio (C_p)	130
5.2.3	An example: Applying control charts and process capability ratio analyses	132
5.3	Business Process Diagnosing Function	137
5.3.1	Automatic diagnosing approach 1: Miniscule Movements of Processes (MMP)	137
5.3.1.1	Miniscule Movements: An approach to discover planets	137
5.3.1.2	MMP method: Definitions and Hypothesis	139
5.3.1.3	Logic of MMP method	141
5.3.1.4	Deducing GM's (generated models)- DIAG meta-model is revisited	142
5.3.1.5	ProDIST algorithm: a novel method for measuring the distance between two process models	147
5.3.2	Automatic diagnosing approach 2: DIAG method	152
5.3.2.1	An example for the second automatic diagnosing approach	155
5.4	Recap	158
5.4.1	SWOT analysis of the chapter	158

"I have self-doubt. I have insecurity. I have fear of failure. I have nights when I show up at the arena and I'm like, 'My back hurts, my feet hurt, my knees hurt. I don't have it. I just want to chill.' We all have self-doubt. You don't deny it, but you also don't capitulate to it. You embrace it".

Kobe Bryant

5.1 Introduction

As it has been mentioned in the first chapter, one of the business challenges for healthcare professionals is to offer processes which are perceived from patient's point of view as efficient processes. Patients have difficulties to anticipate the outcomes of the non-medical healthcare processes.

Non-medically speaking, when a person goes to a healthcare organization, he or she can not expect (i) *how long will be the waiting time*, and (ii) *how much will be the walking distance for finding a certain office*. The responses to such questions, are directly linked to the quality and performance of patient pathways.

These unpredictability characteristic of healthcare processes is a non-trivial criteria. Such uncertainty about the process outcomes affects directly the way patients conceive a process. A process is efficient, if it eliminates the **varying outcomes**, and it is perceived as a **stable process**. This is the main definition of **quality** (Montgomery, 2019).

In light of this, chapter 5 presents two functions of the DIAG methodology. As shown in figure 5.1, the **Business process analyzing** function is positioned here for analyzing the stability of patient pathways outcome. The **Business process diagnosing** function is in charge of automatically detecting the causes of deviations in processes.

This chapter answers three scientific questions of this dissertation.

- How to evaluate the performance of patient pathways and detect the inefficiencies?
- How to measure the distance of a reference behavior from an observed one?
- How to precisely diagnose quantitatively and qualitatively the deviations?

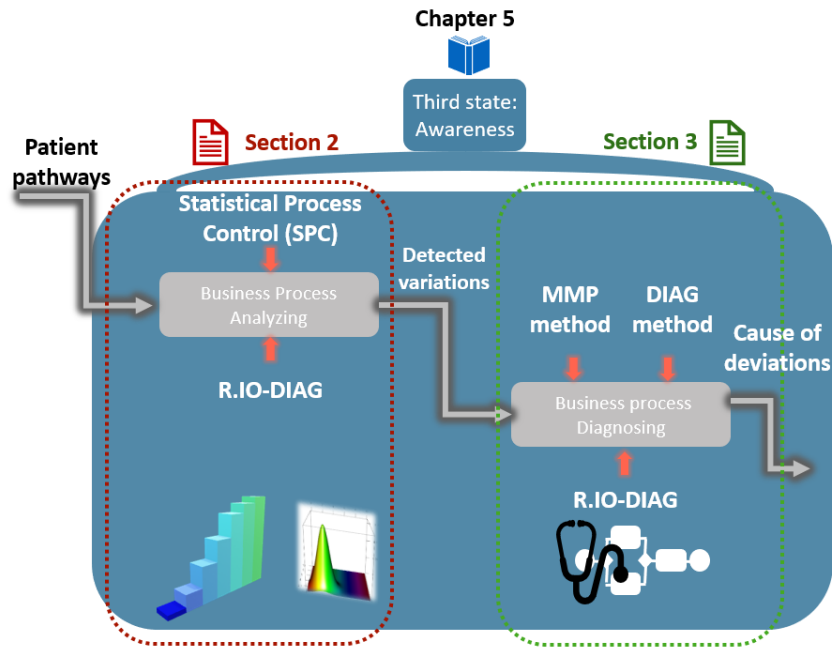


Figure 5.1 – An overall view on top of the third state of the DIAG methodology, Awareness.

Accordingly, a process improvement action is achievable by detecting and removing the cause of process fluctuations.

This chapter includes two parts. First, the business process analyzing function with the application of SPC (statistical process control) will be presented. This approach targets the enhancement activity of process mining.

The second part consists of two approaches for automatically diagnosing deviations of processes. Figure 5.2 illustrate how this chapter is structured to present four main contributions:

1. The application of SPC for evaluating quantitatively the performance of patient pathways.
2. The ProDIST algorithm for measuring the distance of processes from each other.
3. The Miniscule Movements of Processes (MMP) method for automatic diagnosing the patient pathway of individual cases.
4. The DIAG method which contains an algorithm for automatic diagnosing of processes with the addition of domain knowledge into stable heuristic miner algorithms. This method is used to diagnose patient pathways by considering multiple cases (as a reminder: each case represents a patient).

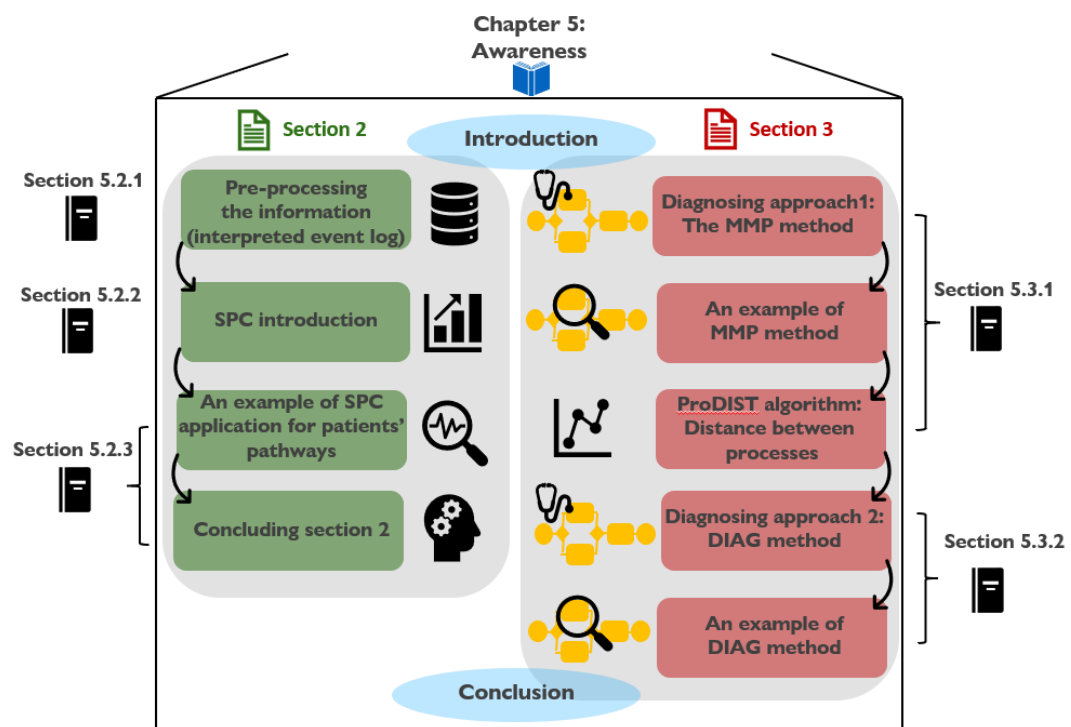


Figure 5.2 – This figure shows how this chapter is structured.

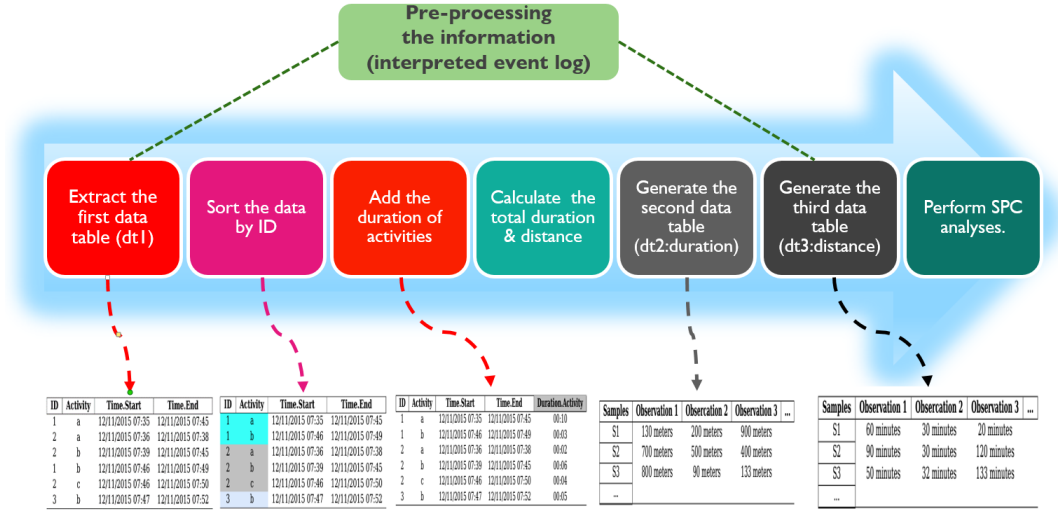


Figure 5.3 – The defined steps to prepare the data.

5.2 Business Process Analyzing Function

It should be kept in mind that, the performance and quality of business processes is heavily relying on the **level of variations in process outcomes**. The presented method in this section (business process analyzing function) proposes several quantitative methods to evaluate the **process variations**.

Accordingly, it would be feasible to assess the quality of patient pathways.

Based on the presented literature review in this dissertation and the work done in (Salimifard et al., 2001); performance analysis of business processes is relying heavily on quantitative analyses and unfortunately it did not receive enough attention.

This function within the DIAG methodology has an important role corresponding to the existing literature gaps.

This statement is valid also for the **enhancement notation (activity)** of process mining paradigm.

Consequently, several quantitative analyses are developed for the **business process analyzing function**:

1. Application of \bar{x} and R **control-charts** to detect inefficiencies in process outcomes.
2. **Process capability ratio (C_p)** analysis to measure the performance of processes.

To carry on with the application of these methods (\bar{x} and R control-charts, and C_p), first, it is required to perform certain data engineering actions. These actions are grouped within the “pre-processing the information” section.

5.2.1 Pre-processing the information

Before applying the quantitative analyses, the event logs should go through several pre-processing actions.

As shown in figure 5.3, these actions are defined here as:

ID	Activity	Time.Start	Time.End
1	a	12/11/2015 07:35	12/11/2015 07:45
2	a	12/11/2015 07:36	12/11/2015 07:38
2	b	12/11/2015 07:39	12/11/2015 07:45
1	b	12/11/2015 07:46	12/11/2015 07:49
2	c	12/11/2015 07:46	12/11/2015 07:50
3	b	12/11/2015 07:47	12/11/2015 07:52

Table 5.1 – An example to illustrate the first extracted data table (dt1).

ID	Activity	Time.Start	Time.End
1	a	12/11/2015 07:35	12/11/2015 07:45
1	b	12/11/2015 07:46	12/11/2015 07:49
2	a	12/11/2015 07:36	12/11/2015 07:38
2	b	12/11/2015 07:39	12/11/2015 07:45
2	c	12/11/2015 07:46	12/11/2015 07:50
3	b	12/11/2015 07:47	12/11/2015 07:52

Table 5.2 – An example to show the sorted data table 1 (dt1) by ID of each case.

1. Extracting a data table from the observed behaviors.
2. Sorting the data by ID of cases and the corresponding activities.
3. Adding the duration of activities.
4. Calculating the total duration and distance of processes.
5. Generate the second data table to classify different samples and the corresponding observations (each observation = duration of a process).
6. Generating the third data table to classify different samples and the corresponding observations (each observation = distance of a process).

These steps are illustrated by the following running example (an event log) in table 5.1, and algorithm 11 presents the formal sequence of pre-processing actions.

5.2.1.1 Extracting a data table from the observed behaviors

At first, the event log should be recorded as a data-table. Table 5.1 illustrates the first data-table (dt1), which is the primary event log. One of the challenges here is to extract the **sequence of activities for a particular ID**. Such a sequence would present one patient pathway.

The next action will sort this data table as required.

5.2.1.2 Sort the data by ID of cases

This function is added to get the individual processes. The objective is to have an organized data table which eases the calculation for the distance and duration of processes. This function results in to a data frame similar to table 5.2 .

5.2.1.3 Add the duration of activities

Next step is to calculate the duration of each activity. Table 5.3 shows the duration of each activity. Next, the duration and distance of each patient pathways should be calculated.

ID	Activity	Time.Start	Time.End	Duration.Activity
1	a	12/11/2015 07:35	12/11/2015 07:45	00:10
1	b	12/11/2015 07:46	12/11/2015 07:49	00:03
2	a	12/11/2015 07:36	12/11/2015 07:38	00:02
2	b	12/11/2015 07:39	12/11/2015 07:45	00:06
2	c	12/11/2015 07:46	12/11/2015 07:50	00:04
3	b	12/11/2015 07:47	12/11/2015 07:52	00:05

Table 5.3 – dt1 with the added duration of activities.

Samples	Observation 1	Observation 2	Observation 3	...
S1	60 minutes	30 minutes	20 minutes	
S2	90 minutes	30 minutes	120 minutes	
S3	50 minutes	32 minutes	133 minutes	
...				

Table 5.4 – dt2 representing the data table for sampling the duration of processes.

5.2.1.4 Calculate the total duration and distances of processes

For this cause, simply a *sum* function is used by considering the duration of activities in each sequence.

The total distance is calculated by considering the *coordination*(x, y) of consecutive events.

If $e_1 = Event.1(x_1, y_1) \& \quad e_2 = Event.2(x_2, y_2)$ then:

$$d_{(e_1, e_2)} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (5.1)$$

5.2.1.5 Generate the second data table for sampling the duration of processes

In order to better perform the statistical process control analyses, it is required to create **samples** from the input data. Therefore, *dt2* is generated to create random samples with unique sizes from the data. Table 5.4 shows a simple example of what a *dt2* could look like.

To apply control-chart and C_p analyses on the sorted and refined event log, it is required to to perform two **sampling** actions.

5.2.1.6 Generate the third data table for sampling the distances of processes

Similar to the previous function (2.5), a data-table called as *dt3* will be extracted. This data-table presents the sampling function for the distance of processes. The sampling will be random on processes with the same profile (date and objective) and all the samples have a unique size. Table 5.5 shows an example of the sampled data about the distance of processes.

The following algorithm embraces these mentioned tasks to prepare the **information** in event logs for the next statistical process control analyses.

Next section will carry on with the task of performing statistical process control analyses.

Samples	Observation 1	Observation 2	Observation 3	...
S1	130 meters	200 meters	900 meters	
S2	700 meters	500 meters	400 meters	
S3	800 meters	90 meters	133 meters	
...				

Table 5.5 – An example of what *dt3* could look like.

Algorithm 11 Pre-processing the information prior to SPC analyses

```

1: procedure Data pre-processing
2:   Input Activity.Set = Structure
3:     < Name >: string
4:     < Id >: integer
5:     < Time >: date
6:     < Coordinate >: float
7:
8:   dt1 = data.table(ID = Activity.Set[ID], Activity = Activity.Set[Name], TimeStart
    = Activity.Set[Time.St], Time.End = Activity.Set[Time.End] ,header = TRUE);
9:                                     ▷ Use shift function and sort
10:
11:   sequences = dt1[,dt1[FirstActivity = dt1[Activity], SecondActivity=
    shift(dt1[Activity], type= "lead")], dt1[ID]] ;
12:                                     ▷ duration of activities
13:   for i in dt1[i] do
14:     ColumnBind(Duration.Activity = dt1[Time.End[i]]- dt1[Time.Start[i]]);
15:   return dt1
16: end for
17:                                     ▷ sampling duration and distance
18:   for i in ID[i] do
19:     Duration.Process = sum(Activity.Duration[i])
20:   end for
21: dt1[ColumnBind(Activity.Set[Coordinate])]
22:   for i in dt1[Coordinate[i]] do distance = equation5.1;
23:   end for
24:   dt2 = sample(Duration.Process, size = n);
25:   dt3 = sample(distance, size = n);
26:   apply SPC analyses
27: end procedure

```

5.2.2 Statistical process control (SPC) application to support the enhancement activity of process mining

To ensure that a service is designed to meet or surpass clients expectations, it should be delivered by a process that is stable and its outcomes are repetitive (Montgomery, 2019). In this context, a process that has high quality level should be able to operate with the least amount of variability around the specification (target values) of the **quality characteristics**. This statement is endorsed by the definition of quality in (Montgomery, 2019). In addition, DIAG meta-model aimed to respect this important criteria.

SPC is a powerful collection of problem-solving tools useful in achieving process stability and improving capability through the reduction of variations in the process.

SPC has already made its way into the healthcare sector (Thor et al., 2007). However, it has been used mainly for analyzing biological experiments but not in a sense of analyzing patient pathways.

As mentioned, **control-charts** and C_p analyses will be applied in this research work. The following will present their application for evaluating the performance and quality of patient pathways.

5.2.2.1 Control charts

The Shewhart control charts are one of the most advanced techniques of SPC. As presented earlier in chapter 4, a typical control chart has three indicators which are known as **center-line (CL)**, **upper control limit (UCL)**, and **lower control limit (LCL)**.

These border lines are presented horizontally in a control chart. They are mainly specifying the limits for the stable outcomes of a process. As long as all of the points of the samples are between LCL and UCL, no action is necessary. But, if a point falls beyond those limits, it could be inferred that the process is out of control due to the high level of variations. Therefore, some inspections on different aspects of the process are required. There are several types of control charts, such as **\bar{x} -chart**, **R -chart (range chart)**, **S -chart**, **P -chart**, **individual charts**, and **C -charts** (Montgomery, 2019).

The application of each of these charts depends on the types of data and analysis that one could require.

In this chapter, for **analyzing the outcome of business processes**, **\bar{x}** , and **R -charts are applied**. The used numerical values are **time** and **distance** for measuring the **stability or in-control state** of processes.

By applying these methods on the **distance and duration samples**, it is feasible to provide a **set of knowledge** about efficient and inefficient patient pathways.

In following the mathematical principles for constructing the control limits will be presented.

Let $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_m$, be the average of each sample, then the process average is the grand average ($\bar{\bar{x}}$)

$$\bar{\bar{x}} = \frac{\bar{x}_1 + \bar{x}_2 + \dots + \bar{x}_m}{m} \quad (5.2)$$

If the range of each sample equals to R then:

$$R = x_{max} - x_{min} \quad (5.3)$$

Now let R_1, R_2, \dots, R_m be the ranges of each sample then the average range is:

$$\bar{R} = \frac{R_1 + R_2 + \dots + R_m}{m} \quad (5.4)$$

Process variability could be monitored by plotting values of the sample range R on a control chart. The control limits for R -chart are as follows:

$$\begin{aligned} UCL &= \bar{R}D_4 \\ CL &= \bar{R} \\ LCL &= \bar{R}D_3 \end{aligned} \quad (5.5)$$

The constants D_3 and D_4 in (5.5) are used based on different values of sample sizes (n).

The control limits for \bar{x} -chart are as follows:

$$\begin{aligned} UCL &= \bar{\bar{x}} + A_2\bar{R} \\ CL &= \bar{\bar{x}} \\ LCL &= \bar{\bar{x}} - A_2\bar{R} \end{aligned} \quad (5.6)$$

The value of A_2 in (5.6) changes based on the size of the samples. A_2 is a customary constant which is used for construction of the \bar{x} -chart.

The values for these customary constants A_2 , D_3 and D_4 are presented in most of mathematical and statistics references (Kuei, 2004).

section 5.2.3 presents the application of these methods by an example.

The results of this application has been published previously in (Araghi et al., 2019; Araghi et al., 2018b). Previous to this research work, process mining applications were not able to propose a method to measure the **quality** and **performance** of the process outcomes.

Such an approach is positioned within the **enhancement activity** of process mining discipline.

Notice

It must be noted that the application of control-charts in this chapter is different with the one presented in chapter 4 for discovering processes. Both applications seek for the statistical stability, but in chapter 4, the data samples did not have **unique sizes**.

Comparably, here, the sample sizes have a unique size. Therefore, the **mathematical approach is different**.

5.2.2.2 Process Capability Ratio (C_p)

Another method to analyze the performance of processes is in terms of process capability ratio (PCR) or C_p (Boyles, 1991). This method helps to understand how well a process is working in accordance with certain result specifications.

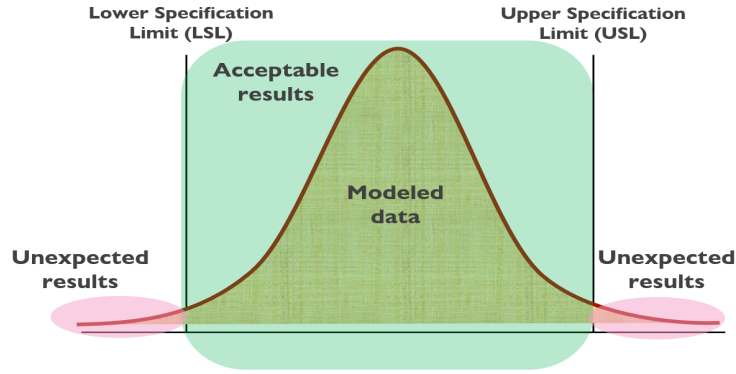


Figure 5.4 – This figure provides an illustration of the process capability ratio analysis.

As presented in figure 5.4, PCR is a statistical method for making a comparison between the output of a process and the specifications limits of the process. A process which all of its outcomes fall between the specification limits, is considered as a capable process. For example, if hospitals define certain specifications as the expected length of stay for patient pathways; C_p ratio helps them to evaluate their performance.

PCR analysis could be defined by adjusting two new limits as the **Upper Specification Limit (USL)** and the **Lower Specification Limit (LSL)**. These limits are specifications relevant to the quality characteristics that one desires to analyze (such as reliability of a process).

In this research work, the USL and LSL could be defined **manually** by the healthcare experts, or **automatically** by calculating and analyzing the distribution of the gathered data.

Equation (5.7) shows the mathematical expressions to calculate C_p . Where σ is the standard deviation of samples (Booker et al., 2001). Equation (5.8) shows how to calculate the **specification limit** regarding the distribution of the data.

$$C_p = \frac{USL - LSL}{6\sigma} \quad (5.7)$$

$$\begin{aligned} USL &= \bar{x} + 3\sigma \\ LSL &= \bar{x} - 3\sigma \end{aligned} \quad (5.8)$$

C_p could have three states, which help experts to analyze the capability of the As-Is processes:

- If $C_p < 1$; it means that process is using up more than 100 % of the tolerance band which means the process is not capable to provide the desired outcome continuously.
- If $C_p = 1$; it means that process is using 100% of its tolerance band. This implies that process may provide some undesirable outcomes, but statistically is predictable and capable of satisfying the current specification defined by the organization.
- If $C_p > 1$; the process is using much less than 100 % of its tolerance band. As a result, relatively few undesirable outcomes could be produced by the process.

These analyses can be seen concretely by the description of an example in the next section.

The figure 5.5, 5.6, and 5.7 show how the data will be modeled in accordance to the process capability analysis.

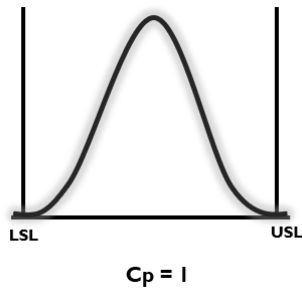


Figure 5.5 – The modeled data when $C_p = 1$

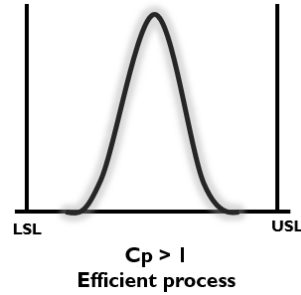


Figure 5.6 – The modeled data when $C_p > 1$

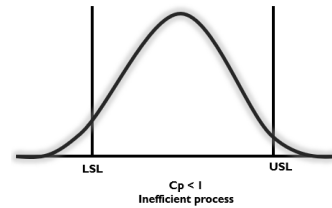


Figure 5.7 – The modeled data when $C_p < 1$.

5.2.3 An example: Applying control charts and process capability ratio analyses

Due to the unpredictability features of patients' pathways outcomes, the objective here is to monitor where are the variations in these processes.

Accordingly, a patient pathway is considered **efficient** if it has **stable outcomes**.

Evidently, the stable outcomes is related to low-level of variations which is in-line with the definition of high-quality outcomes.

This section provides an illustration of applying these methods for measuring the quality and performance of patient pathways in a simulated environment.

The presented experiment here took place within a 10 days period, and the data about 150 individual cases (patients) have been gathered. These individual cases have been divided into 10 samples with a unique size of 15.

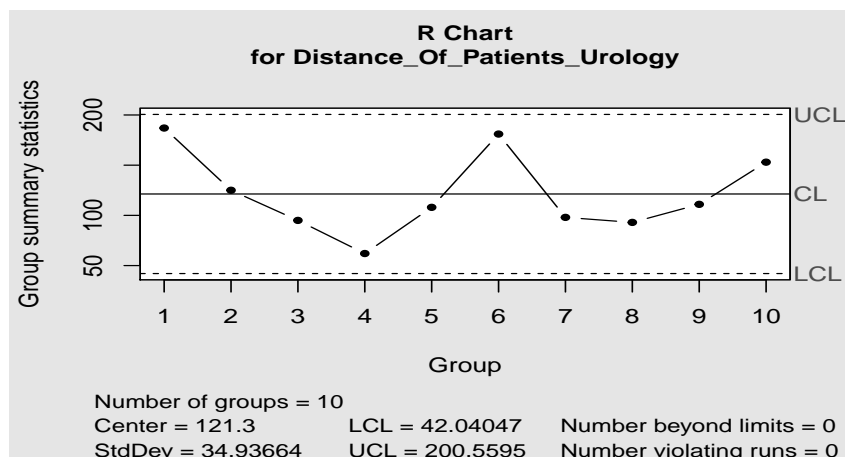


Figure 5.8 – Analyzing the variations of pathways' length

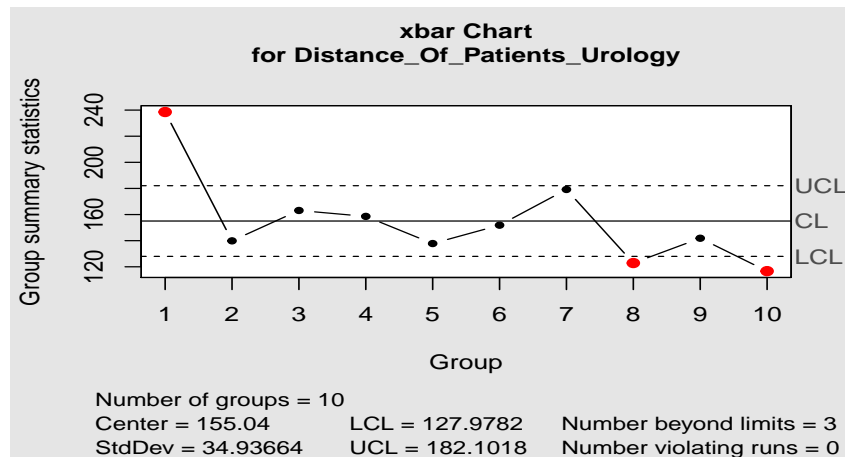


Figure 5.9 – The \bar{x} -chart to monitoring the stability of patients' pathways

It is ideal to begin the quantitative analyses by the R -chart. This is due to the fact that \bar{x} -chart control limits are depending to the process variability. Unless the process variability is in-control, \bar{x} -chart limits will not have accurate results and the process is already **out-of-control** and it is not producing quality results.

The data coming from location event logs gives the advantage of evaluating the quality of processes by measuring the walking-distance of patients.

In similar research works, experts did not address **distance as a metric to evaluate patient pathways** (Araghi et al., 2018b). However, it has been seen that this metric can be correlated with the efficiency of processes.

For instance, if one considers that a normal process outcome is for patients to not get lost inside a facility; a patient who is lost inside a hospital is determined as a deviation. Such a deviation could be detected by monitoring the walking distance.

The R -chart helps to ensure the stability of the extracted data. The CL is the average of all the subgroups' ranges. The other control limits are set by a customary distance of 3σ (standard deviation) above and below the center line. These thresholds define the limits for expected variations in the subgroups ranges. Figure 5.8 shows R -chart for analyzing the range of distances for patients' pathways.

Based on the stability in ranges, it is relevant to construct the \bar{x} -chart presented in figure 5.9. This figure presents the instability of the average traveled distance by patients. The red points in the \bar{x} -chart indicate that within three days of the experiment, there were some assignable causes that affect the distance of pathways for the patients.

These variations could be caused by an increase in the number of admitted patients on those days, the department had the lack of resources to perform medical examination for all the patients.

The length of a patient pathway is a practical indicator for the capability of a hospital in providing efficient services. The reason could be seen as the effect on the length of stay in the hospital. Also, it influences the efficiency of providing emergency treatments to the exact location of a patient with a critical status.

Thus, this research work seeks different means to analyze the capability of processes based on the defined CTQ (Critical To Quality) specifications; "reliability". These specifications (USL, target, LSL) either are defined by the health professionals or could be

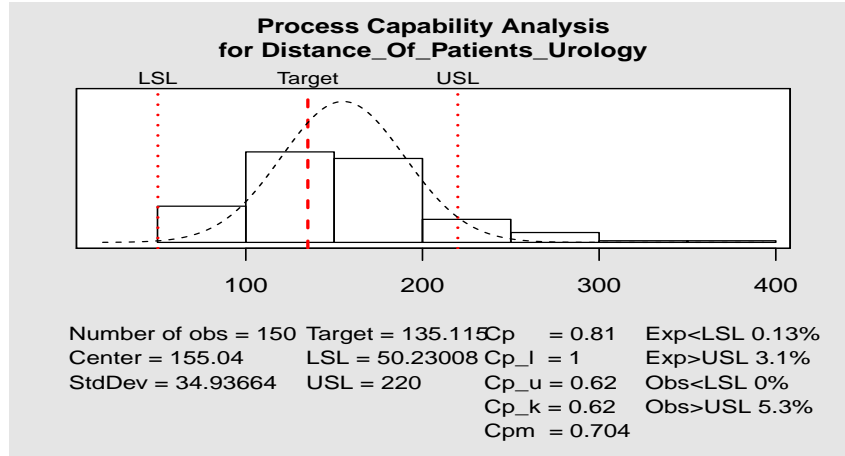


Figure 5.10 – Evaluation of the process capability by the length of patients’ pathways. Note that the average(μ) is not well-centered.

adapted from the way patients are moving in the hospital. Figure 5.10 presents the results on process capability ratio (C_p) of the experiment.

There are several points that should be inferred from *PCR* analysis; first of all, this chart consists of two specification limits that show the possible margins for the length of pathways.

As it has been shown in figure 5.10, there are processes that have the distance higher than USL and lower than LSL. These processes are not satisfying the targeted specifications.

The experts have defined an USL of 220 meters provisionally. The LSL has been adapted **automatically** from the current state of processes with the defined formula in 5.8. The C_p is **less than 1** which shows that processes are being executed in an inefficient way and they are not meeting the expectations. As shown in figure 5.10, 5.3 % of patients are walking more than the upper specification limit.

More details on process capability ratios

The C_p ratio does not consider a situation where a process average is not well-centered between USL and LSL. A process can have a C_p greater than 1, but it will be considered an incapable process if the process average is not centered.

This phenomenon can be seen in figure 5.10.

In this case, it is a better practice to use the C_{pk} ratio (Montgomery, 2019). This ratio considers the behavior of the data close to the two specification limits. It calculates as presented in following:

$$C_{pk} = \min(C_{pu}, C_{pl}) \quad (5.9)$$

$$C_{pu} = \frac{USL - \mu}{3\sigma} \quad (5.10)$$

$$C_{pl} = \frac{\mu - LSL}{3\sigma} \quad (5.11)$$

Note that in the presented example, C_{Pl} and C_{Pu} demonstrate the performance of processes near to the lower specification limit and the upper specification limit.

Generally speaking, if $C_p = C_{pk}$ then the process is well centered, otherwise the process is off center.

On the other hand, C_{pm} measure could be useful if we want to use the average value as the target to reach. C_{pm} could be applied to monitor the difference between the target value and the average of the results. In order to find a center point a value as **target** is defined:

$$Target = T = \frac{1}{2}(USL + LSL) \quad (5.12)$$

Accordingly, C_{pm} ratio is introduced to address the difficulty to apply process capability analysis when the data is not well-centered.

$$C_{pm} = \frac{USL - LSL}{6\sqrt{\sigma^2 + (\mu - T)^2}} \quad (5.13)$$

For instance, the target in this example has been identified as 135.115 meters. However, the average length of pathways is different ($\mu = 155.04$).

Therefore, C_{pm} can have better applicability here.

It is not encouraged to apply *PCR* analyses after indicating that a process is out-of-control (c.f. figure 5.9). Because, if a process is out of control then, obviously it is not capable to meet the expectations. However, this has been presented here as an example to suggest these analyses for process mining applications.

The same analysis are applicable and have been done for the duration of processes which their explanations are beyond the scope of this chapter.

Relevant to the "enhancement" notation of process mining paradigm, the presented methods can provide applicable means in order to detect the process variations from the outcomes of the processes.

Thanks to the presented methods and monitoring process variations, it is now feasible to detect:

- Relevant to duration of processes, which group of patients had effective and ineffective processes.
- Considering the distance of pathways as the numerical value, which sample of patients experienced effective and ineffective processes.
- It is also possible for the domain experts to visualize, how well their expectations from the process were met.

It should be borne in mind that, **detecting the variations that cause a process inefficiency is indeed in line with the definition of "diagnostic"**. Such functionality did not exist in the previous process mining applications and plugins.

After detecting the inefficient patient pathways, next section focuses on introducing methods for **detecting and diagnosing the causes** of such inefficiencies.



Note on DIAG meta-model: the CTQ class

An efficient process or a reliable process is a characteristic of an organization which can be embraced by DIAG meta-model (c.f. figure 5.20). Within the **objective package** a class as **critical to quality (CTQ)** is defined. This class is a inherited from the **characteristic** class.

Therefore, in order to evaluate this characteristic of the organization, one need to define a **KPI** and to use certain **numerical values**, such as time or distance to measure the reliability of the certain **component** of the organization. A **process** can be defined as a inherited object of these components.

It can be inferred from the presented example in this chapter that the mentioned patient pathway was not a **reliable** process.

5.3 Business Process Diagnosing Function

“A state of statistical control is not a natural state for most processes. It is instead an achievement, arrived at by elimination, one by one, by determined effort, of special causes of excessive variation.”

Edwards Deming

This part presents the business process diagnosing function (c.f. figure 5.1). This function proposes two methods to diagnose automatically the structural deviations in business processes. These methods are **Miniscule Movements of Processes (MMP)**, and **DIAG** methods.

5.3.1 Automatic diagnosing approach 1: Miniscule Movements of Processes (MMP)

The purpose for this method is **to diagnose automatically the cause of deviations in the structure of process models**. As mentioned in chapter 1, the **structural problems in processes** cause costly affects for hospitals.

5.3.1.1 Miniscule Movements: An approach to discover planets

This research work considered an inspiring **planet discovery approach from NASA** to **detect the cause of deviations** in process models.

NASA has introduced 5 main ways to discover the existence of a planet in our solar system. These approaches are **watching for wobble, searching for shadows, taking pictures, light in a gravity lens, and Miniscule Movements¹**.

The **Miniscule Movements** became the inspiration here for the developed **automatic business process diagnosing method**.

Let's elaborate on the miniscule movement approach. Figure 5.11 shows an overall image of our beautiful solar system. All planets and their moons are orbiting around the sun and the distances of these elements are already known. However, while these planets are orbiting, their distance from each other are monitored constantly.

Based on the hypothesis of the miniscule movement approach, if the distance between two objects changes regularly, then it can be said that a massive object is located between the two monitored objects. This is shown in figure 5.12.

This research work, tries to propose an approach which is inspired by such a phenomenon in the universe, and it aims to elaborate on it to answer a question within the literature of BPM and process mining on how to **diagnose automatically** the deviating behaviors in a process model.

This method is called as **Miniscule Movements of Processes (MMP)**.

¹<https://exoplanets.nasa.gov/alien-worlds/ways-to-find-a-planet/>

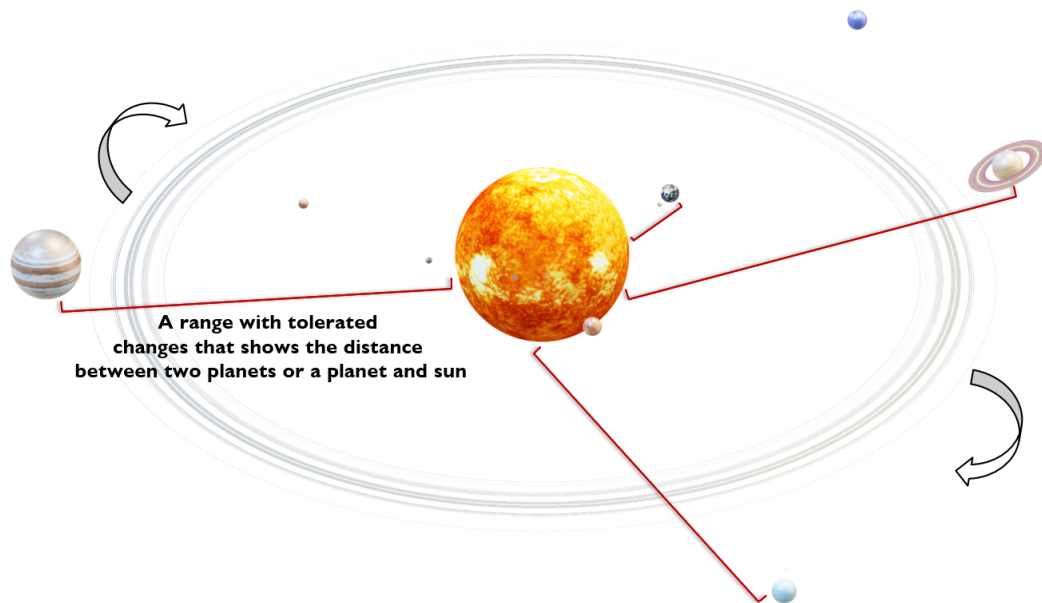


Figure 5.11 – Showing the solar system and the approximated and already measured distances among different elements.

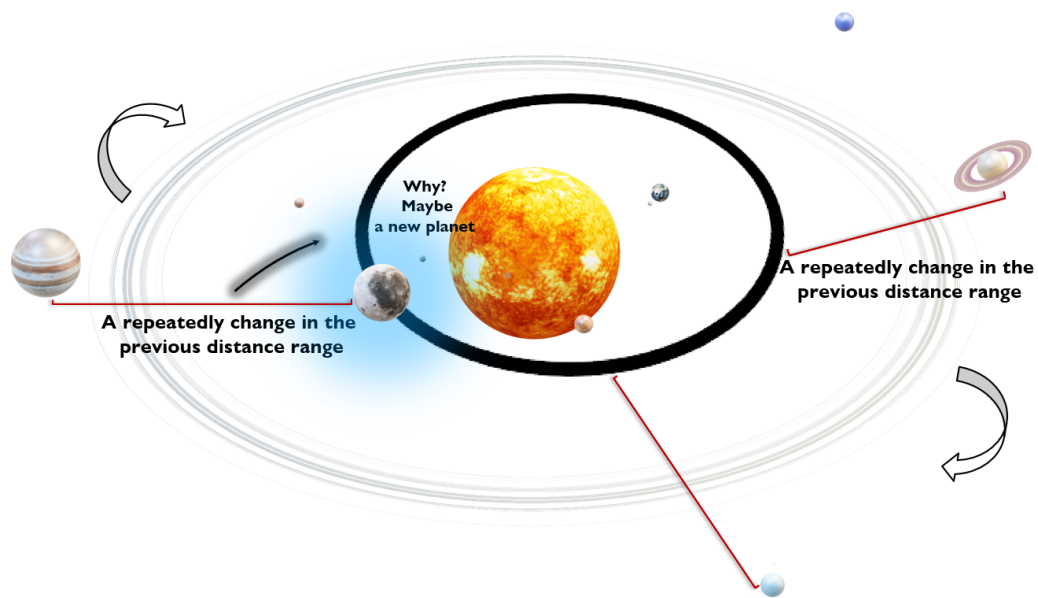


Figure 5.12 – What happens when the previous defined distance range changes? A hypothesis is made that a significant massive object is appeared between the two already discovered objects.

5.3.1.2 MMP method: Definitions and Hypothesis

The **MMP** method is not exactly following the same scientific planet discovery approach. But, it makes an inspired hypothesis.



MMP hypothesis

Domain experts have certain expectations of how a process model should be structured. However, in reality, when they run a process, the process deviates from their expectations. Now, the challenge is to diagnose these deviations.

Have a look at figure 5.13. Imagine the system is able to get a **process model as the reference behavior**; we use RM here to call this model.

RM can be either a normative model designed by the domain experts. This model is indeed sequential.

After extracting the real execution of processes by process discovery methods, it is possible to see certain deviations in descriptive models (DM). Evidently, these deviations cause a **distance** between the RM and DM s.

Until here, figure 5.13 presented the current state of process mining literature for analyzing processes.

Now, consider figure 5.14. By injecting already defined **potential assignable causes (PAC)** into the reference behavior, it is possible to generate deviated processes that we know why are deviating. We call these deviating models as GM . Therefore, it is possible to define GM s as:

$$RM + PAC \Rightarrow GM \quad (5.14)$$

Now, if we are able to find the *minimum* distance between the **descriptive models (DM)** that are extracted from an event log and the GM 's then, there are some (good) chances that the injected PAC into the GM may be the same cause of the deviations of the DM .

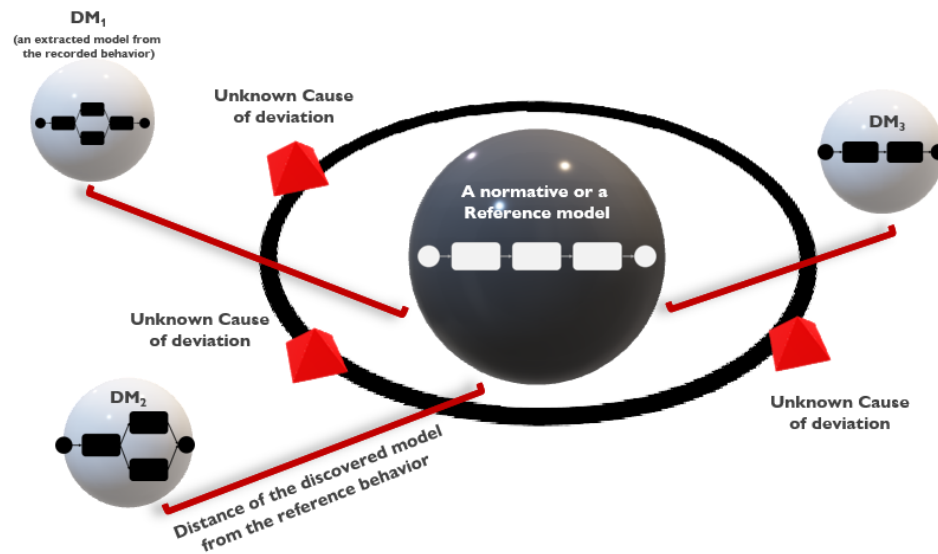


Figure 5.13 – It shows the orientation of the descriptive process models (DM) around a normative or a reference model. Note that the DM's are extracted from existing information and it is possible for them to deviate from the reference behavior due to existence of the unknown causes.

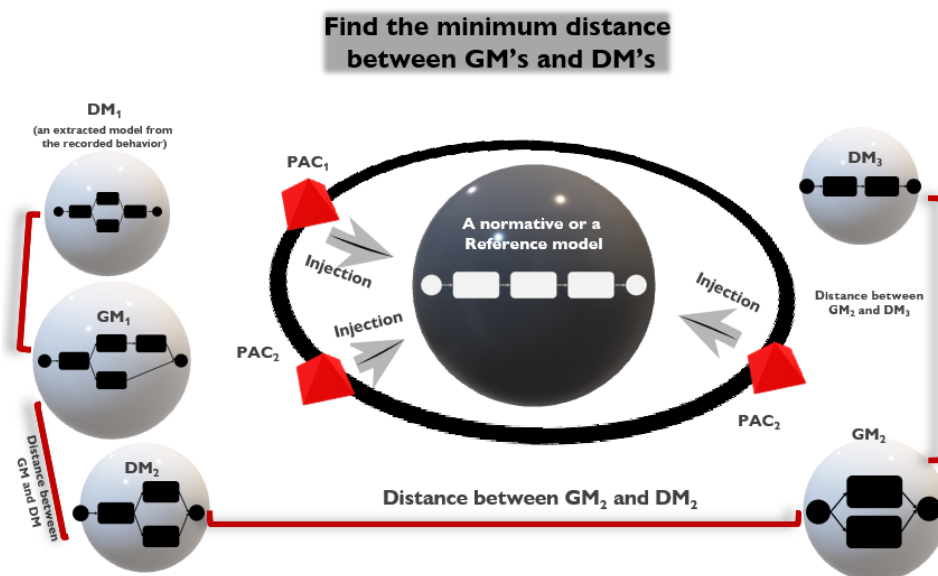


Figure 5.14 – It illustrates the hypothesis of MMP method. This method tries to inject some potential assignable causes (PAC) into the reference behavior to generate several processes (GM). If the minimum distance between a GM and a DM is found, it is possible to indicate they are deviated by the same cause.

Name	Stands for	How to obtain it?
RM	Reference Model (or also called as normative)	Designed by the domain experts
DM	Descriptive Model	Discovered by the discovery methods
PAC	Potential Assignable Causes	Defined by the domain experts
GM	Generated Models	RM + PAC = GM injecting PAC's into a reference (normative) model

Table 5.6 – This table provides certain information about the key definitions within MMP method.

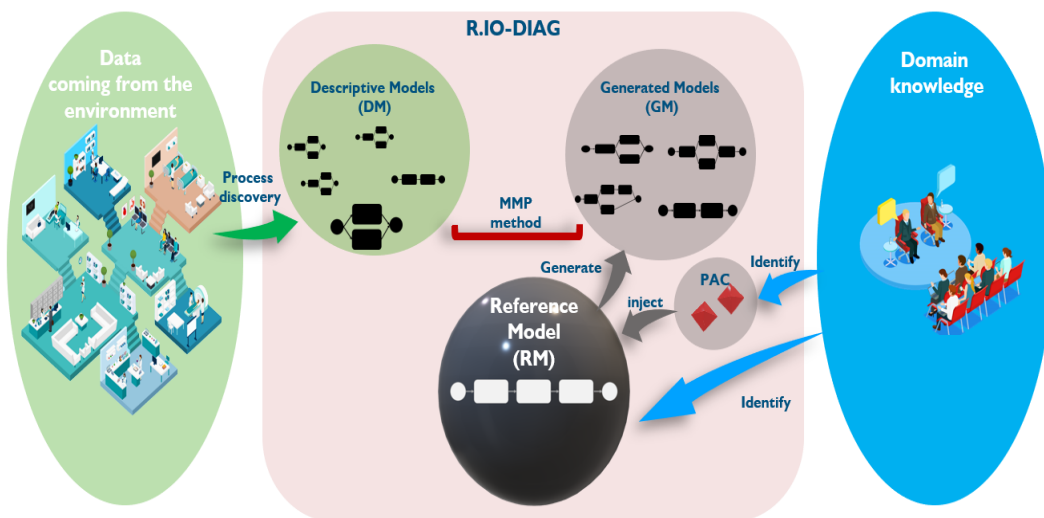


Figure 5.15 – This illustrates how both domain knowledge and data from information systems can be used by the MMP method to diagnose the distance between the modeled and observed behavior.

Table 5.6 provides a summary of the used definitions in this method.

Figure 5.15 shows how the domain knowledge and the obtained data can be used in parallel to apply MMP method.

The obtained knowledge here are used for identifying potential assignable causes (PAC), and a normative model. This added knowledge helps to generate several models (GM).

On the other hand, the real-time data is used in order to extract descriptive models (DM). The MMP method is in charge to find the minimum distance between GM's and DM's, and diagnose the cause of deviations.

5.3.1.3 Logic of MMP method

The following definition will formally illustrate the logic of **MMP** method. It defines how this method diagnoses a deviation in the structure of a process model.

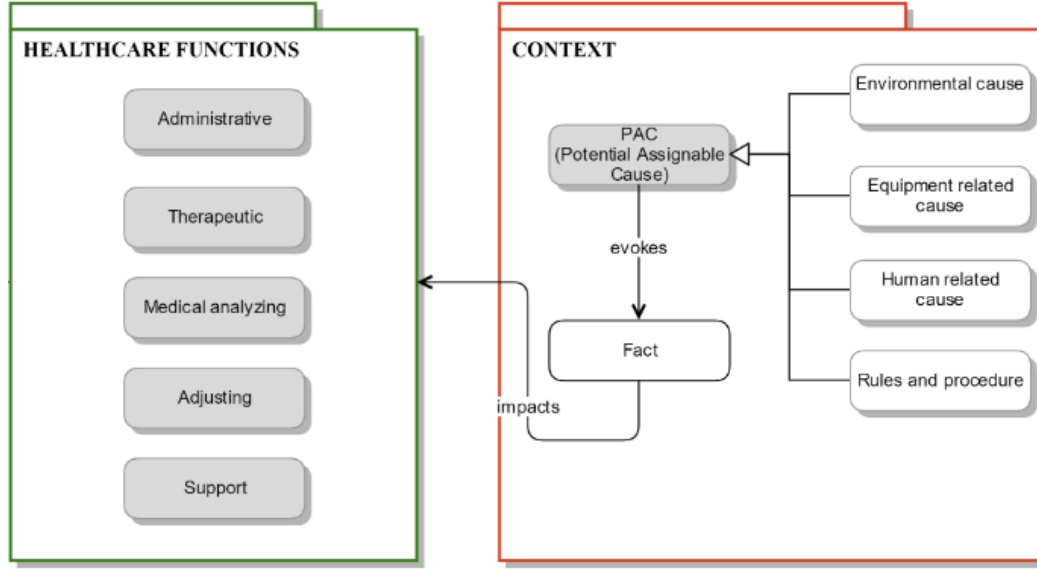


Figure 5.16 – The newly added context package for representing the effect of PAC on the healthcare functions.

$$\begin{aligned}
 & (\forall \ GM_j \ \exists \ PAC_j) \\
 & \quad \therefore \\
 & \quad (GM_j \cap DM_i) = \{PAC_j\} \\
 & \rightarrow Min.Distance\left(\bigcup_{j=1}^n ProDIST(DM_i, GM_j)\right) = ProDIST(DM_i, GM_j)
 \end{aligned}$$

According to the definition 5.3.1.3, for all of the generated models (GM) there exist one or multiple **PAC**'s.

Therefore, the potential assignable cause is shared between the descriptive model (DM_i) and the generated model (GM_j); **if** the distance between DM_i and GM_j is **the minimum distance** among all the existing distance values.

Note, that **ProDIST** is an algorithm developed in this research work to measure the distance between two process model.

Prior to explaining this algorithm, it is necessary to explain how the system can generate GM models by using **PAC**'s and the reference model (**RM**).

5.3.1.4 Deducing GM's (generated models)- DIAG meta-model is revisited

To carry on with this task, the DIAG meta-model is revisited, and a new package as **context** is added. Within this package, the potential assignable causes (PAC) and their relationship with existing functions in a process model are expressed. Figure 5.16 shows this new package on DIAG meta-model.

Within this package, several classes of causes are inherited from **potential assignable cause (PAC)** class. These causes are *environmental causes*, *equipment related causes*,

	Therapeutic	Administrative	Adjusting	Medical analyzing
Environmental cause	Replace (administrative activity)	Add (adjusting activity)	NULL	Add (adjusting activity)
Equipment related cause	Add (adjusting activity)	Add (adjusting activity)	NULL	Add (adjusting activity)
Human related cause	Add (adjusting activity)	Add (adjusting activity)	NULL	Add (adjusting activity)
Rules and procedures	NULL	Remove	Remove	Remove

Table 5.7 – This table shows an example of how the relationship among PAC’s and functions for patient pathways can be defined. It should be noted, these rules can be modified by the domain experts.

human related causes, and *rules and procedure* which can impose deviations on the activities of the reference model.

These causes were adapted from the elements of the fish-bone—cause and effects—diagrams that are applied in quality engineering to perform manual root-cause analyses (Harel et al., 2016; Taner et al., 2007).

The considered **impact actions** (they cause deviations) within the *fact* class are:

- Adding an activity.
- Removing an activity.
- Replacing an activity.
- NULL, for not taking any actions.

As already explained, these activities can be executed by the package of healthcare functions.

Table 5.7 considers the defined potential assignable causes and the different types of functions that are relevant to process activities.

Within this table, several modifiable rules are illustrated. The system generates deviating processes by the defined rules in this table.

For instance, the domain experts can define if the *environmental causes* can affect the *therapeutic functions*, and what are the rules for a certain process.

According to table 5.7, in this situation the activity will be replaced by an administrative activity.

The **fact class and its corresponding method** is realized within algorithm 12. To execute this algorithm, this research work considers a process model as a **string of characters**, and each character represents an activity.

Within algorithm 12, a *public* concept as a list is defined within the “Fact class”. Four functions are devised inside this list. One of these functions is **positionOfActivity** for detecting the position of the activity that is being affected by a *PAC*. The other three functions are the main **impact actions**.

The **Add function** helps the user to define a **fact** which is adding a new activity as the deviation caused by a *PAC*. The **Remove function** is defined for the user to indicate

if a *PAC* exists in the process which activity will be removed. The **Replace function** helps to indicate a rule for replacing an activity by a deviating action caused by a *PAC*.

An illustrative example

Let's illustrate this by an example.

Consider an environment in which a certain list of activities are executable (c.f. table 5.8). This general list includes $\langle "a", "e", "f", "b", "c" \rangle$. Accordingly, a process model represents the main activities that are executed normally. This model is defined as the reference model (RM) and $RM_1 = \langle "a", "b", "c" \rangle$.

Now, by looking at table 5.8, the domain expert is able to see the types of activities which are detected by the system (thanks to the interpretation rules defined in chapter 3). Therefore, he or she defines the potential assignable causes and their impacts on the corresponding activities. This action by the domain expert can be seen in table 5.7.

After this step, the system iterates through the rules defined by the domain expert, and it captures the **facts** and their impacts. Therefore, it deduces the generated models (GM) each time it encounters a new rule on an activity.

The generated models for this example are presented in table 5.9. Now, the question is how to measure the distance of a generated model from a descriptive model.

In light of this, the **ProDIST** algorithm is devised in the next section. This algorithm has a purpose to measure the distance between two process models; however, the ProDIST algorithm can be classified as a novel method for **conformance checking activity of process mining**. Next section explains the ProDIST algorithm.

All Activities	Type	RM	PAC (defined by the user)	Impact Action (defined by the user)
a	Administrative	a	Equipment	Use the add function defined in FACT
b	Adjusting	b	Rules & procedure	Use the Remove function defined in FACT
c	Therapeutic	c	Human related	Use the add function defined in FACT
e	Administrative	–	–	–
f	Adjusting	–	–	–

Table 5.8 – An example to illustrate how user and the system should interact. User based on the type of activity and the potential assignable cause can define which impact action should be triggered.

Algorithm 12 The FACT class

```
1: procedure Define the FACT class
2: input ReferenceModel;
3: input ReferenceActivity & DeviatingActivity & PAC
4: FACT = setClass(
5:   name = "FACT",
6:   private = list(impactActions = c("Add", "Remove", "Replace", "NULL")),
7:   public = list(
8:     positionOfActivity = function(string1, string2, startPosition = 1, n = 1)(
9:       a1 = unlist(strsplit(substring(string1, startPosition), string2))
10:      if (length(a1) < n + 1) then
11:        return(0);
12:      return(sum(nchar(aa[1:n])) + startpos + (n - 1) * nchar(str2) )
13:    end if)
14:     Add = function(ReferenceModel, ReferenceActivity, DeviatingActivity, PAC, After
= TRUE)(
15:       Position = positionOfActivity(ReferenceModel, ReferenceActivity)
16:
17:     for (i in 1:length(ReferenceModel)) do
18:
19:       if (After == TRUE) then
20:         GM = append(ReferenceModel, DeviatingActivity, after = Position)
21:
22:       else
23:         GM = prepend(ReferenceModel, DeviatingActivity, before = Position)
24:         return(GM) ; return(PAC)
25:       end if
26:     end for)
27:     Remove = function (ReferenceModel, DeviatingActivity, PAC)(
28:       Position = positionOfActivity( ReferenceModel, DeviatingActivity)
29:       GM = ReferenceModel[-Position]
30:       return(GM); return(PAC)
31:       ▷ we avoided redefining the positionOfActivity function
32:     )
33:     Replace = function (ReferenceModel, ReferenceActivity, DeviatingActivity, PAC)(
34:       GM = sub(ReferenceActivity, DeviatingActivity, ReferenceModel)
35:       return(GM); return(PAC)
36:     ) )
37: end procedure
```

Order	GM	RM	PAC
1	a, f, b, c		Equipment
2	a, f, b, c		Human related
3	a, f, b, c		Environmental
3	b, c	a, b, c	Rules and procedure
4	a, c		Rules and procedure
5	a, b, f, c		Equipment
6	a, b, e		Environmental
8	a, b, f, c		Human related

Table 5.9 – This table presents the generated models for the mentioned example. Note that the facts and their impacts are addressed each time by using one activity and one PAC.

5.3.1.5 ProDIST algorithm: a novel method for measuring the distance between two process models

Inspired by the **edit distance** algorithms in computational linguistics for quantifying the similarities between two **strings** (Kouylekov, 2006), this research work developed a method for evaluating the distance of two **process models** from each other.

Edit distance algorithms such as the **Levenshtein distance** (Pettersson et al., 2013) are applied mostly in natural language processing (NLP) to address spelling corrections.

In essence, this research work considers each **process model as a string of characters**. Each **character represents an activity** inside the process model. By this approach, **ProDIST algorithm** is able to detect the differences between two process models.

If the distance of two models are greater than 0, then three different situations appear by these actions:

1. An activity is **removed**.
2. An activity is **added**.
3. An activity is **replaced**.

The following represents the steps of this ProDIST algorithm by a running example:

Consider these two processes:

- $Process1 = \langle a, b, c \rangle$
- $Process2 = \langle e, f, c \rangle$

1. **Step 1:** Convert the processes into strings of characters.

$Process1 = \langle "a", "b", "c" \rangle$ $Process2 = \langle "e", "f", "c" \rangle$

2. **Step 2:** Create a custom matrix.

Table 5.10 shows the result of this step. The header of columns will be the name of process # 1 activities.

The rows will be assigned to the name of process # 2 activities. The orders of activities will start from 0. This (0) represents the number of actions needed to transform a *NULLcharacter* into another *NULLcharacter*.

Note the colored cells and the empty ones in table 5.10. In order to fill the empty cells, the algorithm uses those boomerang-shaped areas of the matrix to carry on the calculations (explained in the following).

The final decision to get the distance of the full strings (processes) from each other will be obtained by the last cell where the two furthest characters meet (in the above table these characters are "c" and "c").

3. **Step 3:** Compare the distance of a set of activities of process# 1 from a set of activities of process# 2.

As an example, by considering the order of activities in process# 1, activity sets can be: $\{[a], [a, b], [a, b, c]\}$ and similarly activity sets for process# 2 can be $\{[e], [e, f], [e, f, c]\}$.

	A NULL character	a	b	c
A NULL character	0	1	2	3
e	1			
f	2			
c	3			

Table 5.10 – Creating a matrix of processes activities and their orders.

At each time one member of activity sets will be compared with another one in the other process.

Generally, the goal is to measure the number of actions needed to transform one string to another one (one process model to another model).

For example, as shown in table 5.10, in order to transform a *NULL character* into the second process ($Process2 = \langle "a", "b", "c" \rangle$) 3 **actions** are required; **adding "e", "f", and "c"**.

As another example, the distance between [a] and [e] is 1, because by only one action "a" is **replaced** by "e". Accordingly, the distance between [a,b] and [e] is 2. Since, "a" is **replaced** by "e" and "b" is **added**.

To automatize this comparison, the algorithm considers certain conditions. These conditions are presented in the algorithm 13.

Algorithm 13 ProDIST logic for comparing the distance of activity stes

```

1: procedure ProDIST logic
2:   input Distance.Matrix[i, j];
3:   input Process.1 & Process.2                                ▷ String of processes
4:
5:   for (i&j in 1:n) do
6:     if Process.1[i] == Process.2[j] then
7:       Distance.Matrix[i, j] = Distance.Matrix[i - 1, j - 1]
8:     else
9:       Distance.Matrix[i, j] = min(Distance.Matrix[i - 1][j], Distance.Matrix[i -
10: 1, j - 1], Distance.Matrix[i, j - 1]) + 1
11:     end if
12:   end for
13:   return Distance.Matrix
14: end procedure

```

This means for every intersection between members of activity sets in process#1 and # 2, if two activities are similar, then the distance is equal to the **diagonal value** of that intersection.

Otherwise, the algorithm considers the **minimum value** among the boomerang-shaped cells prior to the intersection and **adds 1**.

Let's have a look at the example in table 5.11 where the previous matrix is evolved accordingly.

In this example, character "a" and "e" are compared. Since, these values are not similar, the algorithm considered the **minimum.value(of the colored cells)+1**.

Now look at table 5.12. As the algorithm continues to fill the matrix, it reaches two activities that are similar. The two ending 'c' activities are repeated in both

	A NULL character	a	b	c
A NULL character	0	1	2	3
e	1	$\min(1,0,1) + 1 = 1$		
f	2			
c	3			

Table 5.11 – It shows how to continue and evolve the comparison between two strings.

	A Null character	a	b	c
A Null character	0	1	2	3
e	1	$\min(1,0,1) + 1 = 1$	$\min(1,1,2) + 1 = 2$	$\min(2, 2, 3) + 1 = 3$
f	2	$\min(1,1,2) + 1 = 2$	$\min(2,1,2) + 1 = 2$	$\min(3,2,2) + 1 = 3$
c	3	$\min(2,2,3) + 1 = 3$	$\min(2,2,3) + 1 = 3$	Diagonal value = 2

Table 5.12 – Final results of the example: the distance between two process models is 2. Two actions are required to convert one model to the other one. By **replacing two activities**.

processes; therefore, the algorithm selected the **diagonal value of this intersection**.

As shown in table 5.12, the distance between these two process models is equal to the value of the furthest cell in the matrix; *distance = 2*. This is due to the need to **replace** 2 primary activities to transform one process to the other one.

The following algorithm 14 presents the implementation of this algorithm from the first step to getting the distance of models. A function as **ProDIST(process1, process2)** is devised to consider a process as a string of characters. These characters represent the process activities.

Let's apply this algorithm on the earlier example, to see the generated models that were obtained. Table 5.13 shows the result for applying the ProDIST algorithm to obtain the distance between each generated model and the descriptive model.

As shown, the descriptive model has a minimum distance from *GM7*. Therefore, it is possible to mention that the deviation within the descriptive model may be caused by the same potential assignable cause on the “therapeutic activity” within the process.

Evidently, it is possible that the distance of *DM* from multiple *GM*'s gets a similar value. In this case, the process should be observed in a certain period of time to see the changes in behavior of the process.

Until here, a new method named as MMP for performing business process diagnostic is introduced. This method used two subjects: (i) A primary domain knowledge about potential assignable causes. (ii) The data coming from execution of real processes.

In the body of this method two functionality is developed: First, a class known as FACT and corresponding functions of it are devised to generate fake processes. These processes are generated by the injection of the domain knowledge.

Second, the ProDIST algorithm is introduced for measuring the distance of real processes from the generated ones. By finding the minimum distance, it is possible to diagnose the real process.

Next section presents the second automatic process diagnosing approach.

Algorithm 14 ProDIST implemetation

```
1: function ProDIST(process1, process2)
2:   process1 = to.string(process1);
3:   process2 = to.string(process2);
4:   process1.activities = string.split(process1);
5:   process2.activities = string.split(process2);
6:    $\triangleright$  if processes have different sizes an index is used
7:   index = sequence(n.min = min((n.process1 = length(process1.activities),
   n.process2 = length(process2.activities)));
8:   n.max = max(n.process1, n.process2);
9:
10:   $\triangleright$  simplifying the results if a process has one activity
11:  if n.min == 1 then
12:    distance = sequence(1, n.max) - (process1.activities[index] ==
   process2.activities[index]);
13:    output = matrix(distance, nrow=n.process1);
14:    return(output)
15:  end if
16:
17:  output = diagonal(cumulative.sum(ifelse(process1.activities[index] ==
   process2.activities[index], 0, 1));
18:
19:  output[2,1] = output[1,2] = output[1,1] + 1 ;
20:  if n.max > 2 then
21:
22:    for i in 3:n.min do
23:      for j in 1:(i-1) do output[i,j] = output[j,i] = output[i,i] - output[i - 1, i - 1] +
   output[i - 1, j];
24:    end for
25:  end for
26: end if
27:   $\triangleright$  processes with different numbers of activities
28:  if n.process1 != n.process2 then
29:    extra.activities = sequence(1, n.max - n.min);
30:    extra.activities = sapply(extra.activities, function(x) x + output[, n.min]);
31:  end if
32:  if n.process1 > n.process2 then
33:    output = row.bind(output, transpose(extra.activities));
34:  else
35:    output = column.bind(output, extra.activities);
36:  end if
37:  Return output
38: end function
```

Order	GM	Affected activity	DM	Distance ProDIST(GM, DM)	PAC
1	a, f, b, c	administrative	a, b, e	3	Equipment
2	a, f, b, c	administrative		3	Human related
3	a, f, b, c	administrative		3	Environmental
4	b, c	administrative		3	Rules and procedure
5	a, c	Adjusting		2	Rules and procedure
6	a, b, f, c	Therapeutic		2	Equipment
7	a, b, e	Therapeutic		0	Environmental
8	a, b, f, c	Therapeutic		2	Human related

Table 5.13 – This table compares the distance of a descriptive model extracted from an event log and the generated models.

More on ProDIST algorithm

Within the fundamental definitions of conformance checking a situation is imagined where both event log and a process model are given. The goal is to detect similarities and discrepancies between the modeled behavior and the observed one.

Therefore, conformance checking methods are developed by an intuition to compare two behaviors (Rozinat et al., 2008). The example of these methods are token replay, alignments, comparing footprints, etc.

The **ProDIST** algorithm developed in this research work shares the same intuition with other conformance checking methods since, it is able to **compare two different behaviors**.

As one may noticed, the MMP method analyzes the individual cases (the process of one patient at a time); what if there is a need to diagnose the ensemble of patient pathways **without having the normative model**?

Well, the next section presents the **DIAG method** to target this challenge.

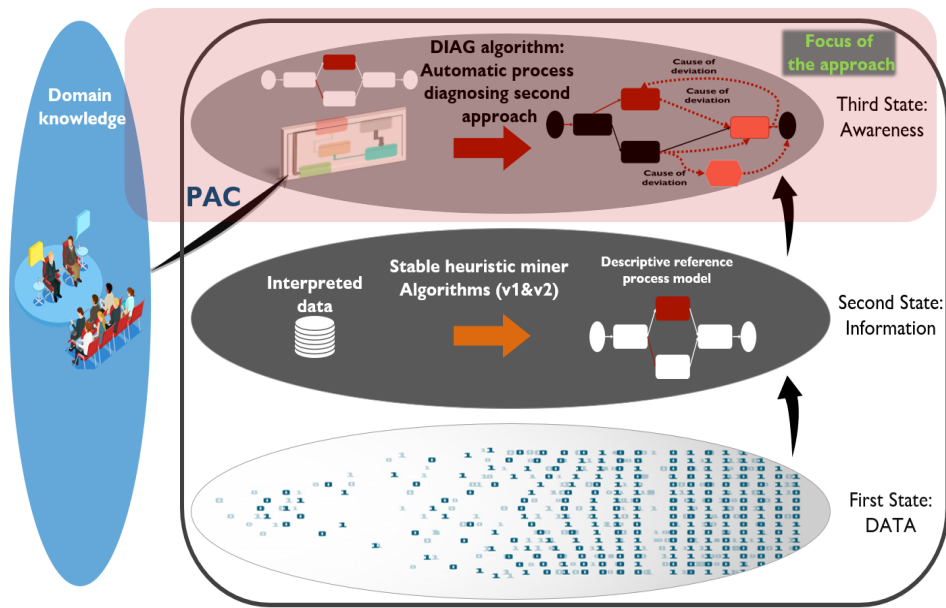


Figure 5.17 – An illustration of the second diagnostic approach devised in this research work. Diagnosing the deviations by applying both the stable heuristic miner algorithms and the corresponding domain knowledge.

5.3.2 Automatic diagnosing approach 2: DIAG method

The second approach for diagnosing business processes is about bringing together the **stable heuristic miner algorithms** and the defined **potential assignable causes (PAC)**. Figure 5.17 illustrates the position of this section in the approach of this thesis.

The intuition of the stable heuristic miner algorithms is related to the **detection of deviating and unstable behaviors**. From this perspective, these algorithms are capable to diagnose automatically the causes of deviations, **if they get enriched by the domain knowledge**.

As mentioned by the previous method, **potential assignable causes (PAC)** are modeled thanks to the **DIAG meta-model**. The last version of this meta-model is shown in figure 5.20.

Now, the objective is to develop the stable heuristic miners by not only the *data* coming from the information systems, but also, by the addition of the *knowledge* obtained from domain experts (c.f. figure 5.18).

Since the stable heuristic miner algorithms were introduced in chapter 4, their definitions are not revived here.

As a reminder, based on the approach of the **stable heuristic miner V2**, only certain behaviors are discovered that respect these conditions :

1. **Activities** that are stable and their corresponding observations are between the two thresholds ($[LCL < \bar{x}_s < UCL]$).
2. **Activities** that have high variations in their behaviors. These activities are above the upper control limit threshold ($[UCL \leq \bar{x}_s]$).

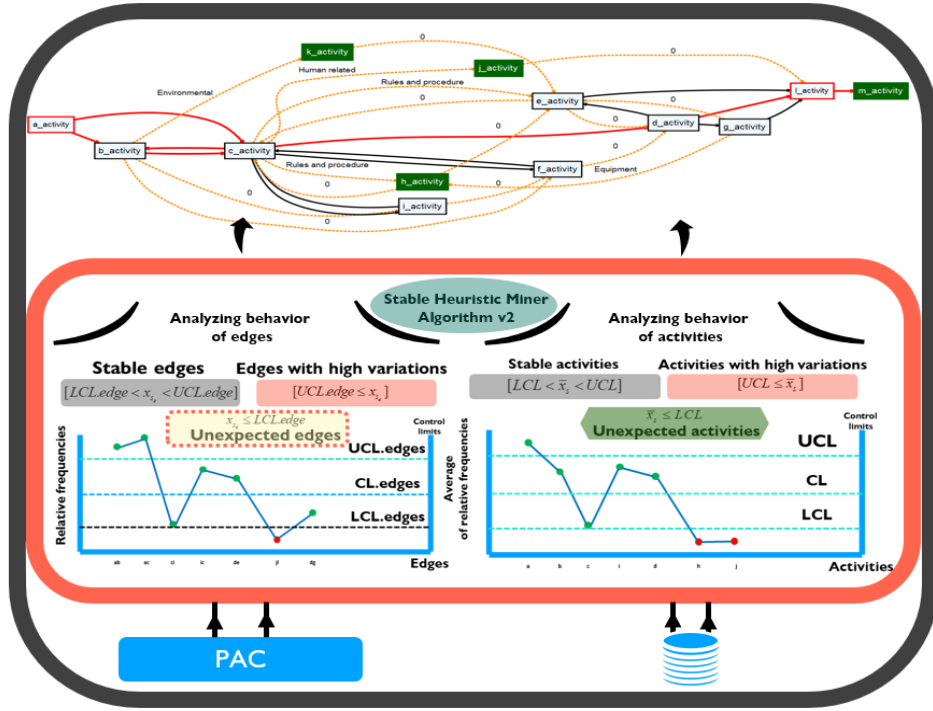


Figure 5.18 – This figure shows the adjustment of the stable heuristic miner v2 algorithm by injecting potential assignable causes as the domain knowledge.

3. **Edges** that show stable behaviors ($[LCL.edge < x_{se} < UCL.edge]$).
4. **Edges** that have high variations and are above the UCL ($[UCL.edge \leq x_{se}]$).

In contrast, the new method—DIAG— considers the activities and edges that are **lower than LCL and LCL.edge**; then, it uses the **domain knowledge** to diagnose the causes of deviations.

To do so, it uses a *matching procedure*. According to this procedure, if these deviating behaviors correspond to the a-priori knowledge—provided by the domain expert— the algorithm will show the cause of deviations for each anomaly.

Figure 5.18 shows how the application works by the DIAG method. After receiving the event log and the domain knowledge (PAC), the stable heuristic miner algorithm will adjust its functionality and result in a process model with diagnosed deviations.

Algorithm 15 shows the formal construction of this method.

Let's demonstrate the notion of this algorithm by an example that earlier was demonstrated in chapter 4; the 'L' event log .

Algorithm 15 DIAG algorithm for predictive discovery of process models.

```

1: procedure DIAG
2:   input DomainKnowledge;
3:   input EventLog
4:   ▷ previous algorithms of stable heuristic miners will be executed
5:   Execute "FootprintMatrix generation";
6:   Execute "extract the thresholds";
7:   Execute "identify the statuses of activities and edges";
8:   Domain.Knowledge.df = data.frame(DomainKnowledge["Activity"], DomainKnowl-
     edge["Deviation"], DomainKnowledge["PAC"]);
9:
10:  Merged.DomainKnowledge.UnstableActivitiesEdges =
    as.matrix(merge(DomainKnowledge, UnstableEdges, by.x= c("Activity", "Devi-
      ation"), by.y = c("FromActivity", "ToActivity"), all.y = TRUE));
11:
12:  Deviating.Behaviors = Merged.DomainKnowledge.UnstableActivitiesEdges %>%
13:  mutate.all(replace(as.character(.), is.na(.), "0"));
14:  ▷ Nodes
15:  Node.Stable = data.frame(Union(unique(StableBehavior.Activity),
    unique(StableBehavior.Edge)), attribute= "normal");
16:  Node.Unstable = data.frame(Union(unique(UnstableBehavior.Activity),
    unique(UnstableBehavior.Edge)), Attribute.Shape = "tripleOctagonal", At-
    tribute.Color = "darkgreen");
17:  Node.Hot = data.frame(Union(unique(HotZone.Activity), unique(HotZone.Edge)),
    Attribute.Color = "red");
18:  All.Nodes= combine(Node.Unstable, Node.Stable, Node.Hot);
19:  ▷ Edges
20:  Stable.Edge = data.frame(match(ProcessActivity, All.Nodes), attribute = "nor-
    mal");
21:  Hot.Edge = data.frame(match(HotZones, All.Nodes), Attribute.Color = "red");
22:  deviation = data.frame(match(Deviating.Behaviors, All.Nodes), Attribute.Style
    = "dashed", Attribute.Color = "orange"); All.Edges = combine(deviation, Hot.Edge,
    Stable.Edge)
23:
24:  devise.graph(All.Nodes, All.Edges)
25: end procedure

```

Activity	Deviation	PAC
c	j	Human related
b	k	Environmental
c	h	Rules and procedure
j	i	Human related
c	e	Rules and procedure
g	h	Equipment
...

Table 5.14 – This table shows an illustration of how the **domain knowledge** will be recognized by the application.

5.3.2.1 An example for the second automatic diagnosing approach

Presume that domain experts are able to provide some information for certain activities inside an organization, to capture what are the **potential assignable causes**, and what would be their **impacts**. This is similar to what has been shown in table 5.7, where the domain experts defined the causes and the facts corresponding to the impact of each cause.

Table 5.14 shows an example of the *mentioned domain knowledge*. It is possible for the user to define one or several number of entries to identify how one or several PAC can affect activities.

Now, imagine during a data gathering procedure an event log is obtained.

$$L = [< a, b, c, d, e, l, m >^{12}, < a, b, f, d, e, l, m >^2, < a, b, c, d, g, e, l, m >, \\ < a, b, c, d, g, h, e, l, m >^5, < a, c, b, c, d, l, m >^6, < a, c, f, c, d, l, m >^3, \\ < a, c, b, i, c, e, c, d, g, l, m >^5, < a, c, b, c, f, c, d, l, m >^6, \\ < a, c, b, c, i, c, h, c, d, l, m >^4, < a, c, b, c, i, f, c, d, g, l, m >^6, < a, b, c, j, l, m >^6, \\ < a, c, b, k, e, d, l, m >^4]$$

This event log is obtained at the data state from an information system (c.f. figure 5.17). This event log has 13 activities and the hidden information corresponding to their behaviors.

Now, by the addition of the domain knowledge, if the new diagnosing algorithm receives this event log, it will evoke the “matching procedure”, and it will generate the model shown in figure 5.19.

The activities and edges that are shown in **black** are representing the **stable behaviors**. The **red** color corresponds to activities and edges that have **higher variation** than their normal (stable) value.

The edges in the **dashed** form are **deviating connections** among activities. The activities in **green** are **deviating activities**.

As presented in figure 5.19 the deviation between **activity ‘b’** and **activity ‘k’** corresponds to an **environmental cause**.

The deviation between **activity ‘c’** and **‘f’** is related to a **human error**. The edge between **activity ‘c’** and **‘h’** is related to a change in **rules and procedure**.

When some edges demonstrate 0 values, it means that the provided domain knowledge did not match these deviations. Simply put, the domain knowledge was not adequate.

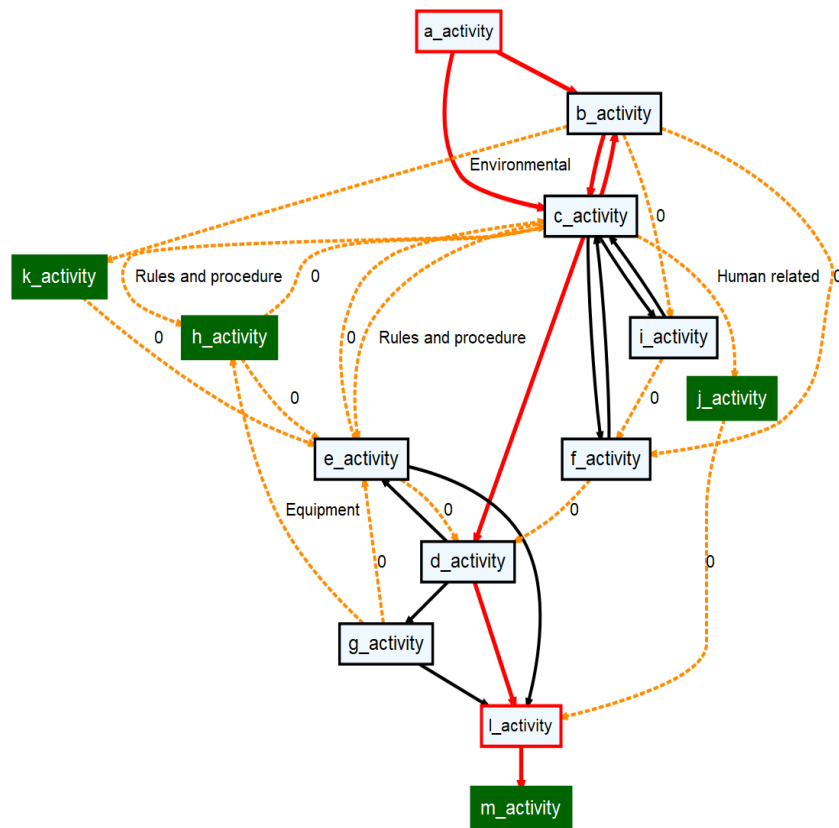


Figure 5.19 – The discovered model which shows the causes of deviations. The activities and edges that cause higher variations –than the normal value– are shown in red. The deviating activities are shown in green. The dashed edges show the unexpected connections among activities. The cause of these deviations are shown on each behavior. If a deviation does not correspond to the domain knowledge, it gets a 0 value.

Note that, it is possible to go even further and instantiate the causes of deviations. For example, an “Equipment” cause of deviations can be instantiated as a “malfunctioning MRI machine”.

To best of our knowledge, such a functionality in discovering predictive models from event logs has never been addressed in the literature of process mining.

Such an approach and including methods permit not only for discovering business processes but to diagnose them automatically as well.

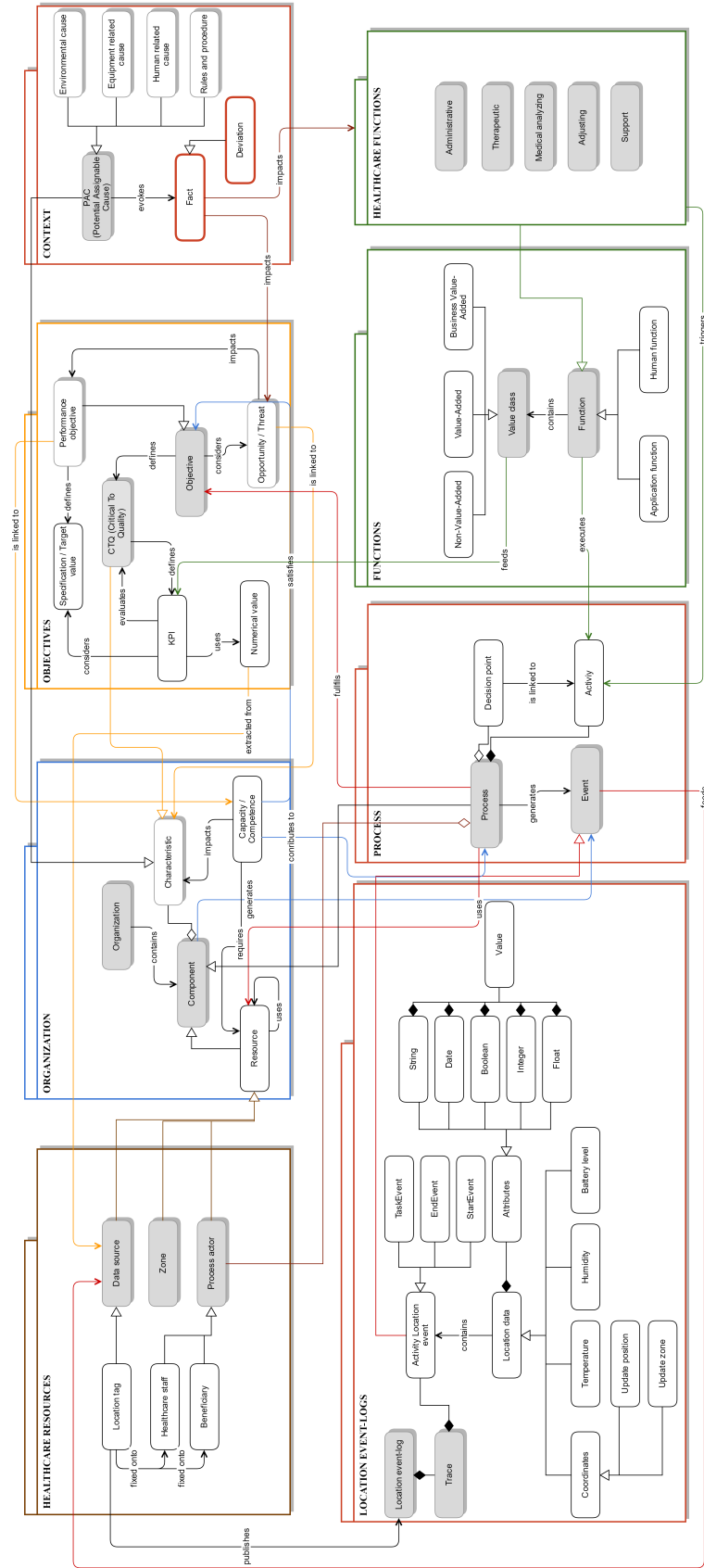


Figure 5.20 – The second version of DIAG meta-model to support automatic diagnosing of patient pathways.

5.4 Recap

The presented chapter described two main functions of the DIAG methodology.

In section 2, the business process analyzing function targets the lack of sophisticated methods for quantitative performance analyses, and proposes the application of statistical process control.

This approach is positioned within **enhancement** activity of process mining discipline.

Two presented methods for business process analyzing function were the **control charts** and the **process capability ratio** analyses. Their objective is to detect variations in the extracted data. This aids the domain expert to have an overall view of the performance of the processes.

Inside the section 3 of this chapter, two automatic process diagnosing approaches were introduced. The first one was the **MMP method** which was looking for the distance between the recorded behaviors in the information system and a simulated behavior by known causes of deviations.

Within this method, a novel algorithm (**ProDIST**) for measuring the distance of processes from each other was introduced.

The second process diagnosing approach presented as **DIAG method**, adjusted the stable heuristic miner algorithm (V2) by the possibility of receiving two sets of information instead of one.

The first information was the event log, and the second one was a-priori knowledge about the causes of deviations. This helped the experts to load the event log into the system and receive **not only a discovered model of the process execution, but a diagnosis on top of the model too**.

5.4.1 SWOT analysis of the chapter

Strength

Methods presented in this chapter targeted the described gaps in the literature according to chapter 2.

Lack of quantitative analyses beside modeling processes is an important issue in the state of the art of process mining. The application of SPC provides the mean to fill this gap.

Automatic diagnosing of processes is neglected in the literature. Most of researchers mistakenly considered analyzing a process as a synonym to diagnosing the causes of inefficiencies in a process.

The two methods of MMP and DIAG are developed here to address this issue.

The ProDIST algorithm introduces a new and fast approach for measuring the distance of behaviors from each other.



Weakness

Both MMP and DIAG methods are introduced primarily in this dissertation. They could lead to several causes for a certain deviation in a process. Therefore, their accuracy must be endorsed by predictive models.

In addition, the MMP method is not applicable for the process that contains decision points.



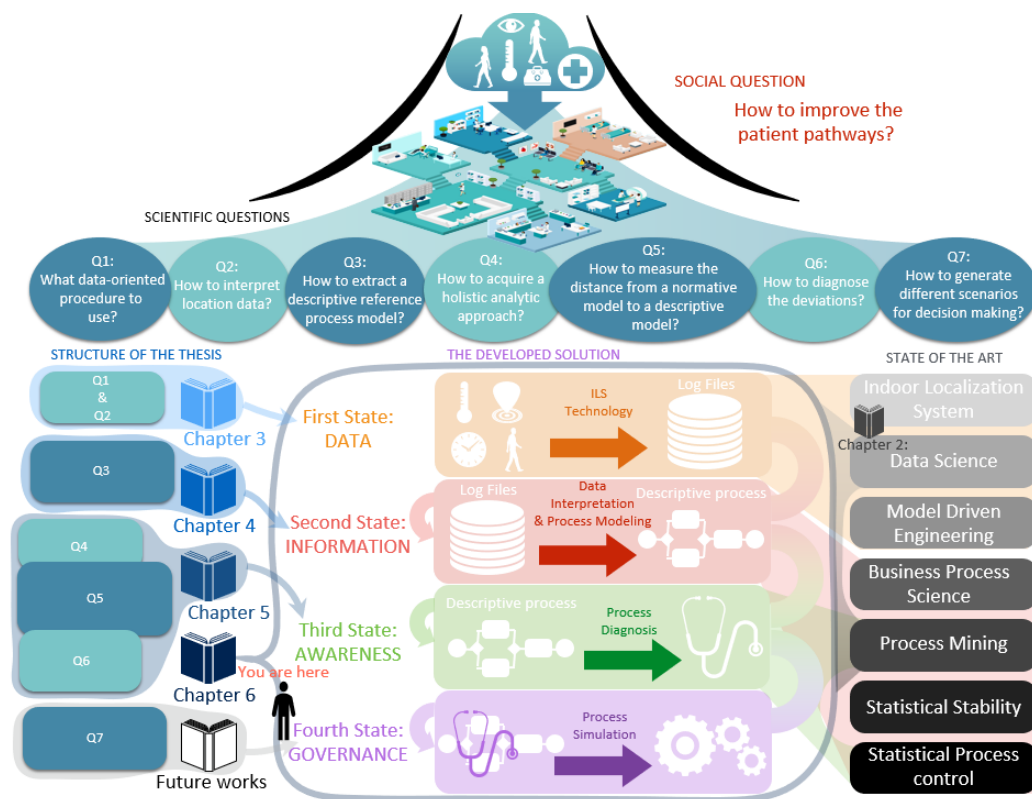
Opportunity

It would be a good approach to improve the accuracy issue of the diagnosing methods by performing multiple experiments. Such an approach leads to gathering of a training data set. This data can be used to devise a predictive model (modeled by supervised learning methods) to ensure about the performance of the methods.



Threat

The ProDIST algorithm have not been applied for process models containing decision points. This must be addressed in future works.



III

Third part: the experimentation

6

A case study: experimental results of the thesis

6.1	Motivation	163
6.2	Presentation of the case study	164
6.3	Preparation for interpreting the location data	166
6.4	Discovering and modeling patient pathways	171
6.4.1	Results of discovery algorithms	171
6.5	Analyzing the quality and performance of patient pathways	182
6.6	Automatic diagnosing of patient pathways	188
6.6.1	Miniscule Movements of Processes (MMP) method	188
6.6.2	Application of DIAG method	190

“Don’t walk through life just playing football. Don’t walk through life just being an athlete. Athletics will fade. Character and integrity are really making an impact on someone’s life, that’s the ultimate vision, that’s the ultimate goal – bottom line.”

Ray Lewis

6.1 Motivation

The objective of the presented case study in this chapter is to illustrate the applicability of the proposed methods within this research work. As illustrated in figure 6.1, this chapter presents the experimental results related to:

1. Modeling the existing concepts in patient pathways (seen in chapter 3).
2. The applicability of the stable heuristic miner algorithms ($V1, V2$) and a comparison with the classic heuristic miner algorithm (seen in chapter 4).
3. Presenting the application of SPC for evaluating the performance of the processes (seen in chapter 5).
4. The application of the two automatic diagnosing methods: **MMP** and **DIAG** (seen in chapter 5).

6.2 Presentation of the case study

During this concrete example, 7 scenarios in 7 departments of Toulouse hospital were simulated.

The Toulouse Hospital University is located in south of France with several establishments. More than 3900 physicians and 11600 hospital staff are welcoming around 280000 patients annually. It has been estimated that more than 800000 **medical appointments** are being registered each year.

Approximately, 400 patients are admitted to the emergency department daily. Due to its high volume of patients arrival and high-risk operations, this research was inducted in a much more smaller dimension by simulating the pathways of only 7 departments and around 300 patients.

During the initializing days of the experiment, the primary information of the hospital were gathered. These information correspond to the map of the facility and the monitored departments.

Moreover, certain required information were obtained by the domain experts. These information were about:

- A simple introduction of the environment.
- Identifying the possible executing operations in each zone of the facility?
- Verifying the types of each of these activities.
- What are the potential causes that could affect the execution of the identified activities.
- Considering the duration and the walking distance of patient pathways, what are the expected, desirable and outcast results of the patient pathways.

After discussing these issues, the objectives of the experiment were presented to the domain experts. This is an important topic to be considered for future studies. Based on the gained experiences, hospital staff (in particular physicians) are not eager to use the ILS technology.

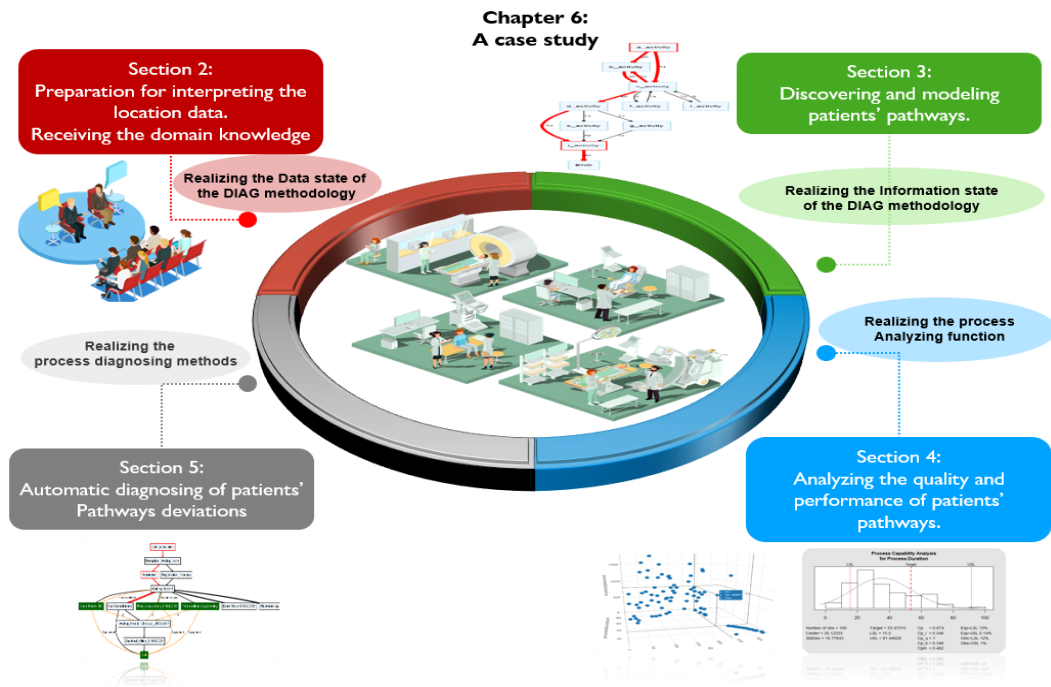


Figure 6.1 – Structure of chapter six.

Hence, the special advantages of applying ILS in healthcare environments must be crystal-clear for the staff.

The mentioned benefits were:

- Increase in safety and security of patients.
- Significant aid in monitoring patients who are suffering from dementia.
- Real-time localization of patients.
- Increase the agility of staff while operating in emergency situations.
- Potential advantages in **improving patients pathways**.

Finally, after this exchange of information, the case study was devised in conformity with the **DIAG methodology** states. This implies the realization of each state of this methodology.

Hence, by considering the primary information provided by the domain expert, the resources, functions, and potential causes of deviations were modeled.

The simulated experiment led in to generation of location data from the processes of 261 patients. These gathered data were used to apply other functions of the DIAG methodology.

This action is envisioned by figure 6.1, which presents the structure of this chapter.

6.3 Preparation for interpreting the location data

As shown in chapter 3, the task for modeling the existing concepts within the organization, is dedicated to **configuring the environment and the systems function**.

Figure 6.2 shows an abstracted version of the modeled resources. Figure 6.3 illustrates the functions that are defined to be executed in the hospital.

To start with gathering of the location data, each patient received a tag with a particular “eui (the tag id)”.

Afterwards, according to DIAG methodology, the **location data interpretation rules** will identify the activities that a patient has executed. Figure 6.3 shows a screen shot of the application when it replays the event logs by considering the location data interpretation rules. As shown, the map corresponds to the position of the patients. Aside from the map, the graph shows the instant that an **in/out zone event** is detected.

As illustrated in figure 6.6 and 6.7, the interpreted information will be used in the next function for modeling patient pathways.

In addition, a series of information are defined by the experts as an input to be used for the business *process diagnosing function*. Table 6.1 presents such an input which will be used for the MMP method. This domain knowledge is relevant to the existence of certain potential assignable causes (**PAC**) for the activities inside the UROLOGY department.

Accordingly, if such causes are affecting the activities, particular impact actions will be defined. For example, the “*Registration*” activity can be affected by **Human related causes** (like unavailable staff); therefore, if such a cause exists, then the corresponding fact is to **add a supplementary “Reception_ Waiting_ room”** activity.

As mentioned in chapter 5, it should be noted that the expert can identify several *causes* for a particular activity. This is why some activities are shown in **red** in the table 6.1. These activities that can be affected by one or multiple causes are:

- Registration
- Box Consultation
- Checkout_ Office _ UROLOGY
- Post_ consultation

This table will be used further down in section 6 of this chapter.

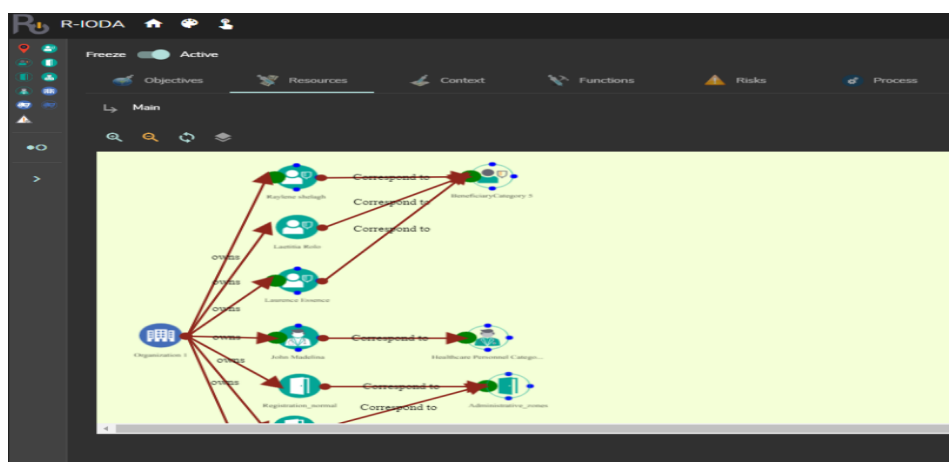


Figure 6.2 – Modeling existing resources in the hospital.



Figure 6.3 – Modeling provisioned human functions in the hospital.

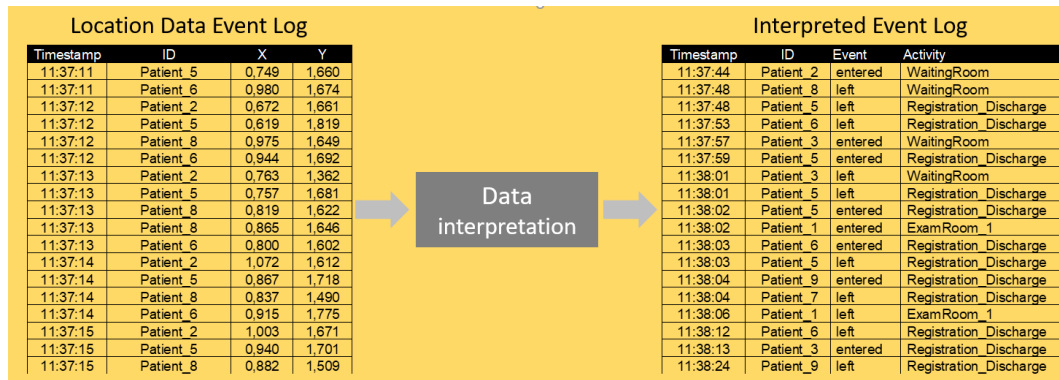


Figure 6.6 – An illustration of the interpretation procedure.

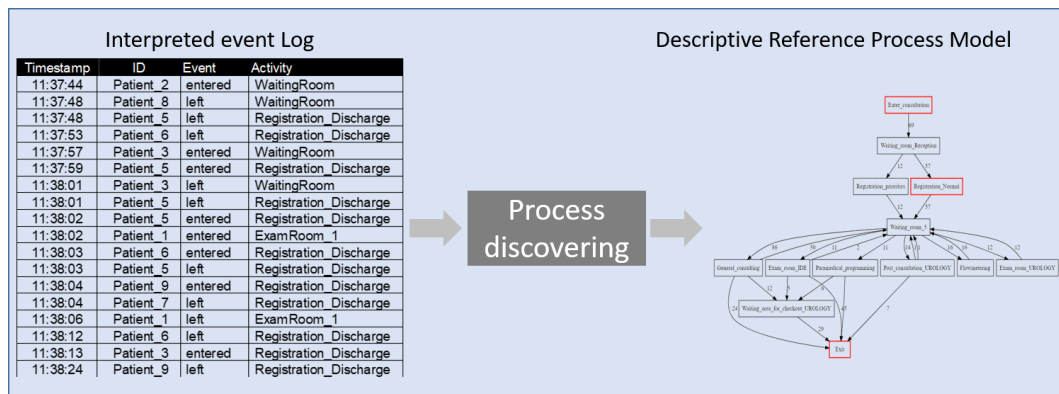


Figure 6.7 – An illustration of the process discovery procedure.

All possible activities in the organization	Type	PAC (defined by user)	Facts impact actions (defined by user)
Enter_consultation	Adjusting	NULL	NULL
Reception_Waiting_room	Adjusting	Rules and procedure	Remove
Registration	Administrative	Human related	Remove
Registration	Administrative	Human related	Add
Registration_priorities	Administrative	Human related	Reception_Waiting_room Add
Waiting_room5	Adjusting	Rules and procedure	Reception_Waiting_room Remove
Box Consultation	Therapeutic	Human related	Add
Box Consultation	Therapeutic	Environmental	Waiting_room5 Add
Exam Room UROLOGY	Medical Analyzing	Human related	Reception_Waiting_room Add
Flowmetering	Medical Analyzing	Human related	Waiting_room5 Add
Waiting area for checkout UROLOGY	Adjusting	Rules and procedure	Waiting_room5 Remove
Checkout_Office_UROLOGY	Administrative	Rules and procedure	Remove
Checkout_Office_UROLOGY	Administrative	Human related	Add
Post_consultation	Medical Analyzing	Rules and procedure	Reception_Waiting_room Remove
Post_consultation	Medical Analyzing	Human related	Add
Post_consultation	Medical Analyzing	Equipment	Waiting_room5 Add
Exam Room IDE	Medical Analyzing	Environmental	Waiting_room5 Add
Paramedical programming	Administrative	Rules and procedure	Reception_Waiting_room Remove
Exit	Adjusting	Rules and procedure	Remove

Table 6.1 – A table corresponding to the added domain knowledge by the experts. This will be used as an input for the MMP method to diagnose the deviations of patient pathways.

6.4 Discovering and modeling patient pathways

The objective for presenting the results of this section is in threefold.

- **First**, to obtain a descriptive reference process model showing the normal and stable pathways for all the departments in the hospital. Simply put, to visualize what is the **common pathways** for the patients, and which zones are *normally* being occupied by patients during the process execution.
- **Second**, to compare the results of the classic heuristic miner algorithm with the new methods presented in this research work.
- **Third**, performing the same methods to visualize the descriptive reference process model for each department. This helps to capture an image of common pathways for patients who have almost similar profiles. Ultimately, this model will be used at the **fourth state —awareness—** for the **business process diagnosing** function.

6.4.1 Results of discovery algorithms

In light of the mentioned objectives, three algorithms were applied to discover a model which represent the **descriptive reference process model**. The following will present the result of these algorithms:

1. Classic heuristic miner
2. Stable Heuristic Miner V1
3. Stable Heuristic Miner V2

In this order, figure 6.9, illustrates the result of the **classic heuristic miner algorithm**. There were 35 activities registered in the main event log.

Similar to what has been discussed in chapter 4, the classic algorithm only considers the frequency of relations among activities. Then, by calculating the **dependency measure** values it assigns a score to identify the dependency of activities to each other. Figure 6.9 presents 100 percent of the recorded information.

As one can observe in figure 6.9, this model is far from being suitable for further analyses. This is due to the **existence of noises** in the model and **the inability of this algorithm for removing them rationally**.

This algorithm tends to filter paths and activities in a counter-intuitive and arbitrary way, which is seen as a **principal flaw** for its performance (De Cnudde et al., 2014).

For example, figure 6.10 shows the filtered behavior by a threshold of 20. This means only activities that have higher dependency than 0.2 will be discovered.

Accordingly, figure 6.11 shows the process model discovered by a threshold of 50%. In the same vein, figure 6.12 illustrates only activities with a dependency higher than 0.8.

Obviously, the values of these thresholds are decided by a **random** approach. **It is not clear** how the healthcare experts can obtain the reference descriptive behavior of an event log.

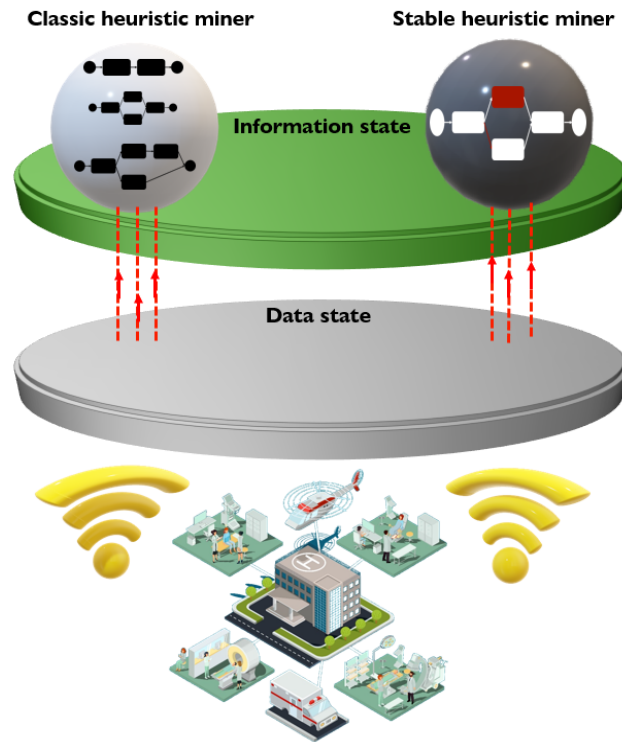


Figure 6.8 – A comparison between the classic heuristic miner and stable heuristic miner algorithms results.

Likewise, this logic of filtering information, based only on frequency not stability is seen in many algorithms such as fuzzy miner (Gunther et al., 2007) as well.

Most of users of the commercial process mining tools like DISCO (<https://fluxicon.com/disco/>) are familiar with the two thresholds that are dedicated for removing some information from the model. However, this action is completely random. For this reason, the domain experts fail at deciding which model would represent the common patient pathways.

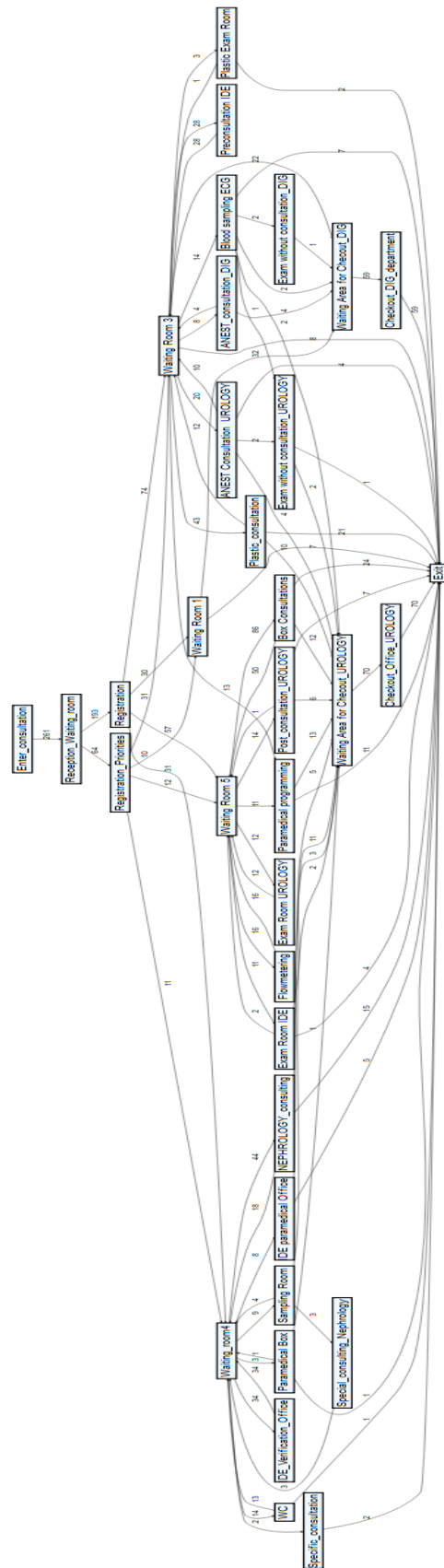


Figure 6.9 – The discovered model of the hospital by applying the classic heuristic miner approach. In this model, the threshold value is set at 0percent This model is hardly analyzable.

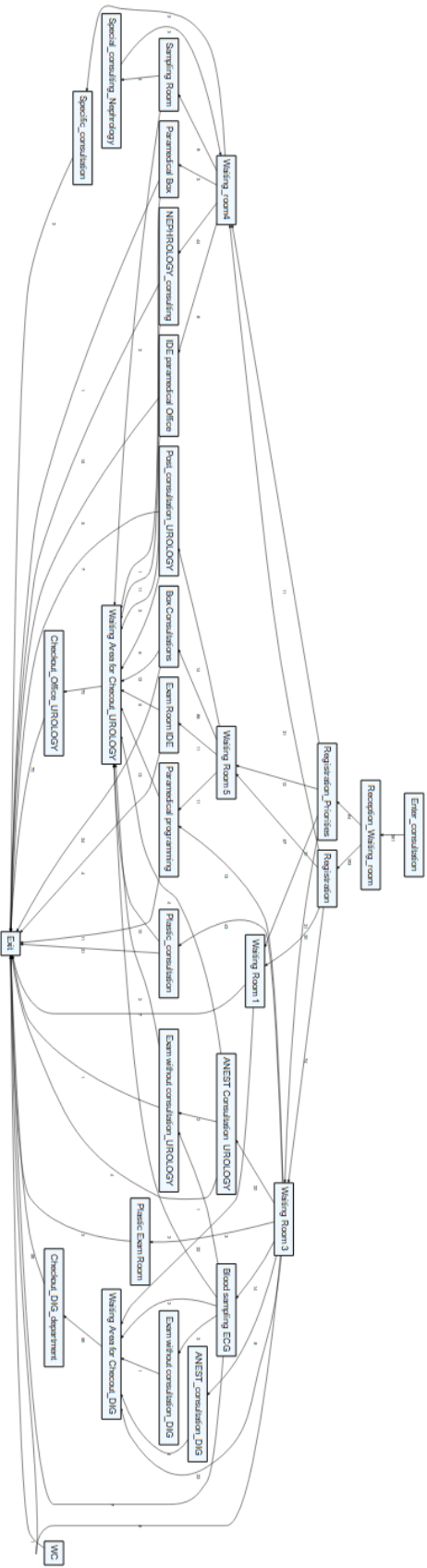


Figure 6.10 – The discovered model of the hospital by applying the classic heuristic miner approach. In this model, the threshold value is set at 20%.

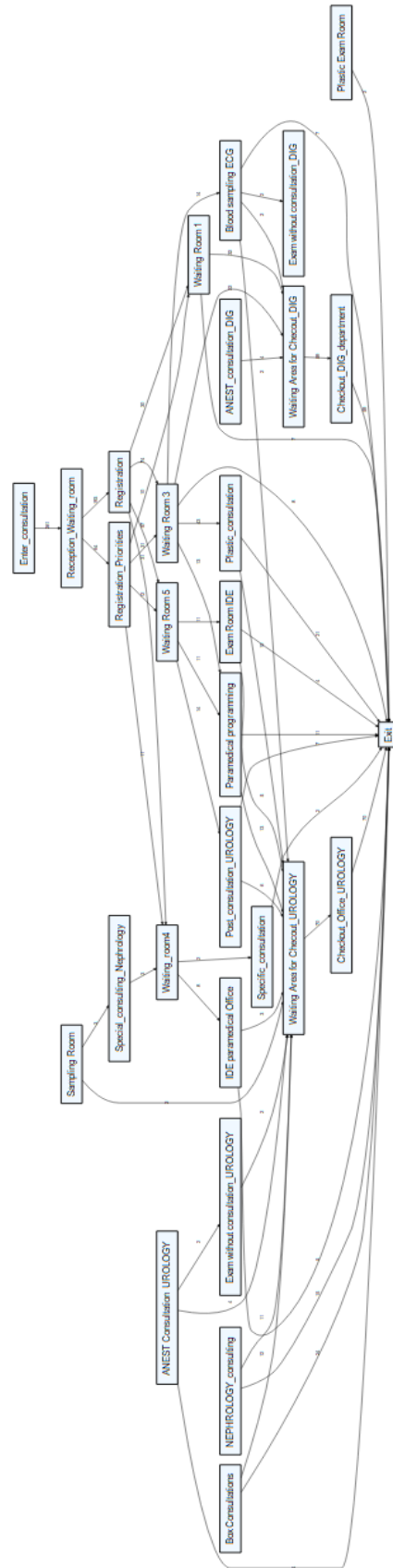


Figure 6.11 – The discovered model of the hospital by applying the classic heuristic miner approach. In this model, the threshold value is set at 50%.

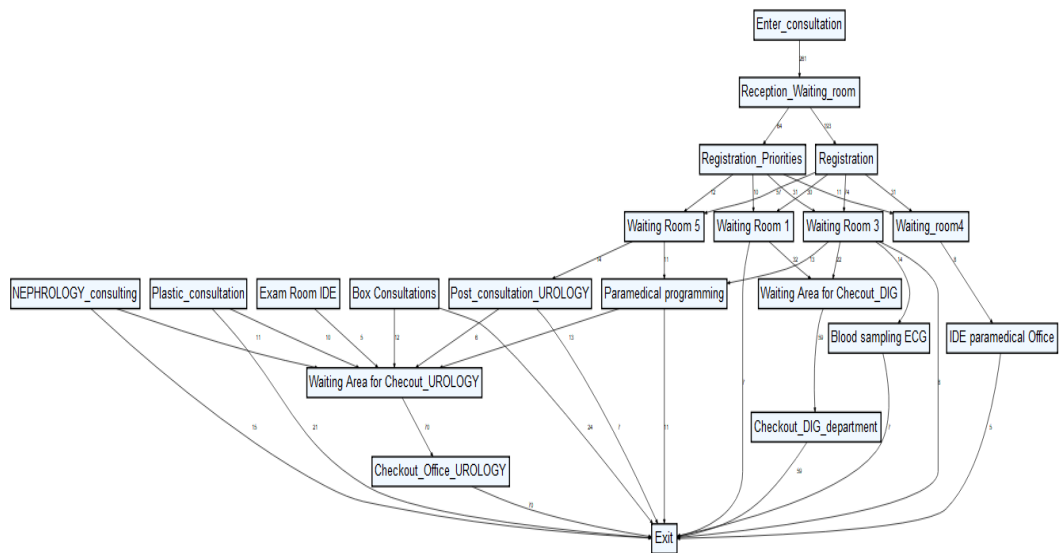


Figure 6.12 – The discovered model of the hospital by applying the classic heuristic miner approach. In this model, the threshold value is set at 80%.

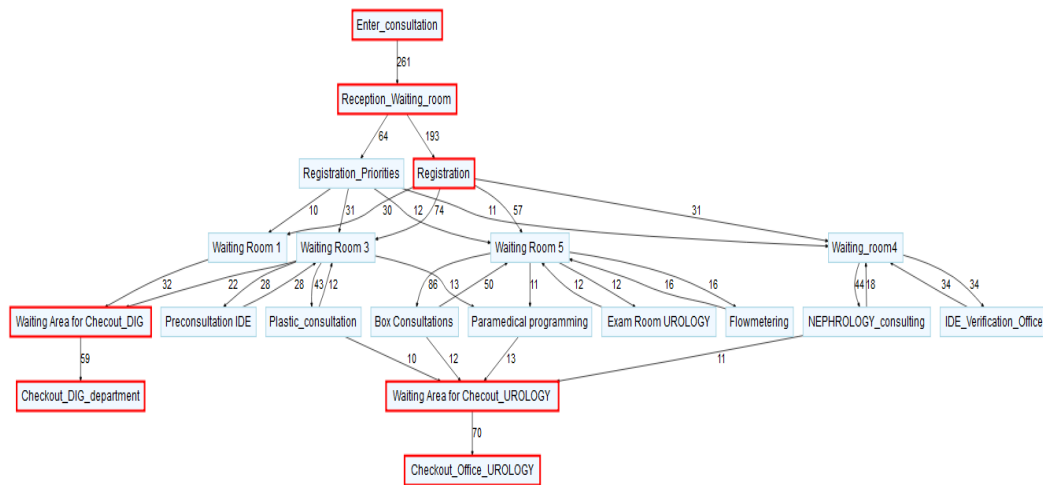


Figure 6.13 – The discovered model by the **stable heuristic miner V1**. Note that, the statistical stability is not considered for edges. This has been addressed by the second version of the algorithm.

	Lower than LCL	In the stable State	Higher than UCL
Number of activities	15	13	7
Total number of modeled activities		20	
Total number of observed activities (appeared within event log)		35	

Table 6.2 – A comparison between the number of observed behaviors in the event log (of all of the departments) and the modeled behaviors in the descriptive reference process model, thanks to stable heuristic miner V1.

In contrast, the stable heuristic miner algorithms are embracing the **statistical stability phenomenon** and determines the thresholds by extracting the stability level of behaviors from an event log. The two developed discovery algorithms in this research work (**stable heuristic miners; V1 and V2**) help to discover the common pathways for patients.

The result of the first version of **stable heuristic miner algorithm (V1)** is shown in figure 6.13.

By comparing the discovered models by the classic heuristic miner algorithm (figure 6.9, 6.10, 6.11, 6.12) and the model discovered by the stable heuristic miner V1 (figure 6.13), it can be observed that the model in figure 6.13 is much more pragmatically eligible for performing different analyses.

Accordingly, the domain experts can consider the descriptive reference process model to determine which activities are being executed normally by patients. The **automatic extraction of thresholds** from the event log removes the need to randomly define the value of thresholds.

As shown in table 6.2, out of the total number of 35 activities in the first model, 15 activities were detected with an instability lower than the lower control limit (LCL) and they are not shown in the descriptive reference process model. From the 20 remaining activities in the descriptive reference process model, 7 activities are considered as hot zones, which impose high instability and variation to the normal behavior of the process.

These hot zones are indicated in red. These are the activities whose average behavior values are higher than the upper control limit (UCL). This implies that such activities represent unusual and eccentric behaviors in the log which could lead to future problems.

To exemplify these statements, the “*waiting_room_5*” provides a good illustration. Based on the method of the stable heuristic miner, one can ensure that the probability of receiving the same behavior for this activity is high and all the activities related to “*waiting_room_5*” could be regenerated in the future. Therefore, the experts can plan the requirements for running such activities in the future. On the other hand, hot-zone activities such as “*Registration*” indicate that these activities are generating behaviors that are beyond the usual and stable behavior of all of the other defined activities.

As a further illustration, the incoming flow into the “*Registration*” activity can be considered. It can be seen that 193 cases enter this activity, which is higher than most of the existing flows in the process. Also, this outgoing flow has a high variation in comparison with the outgoing flows from “*Registration*”. Therefore, such behavior could cause potential bottlenecks in this activity and consequent instability in the process.

Similarly, the activity “*Waiting_room_Reception*” receives 261 cases and it has two outgoing flows with values of 193 and 68. It is obvious that **variation among these behaviors is significant**, therefore, the **unstable behavior** of this activity is detected.

As readers might notice, this version **only** applied the statistical stability approach for evaluating the **behavior of activities not edges**. This issue is covered by the **second version** of the algorithm.

Figure 6.14 shows the result of the **stable heuristic miner V2**. This algorithm discovers a descriptive reference process model by considering the stable behavior of the **both edges and activities together**.

Out of 88 edges existing in the model extracted by the classic heuristic miner approach (c.f. figure 6.9), only 35 edges are considered in the descriptive reference process model (c.f. figure 6.14). These edges are related to the stable and high variation behaviors.

The red edges between activities indicate that those edges whose values are higher than the *UCL* (upper control limit) threshold and are expressing behaviors that are not aligned with the tolerance of the process.

Thanks to this algorithm, two highly sensible areas in the process model are detected. By looking at the beginning and the end of the model, it is visible that the most flows of patients are in these areas. This issues several potential bottlenecks in the process model.

Clearly, when the attention of the highly unstable behaviors are focused in those areas, this creates potential causes of inefficiencies for the process.

By applying the new version of the algorithm, the domain experts can obtain more detailed information. For instance, the path among “*Enter_consultation*”, “*Reception_waiting_room*”, “*Registration*”, “*Waiting_Room_5*”, and “*Box_consultation*” activities creates high instability in the process. The behavior of this path should be normalized, since the algorithm detected highly abnormal behaviors.

This path *injects high variations* into the overall behavior of the process.

Furthermore, recalling the challenge of process discovery algorithms in chapter 4, the stable heuristic miner statistically addresses the issue of **noise in a discovered model**. This is done by automatically determining the stable level of information. This information is needed for the expert to have an overall understanding of process executions by patients.

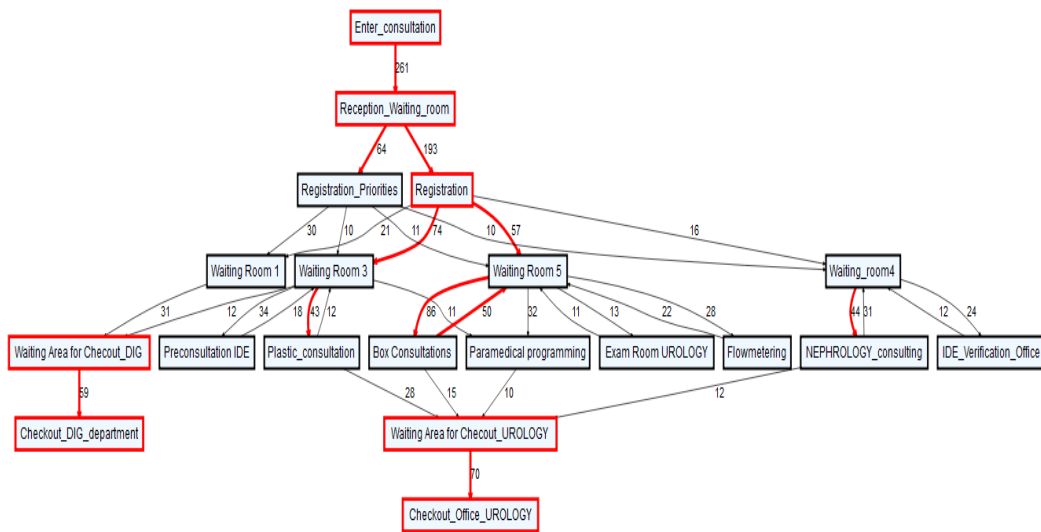


Figure 6.14 – The discovered model by the **stable heuristic miner V2**. **This model is related to the behavior of all of the 7 departments. Red activities and edges correspond to behaviors which represent where are the high variation behavior in the process.**

To address the third objective of this section, the processes of each department were investigated individually. As an example, the process model shown in figure 6.15 can be considered. It shows the patient pathways for the “**Urology**” department according to the total existing events. This model is extracted by the **classic heuristic miner**. Similarly to the previous example, **it is not clear** which level of information depicts the common behavior. In order to mine the descriptive reference process model for this department, the stable heuristic miner algorithms were used. Figure 6.16 shows the main behavior of patients within the urology department only by considering stability for **activities**.

From the 14 activities mined by the classic heuristic miner, 10 of them are shown by the result of the first version of the stable heuristic miner algorithm. Table 6.3 summarizes this observation. In addition, table 6.4 presents in detail the result of applying stable heuristic miner V1.

Within these 10 activities, 8 of them are expressing stable behaviors and 2 of them are showing high instability in comparison with the total number of recorded behaviors.

However, the *statistical stability of edges are not addressed* here. To get the stability of both activities and edges, second version of the stable heuristic miner algorithms is used. Figure 6.17 illustrates the result.

By means of this algorithm **not only unstable activities, but unstable edges are discovered** as well. Now, by considering the “**Registration**” activity, it is possible to discern the source of instability in this department.

The high level of fluctuations caused by this activity is seen by the **edge** between “**Registration**” and “**Reception_ Waiting_ room**”. The value of this edges is 57 which is **higher than the normal value for all the registered behaviors**. As a result, it puts the process in **unstable situations**.

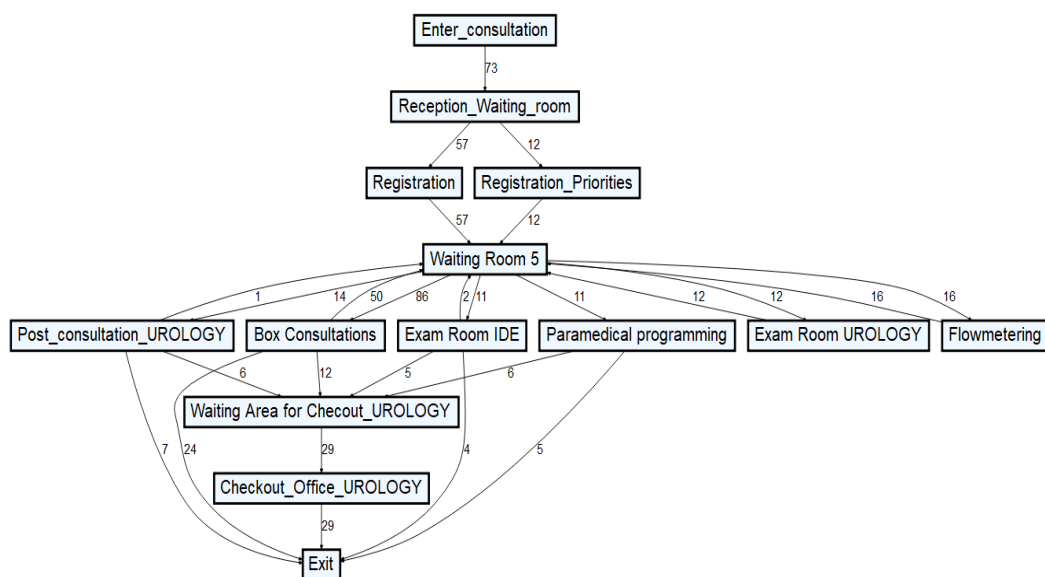


Figure 6.15 – The discovered model of the urology department by the classic heuristic miner.

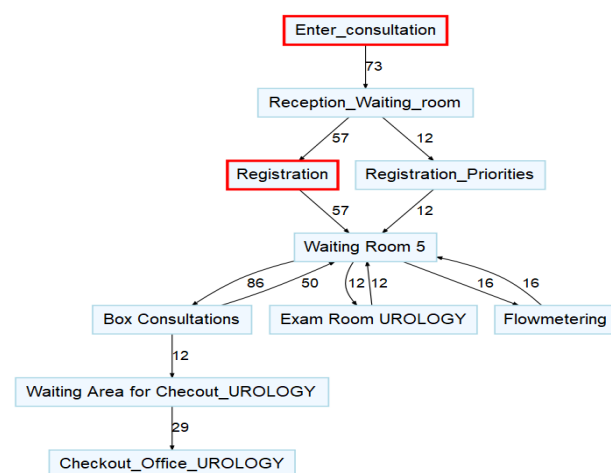


Figure 6.16 – The discovered model of the urology department by the stable heuristic miner V1. The Statistical stability method is applied only on activities behavior.

	Lower than LCL	In the stable State	Higher than UCL
Number of activities	4	8	2
Total number of modeled activities	10		
Total number of observed activities (appeared within event log)	14		

Table 6.3 – Comparing the number of observed behavior in the event log, with the result of stable heuristic miner V1.

All possible activities in the organization	Insignificant and unstable behavior (Lower than LCL)	Stable behavior $LCL < x < UCL$	High varying and unstable behavior (Higher than UCL)
Enter_consultation	—	—	Yes
Reception_Waiting_room	—	Yes	—
Registration	—	—	Yes
Registration_priorities	—	Yes	—
Waiting_room5	—	Yes	—
Box Consultation	—	Yes	—
Exam Room UROLOGY	—	Yes	—
Flowmetering	—	Yes	—
Waiting area for checkout_UROLOGY	—	Yes	—
Checkout_Office_UROLOGY	—	Yes	—
Post_consultation	Yes	—	—
Post_consultation_UROLOGY	Yes	—	—
Exam Room IDE	Yes	—	—
Paramedical programming	Yes	—	—
Exit	Yes	—	—

Table 6.4 – This table presents the total number of activities seen in the event log of the UROLOGY department. Also, it presents the result of applying **stable heuristic miner algorithm V1** for evaluating the statistical stability of activities.

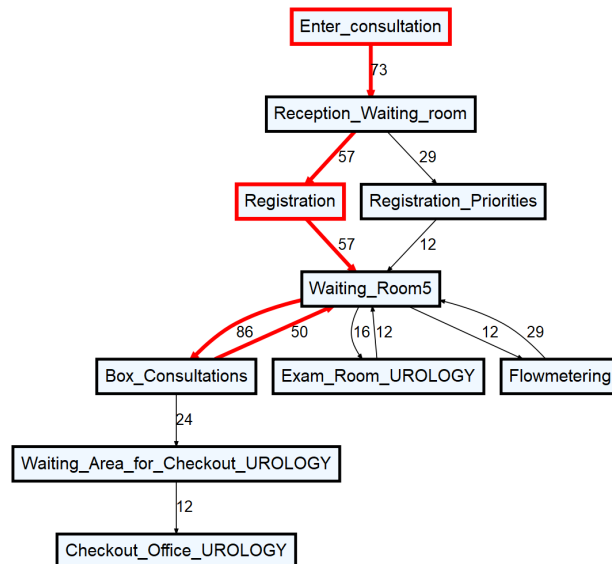


Figure 6.17 – The discovered model of the urology department by the stable heuristic miner V2. The Statistical stability methods are applied for both activities and edges behavior.

The same behavior causes the instability for edges between “Waiting_Room5” and “Box Consultation” activities.

Since, these deviating behaviors are mined and highlighted, the domain experts must focus on **stabilizing these activities and edges** to ensure that they will observe any unexpected behavior in patient pathways.

Acquiring such diagnosis is only feasible if experts are sure that the extracted model does indeed show the descriptive reference process model of patients within this department. **This was not a possible outcome of the previous process mining algorithms.**

Moreover, the application of stable heuristic miner algorithms helped experts **to detect deviating behaviors automatically and to capture an image of what patients do normally**, even if the experts did not particularly have complete knowledge about the process.

Previously, it was a necessity for the domain experts to know the process by heart. This was due to the reason that **the discovered model must have been filtered manually**. And this filtration should have happened *by low level of uncertainty*.

Such a process knowledge will not be required anymore while applying the novel stable heuristic miner algorithms.

6.5 Analyzing the quality and performance of patient pathways

This section will probe the applicability of the presented methods of the **business process analyzing function** in the context of the already known case study.

For this matter, the data related to the patient pathways within the “urology” department is used.

After receiving the data, “distance” and “process duration” were chosen as the **numeric value** for measuring the **reliability** of the processes (Note that the reliability is a quality characteristic, try to find their relations within the DIAG meta-model).

At first, a basic statistical analysis is performed considering these two numeric values. Figure 6.18 presents the duration of processes within the urology department for each individual case. In this figure, the size of each observation corresponding to the number of events is registered for that particular process. As an example to interpret this scatter plot, consider $ID = 238$, this case has finished its process approximately in 60 minutes. Also, it can be inferred that this process consisted of 5 events.

At the far right side of this figure, certain cases either have uncompleted processes or unexpected behaviors. Note that, due to the numerous number of cases, some of the ID numbers are missing on the x axis for better visualization. However, they are modeled in the scatter plot.

Similarly, figure 6.19 presents the distance taken by each patient to finalize his or her process.

Moreover, figure 6.20 illustrates the behavior of each case by considering two dimensions; *process duration* and *distance*. Despite the fact that these presented plots provide important information about the execution of processes, **they fail at providing rigorous analyses about the performance and the quality of processes.**

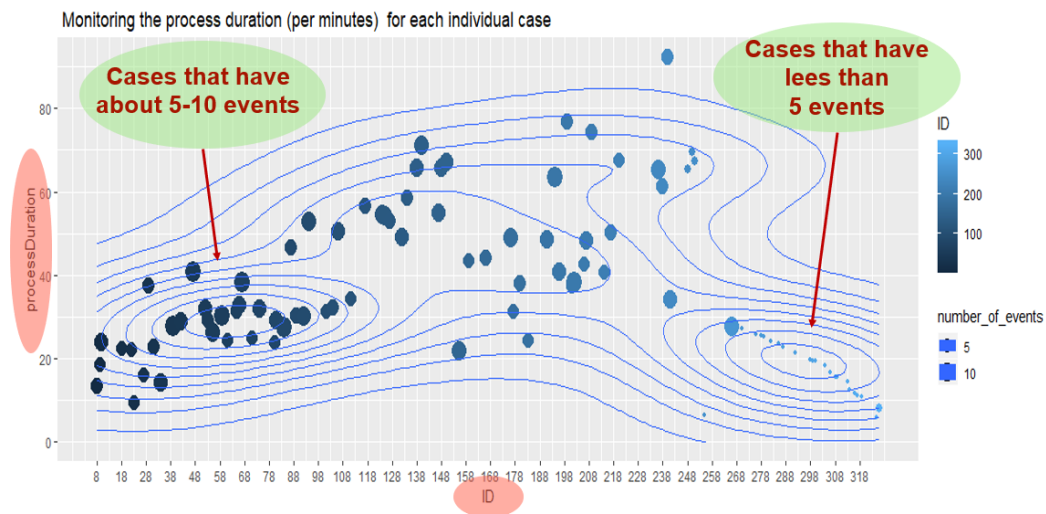


Figure 6.18 – This figure shows the duration of patient pathways for each individual case. The size of each observation changes based on the number of events that were registered for that particular process.

Therefore, **to address this issue**, the statistical process control methods presented in this research work are exploited.

Primarily, the duration of processes is monitored. As presented in chapter 5, several steps were taken to prepare the data. Related to this, 10 samples with a unique size of 10 were extracted (readers might find a difference between this number and the presented frequencies in the process model, the reason is that only patients with completed processes were modeled in the descriptive process models).

Then, the R -chart was applied to ensure about the **in-control status** of the duration range (c.f. figure 6.21). As shown, the range in which the process duration changes is in-control. Thus, by the next action, the \bar{x} -chart analysis was applied.

As presented in figure 6.22, the average duration of each sample is expressing a **normal** and **in-control** behavior. This implies that the process is **reliable** by considering the duration of the processes as the *numeric value* for evaluation.

This explanation for the duration of processes is comparable for the distance of processes as well. By considering figure 6.23, and 6.24 one can observe that the process is indeed in-control statistically and there is no sample with a highly fluctuating behavior.

Statistically, this process is in-control. But is this process efficient? Does this process meet the expectations of the patients and domain experts?

To answer these questions another complementary method is used as well. For this reason, the pursuing paragraphs present the application of **process capability ratio** analysis.

Within the presented control charts, there are samples whose behavior are showing a bit variation from the **central line** (most stable and normal behavior). These behaviors might need further analyses.

Accordingly, figure 6.25 analyzes the capability ratio of the process related to the **duration** of all of the recorded cases. Although, the process is statistically in control, but, **it is not efficient**.

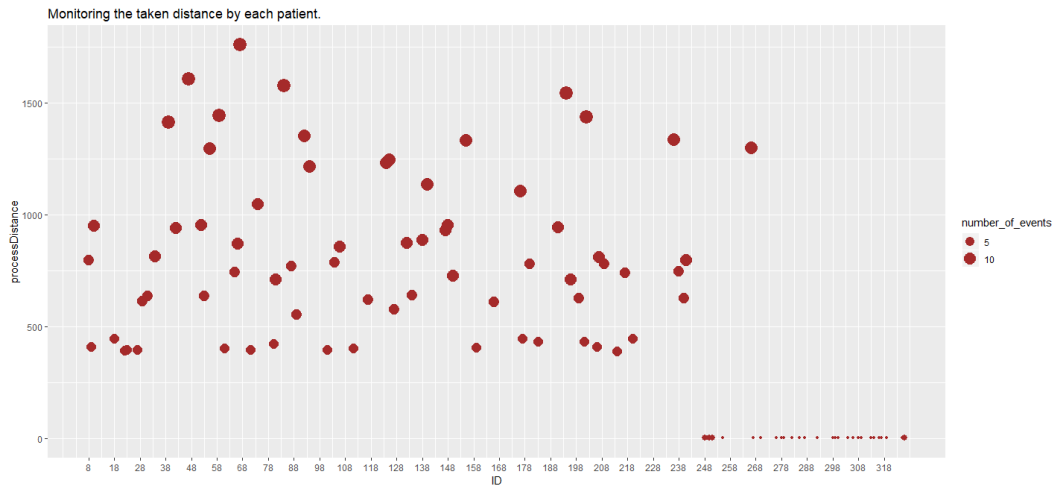


Figure 6.19 – This figure shows the distance of patient pathways for each individual case. In addition, it provides a density analysis as well.

As shown, the **lower specification limit** for the duration of the process defined by the domain experts and is about 16 minutes. However, this limit was not respected by 12% of cases. The C_p value is lower than 1 which **indicates the inefficiency of the process**. That being said, the process is very efficient in respecting the **upper specification limit (USL)**. Only 1% of observations are fallen outside of this limit, but, this process fails at respecting both specification values.

Likewise, the process is not efficient by considering the **distance** as the evaluating metric. As presented in figure 6.26, around 27% of the observation fall below the **lower specification limit (LSL)**. This can imply the existence of several cases who did not finish their processes. Again, the process is highly efficient in respecting the **USL** ($C_{pu} > 1$), but not so reliable close to the lower specification limit ($C_{pl} < 1$).

Control chart analyses are suitable for detecting the variations in the data; however, just expressing a statistically stable behavior is not a sufficient metric for evaluating the performance of the processes. Thereby, these behaviors must be evaluated by certain specification limits. With this objective, the **process capability ratio analysis** permits for such evaluation.

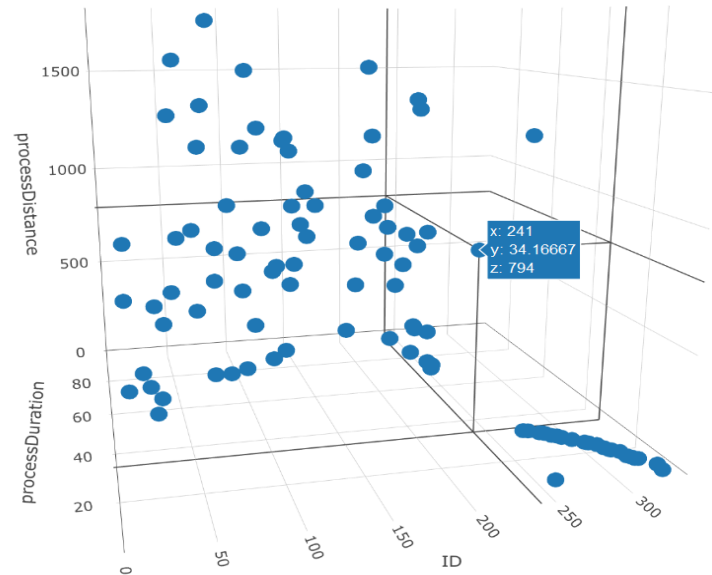


Figure 6.20 – Visualizing patients pathways behaviors by considering the distance and duration of each case. y axis is related to the process duration and z axis is presenting the process distance.

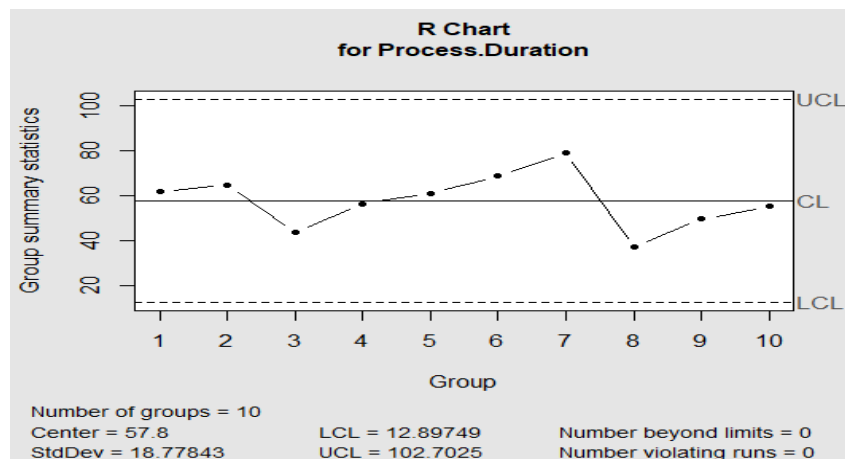


Figure 6.21 – The R -chart monitoring the **in/outOf-control** status of the process according to the range in which duration of processes change.

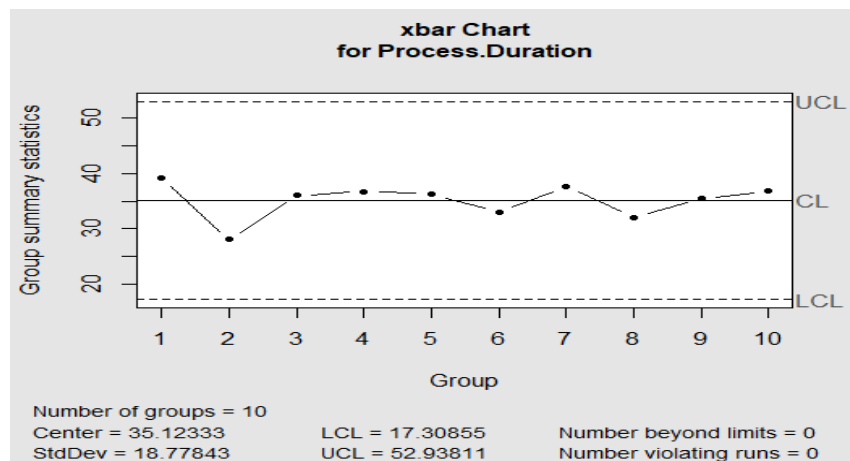


Figure 6.22 - The \bar{x} -chart monitoring the **in/outOf-control** status of the process according to the average duration of each sample.

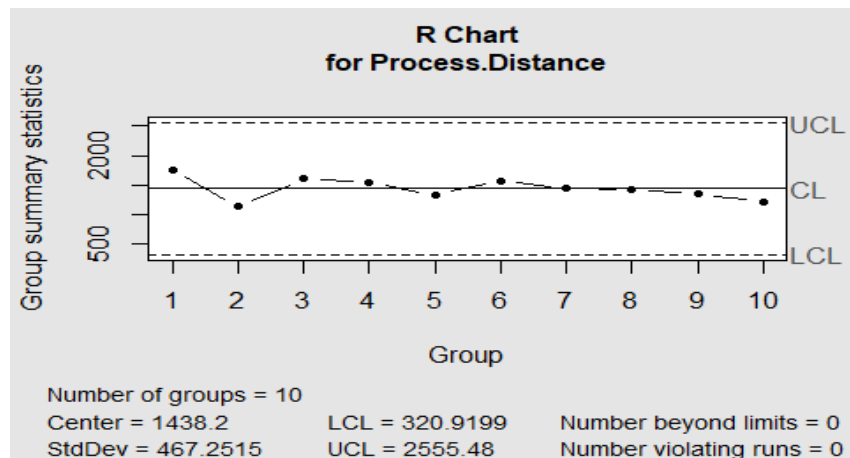


Figure 6.23 - The R -chart monitoring the **in/outOf-control** status of the process according to the range in which distance of processes change.

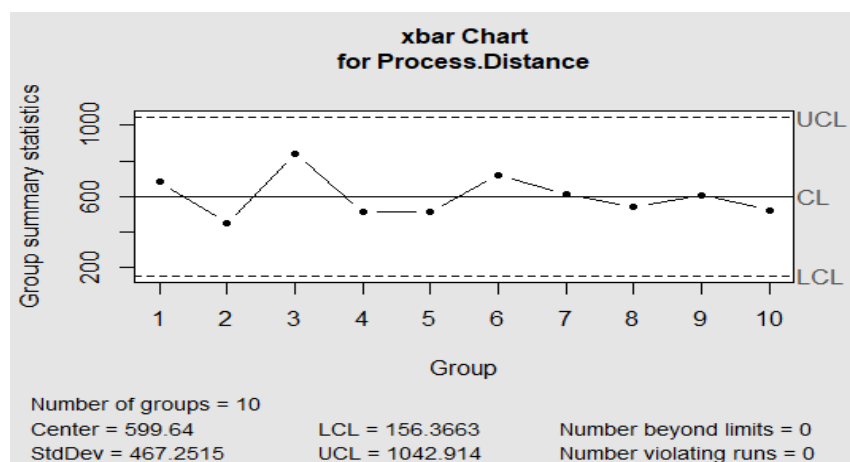


Figure 6.24 - The \bar{x} -chart monitoring the **in/outOf-control** status of the process according to the average distance of each sample.

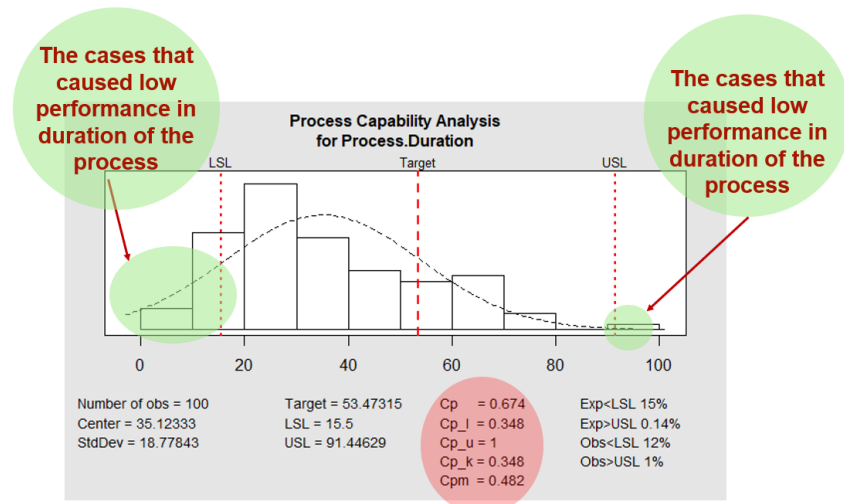


Figure 6.25 – Evaluating the performance of the process by analyzing duration of processes. This evaluation is based on the specification limits and the behavior of the modeled data according to these specification limits.

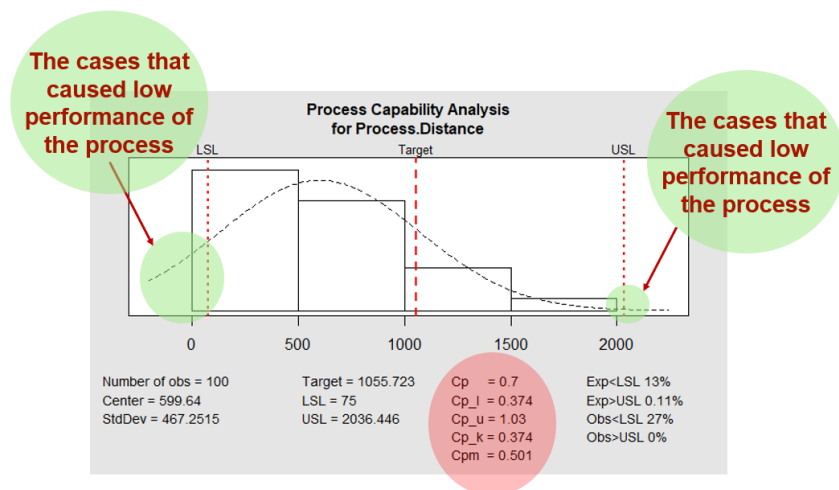


Figure 6.26 – Evaluating the performance of the process by analyzing distance of processes.

6.6 Automatic diagnosing of patient pathways

In this section the two introduced process diagnostic methods are applied to diagnose the **structural deviations** of the processes.

6.6.1 Miniscule Movements of Processes (MMP) method

The MMP method seems more applicable for diagnosing deviations of individual case pathways. It should be noted, that the objective of this method is on diagnosing the structural deviations in the process.

As an example, the process model for the patient whose process duration was higher than *USL* is considered (c.f. figure 6.25). This duration is related to the patient with the *ID* = 240. The duration of process was 92.3 which is higher than the *USL*. The process model DM_{240} includes this trace of activities:

$DM_{240} = \langle \text{"Enter_ Consultation", "Reception_ Waiting_ room", "Registration", "Waiting_ room5", "Box_ consultation", "Waiting_ room5", "Post_ consultation", "Exit"} \rangle$

For the process of the urology department a normative model is determined by the domain experts. This model is devised by *considering* the common behaviors in figure 6.17. Therefore, *RM* is presented by the following trace:

$RM = \langle \text{"Enter_ Consultation", "Reception_ Waiting_ room", "Registration", "Waiting_ room5", "Box_ consultation", "Post_ consultation", "Checkout_ Office_ UROL- OGY", "Exit"} \rangle$

The **MMP** method needs several requirements to initialize diagnosing DM_{240} . The **potential assignable causes** and corresponding **facts** are presented by the domain experts. Recall table 6.1, where the required information were defined.

Consequently, a subset of this information will be considered for the normative model of the urology department (*RM*). This subset is similar to table 6.5.

The first step for the application is to generate the fake processes which are carrying the potential assignable causes. According to table 6.5, there are 12 potential actions to take on 8 activities. Therefore, by combining all the possibilities, 3790 process models will be generated. This is a large number in comparison with the small number of defined **deviations**. Now, the objective is to find a process model in this set which has the minimum distance from DM_{240} .

To do so, the **ProDIST algorithm** is used. As shown in table 6.7, the generated model $GM36$ has the minimum distance from the patient pathway (DM_{240}). Therefore, pursuant to the primary information given by the domain expert, the extracted patient pathway

All possible activities in the organization	Type	PAC (defined by user)	Facts impact actions (defined by user)
Enter_consultation	Adjusting	NULL	NULL
Reception_Waiting_room	Adjusting	Rules and proceure	Remove
Registration	Administrative	Rules and proceure	Remove
Registration	Administrative	Human related	Add
Waiting_room5	Adjusting	Rules and proceure	Reception_Waiting_room
Box Consultation	Therapeutic	Human related	Remove
Box Consultation	Therapeutic	Environmental	Add_Waiting_room 5
Checkout_Office_UROLOGY	Administrative	Rules and procedure	Add
Checkout_Office_UROLOGY	Administrative	Human related	Reception_Waiting_room
Post_consultation	Medical Analyzing	Rules and procedure	Remove
Post_consultation	Medical Analyzing	Human related	Add_Waiting_room 5
Post_consultation	Medical Analyzing	Equipment	Add_Waiting_room 5
Exit	Adjusting	Rules and procedure	Remove

Table 6.5 – The used information for generating fake process models from the normative model (RM).

Generated Models (GM)	Trace of Activities
GM1	Enter_Consultation, Registration, Waiting_room5, Box_consultation, Post_consultation, Checkout_Office_UROLOGY, Exit
GM2	Enter_Consultation, Reception_Waiting_room, Registration, Reception_Waiting_room, Waiting_room5, Box_consultation, Post_consultation, Checkout_Office_UROLOGY, Exit
...	...
GM9	Enter_Consultation, Waiting_room5, Box_consultation, Post_consultation, Checkout_Office_UROLOGY, Exit
...	...
GM36	Enter_Consultation, Reception_Waiting_room, Registration, Waiting_room5, Box_consultation, Waiting_room5, Post_consultation, Exit
...	...

Table 6.6 – The generated models by injecting potential assignable causes into certain activities of the normative model (RM).

Distance of the generated models from the descriptive model	PAC	Affected activity
ProDIST(DM, GM1) = 7	Rules and procedure	Reception_Waiting_room
ProDIST(DM, GM2) = 5	Human related	Registration
...
ProDIST(DM, GM9) = 7	Rules and procedure	Reception_Waiting_room
	&	&
	Rules and procedure	Registration
...
ProDIST(DM, GM36) = 0	Human related	Box_Consultation
	&	&
	Rules and procedure	Checkout_Office_UROLOGY
...

Table 6.7 – This table presents the final result of the MMP method where a process was detected with minimum distance from the descriptive model (DM_{240}). Since there were more than 3000 generated models, not all the processes are presented. This table shows the descriptive model has minimum distance from GM36. This could indicate the existence of same causes on DM_{240} .

(DM_{240}) is expressing deviating behaviors, due to the existence of two causes ("**Human related**" and "**Rules and procedure**") for two activities: "**Box_Consultation**" and "**Checkout_Office_UROLOGY**".

Thanks to this method, the domain experts are able to detect the causes of structural deviations in a patient pathway.

6.6.2 Application of DIAG method

As introduced in chapter 5, this method performs the business process diagnosis on top of a **descriptive reference process model**. Unlike the previous method, DIAG is not suitable for diagnosing individual cases. This is due to the use of **statistical methods**. Therefore, it is a good practice to apply the DIAG method for diagnosing the common patient pathways.

To employ this method, the domain knowledge is needed. The volume of the input information depends on the decision of the domain expert.

For this specific study, a knowledge set is provided similar to table 6.8. Evidently, by changing the method from MMP to DIAG, the way to provide this knowledge set changes.

Contrary to the previous method, the user should not define the 3 impact actions; *add*, *remove*, *replace*.

Here, for the DIAG method, user must consider the relationships among activities and mention *what will be the deviation if a certain cause is present*.

After receiving the event log and the knowledge set, **DIAG method** will discover the descriptive reference process model and the causes of deviations for unexpected behaviors.

The result of applying this method is represented in figure 6.27. This model presents the corresponding diagnosis for the urology department.

As demonstrated in this model, there are three types of behaviors:

- **Stable activities and edges**: these behaviors are shown in black. They are presenting the most common and normal behaviors.

Activity	Deviation	PAC
Enter_consultation	Box Consultation	Rules and procedure
Reception_Waiting_room	Registration_Priorities	Rules and procedure
Registration	Reception_Waiting_room	Rules and procedure
Registration	Paramedical programming	Human related
Waiting_room 5	Registration	Rules and procedure
Waiting_room 5	Exam Room UROLOGY	Rules and procedure
Box_Consultation	Waiting_room 5	Human related
Box_Consultation	Registration	Environmental
Checkout_Office_UROLOGY	Registration	Rules and procedure
Paramedical programming	Exit	Equipment
Flowmetering	Waiting_room 5	Human related
Post_consultation	Box_Consultation	Rules and procedure
Post_consultation	Waiting_room 5	Human related
Post_consultation	Exit	Equipment
Exit	Checkout_Office_UROLOGY	Rules and procedure

Table 6.8 – The knowledge set provided by the domain expert to be used by DIAG method.

- **Activities and edges with high variations:** these behaviors are shown in **red**. They correspond to observations with a higher level of variations than the upper control limit of the stability state.
- **Deviations:** these behaviors are represented by activities modeled in **green**, and **dashed edges**. They are illustrating abnormal behaviors recorded in the event log.

Moreover, the causes of deviations among activities are presented beside deviating edges. If in the knowledge set, information does not correspond to the extracted deviations, then, a **0 value** is given to those deviations.

Consequently, it can be inferred that some cases in the “**Exam Room IDE**” finished their processes because of a problem related to the “**Equipments**”. Similarly, because of “**Human related**” errors, certain patients went back to “**Waiting Room 5**”.

Thanks to this method, the user not only is able **to discover patient pathways automatically**, but is possible to visualize unexpected and deviating behaviors and **diagnose them automatically**.

Moreover, it is possible to instantiate such causes to understand in more detail what type of equipment caused a deviation. However, this matter does not influence the applicability of the algorithm. This instantiation depends on the level of detail that the domain expert wants to provide.

To best of our knowledge, attaining such results has never been addressed by previous process mining activities. In addition, previous process discovery algorithms **were not able to determine deviating behaviors** in an event log nor **diagnosing them automatically**.

Traditional process mining activities only addressed *process discovery*, *enhancement of business processes* and *conformance checking*. The presented methods in this chapter can be seen as a milestone, since it targets the **automatic business process diagnosing**.

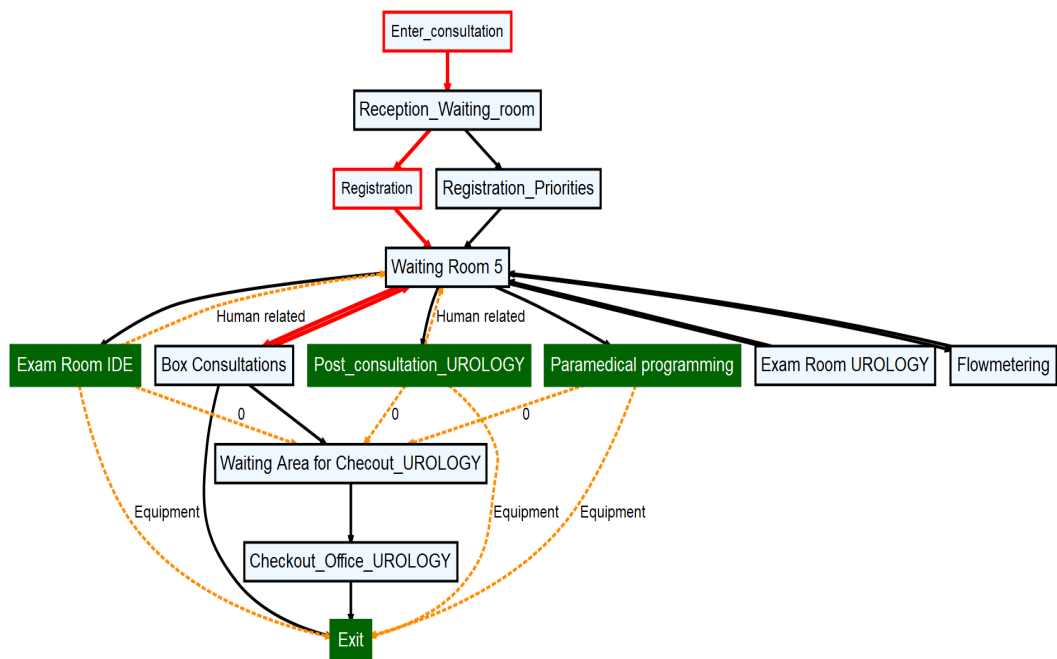


Figure 6.27 – The discovered model of the urology department which indicates the causes of deviations. The activities and edges with high variations are shown by a red color. The dashed edges show the **unexpected and deviating behaviors** among activities of the patient pathways. The causes of these deviations are indicated on each edge. The activities that are shown in green are unexpected activities that do not correspond to the common behavior of patients.

Conclusion

The work presented in this dissertation addresses a social question about “**how to improve patient pathways?**”

These pathways –as a subclass of care pathways– can express several information that can be used for the improvement procedure.

Traditionally, patient pathways were (and are) monitored by interview-based approaches Eisenthal et al. (1983) and Litchfield et al. (2018). Needless to say, this approach is extremely time-consuming. On top of interviewing staff, the patients should be questioned. It could be expected that the extracted information from each interview can be in contradiction with another one. This is due to the fact that each person can perceive a process differently.

To overpass such a difficulty, this research work suggests to use an indoor technological solution. The applied indoor localization system permits for a real-time tracking of patients activities. As a result, there would not be required for patients to participate in long interviews. After all, participation in interviews could be the last thing that an anxious patient looks for.

However, the ILS solution by itself can not be adequate to address the process improvement issue. It can offer significant data about the execution of processes, but, this data must be treated and mined prior to any improvement action.

As shown in figure 6.28, this research work conducted a literature review about the usage of location data in different industries during the last 5 years; consequently, it has been highlighted that the **interpretation and sense-making procedures** of location data did not receive sufficient attention.

The lion’s share of the research in this area proposed different methods and case studies for **extracting knowledge** from location data.

In addition to this, a formal data science-oriented procedure was missing. Such a procedure should indicate all the necessary steps from gathering the data until the knowledge extraction.

As a result of this study of the literature, the **DIAG methodology** has been introduced in this thesis. This methodology constructed by four different phases; *data*, *information*, *awareness*, and *governance*.

Each state has one or multiple *functions*. The first three functions of this methodology addressed the location data interpretation issue. Within the first function a **meta-model**

was devised to prepare a solid foundation prior to the **location data interpreting function**.

The DIAG meta-model allows for modeling of the resources, functions and other concepts that are involved in patient pathways.

This facilitates the sense-making procedure. As a consequence the location data interpretation rules would be capable to link the data registered in event logs with the real concepts within patient pathways (process actor, resources, and etc.). Chapter 3 illustrated these functionality.

Another experienced issue was related to the extraction of the **descriptive reference process model**. This model would express the **common pathway** which is *normally* taken by patients.

The classic heuristic miner algorithm showed promising results while dealing with healthcare data Rojas et al., 2016. However, like many other process discovery algorithms, it uses an arbitrary threshold to filter some behaviors from the model. The objective of this filtration was to present a level of information which is analyzable by the experts. Meanwhile, there is no ascertaining logic behind this action to indicate the discovered model is the descriptive reference process model. As a result, heuristic miner algorithm fails at extracting the “common pathways”.

This research work evoked the **statistical stability phenomenon** to introduce two novel algorithms which are capable to automatically extract the descriptive reference process model. These algorithms are **stable heuristic miner V1 and V2**. Chapter 4 described the development of these algorithms.

These algorithms targeted the well-known structural problem of the family of heuristic mining algorithms which was related to the arbitrary selection of thresholds values De Cnudde et al., 2014.

According to the provided analysis of the literature in chapter 2, most of the data and process mining methods tackled the pattern and process discovery topic.

Therefore, there exist a lack of quantitative and qualitative methods that can target the issue of **enhancing business processes**. Moreover, the existing work in the literature did not directly address **business process diagnosing**.

As defined by Merriam-Webster dictionary, **diagnosing** is an action to *recognize cause or nature of a potential problem*. And, most of current works in the literature of process mining addressed this issue only by providing a basic statistical analysis of the AS-IS situation of the process.

In substance, the previous methods fail at detecting the causes of deviations and inefficiencies in processes.

In view of this problem, the present dissertation invested heavily in chapter 5 on exploring proper solutions to **measure and evaluate the quality and performance of patients pathways and eventually the automatic diagnosis of processes**.

Consequently, the application of **statistical process control** offered a substantial contribution to the application of process mining in healthcare sector. The results of such an application were appreciated by the healthcare community Araghi et al., 2018b, **araghi_Evaluating_2019**, Namaki Araghi et al., 2018.

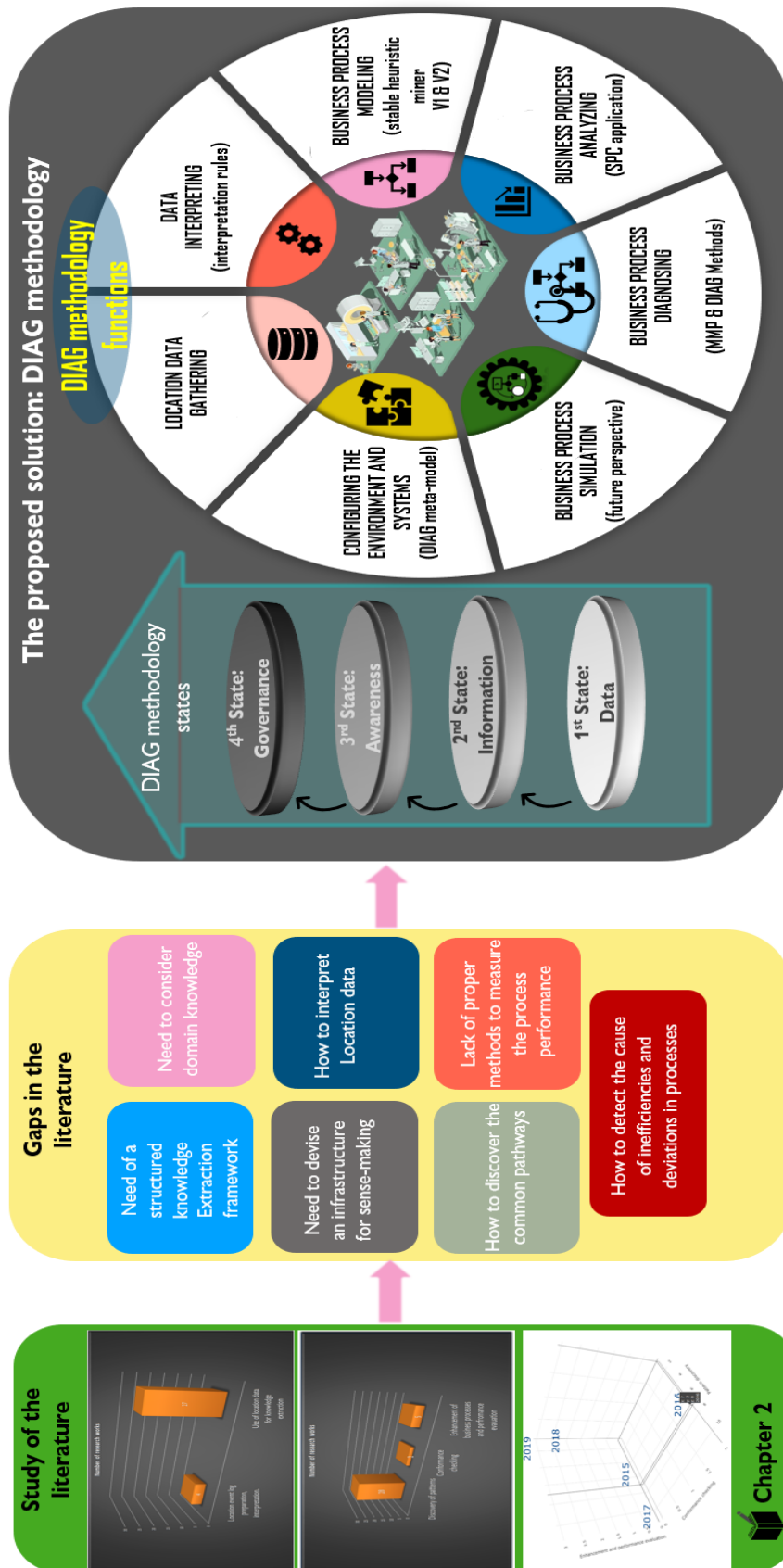


Figure 6.28 – A summary of the gaps found in the literature (presented in chapter2) and the proposed solution in this research work.

Relevant to the automatic business process diagnosing topic, this research work demonstrated two new methods. The **MMP (Miniscule Movements of Processes)** was introduced as a method to diagnose deviations in the process of each individual case in comparison with the normative process. Within this method the **ProDIST algorithm** was presented to measure the distance of two processes from each other.

In addition, the **DIAG** method employed and enhanced the second version of the stable heuristic miner algorithm to automatically diagnose the deviations within the “common pathways” of patients.

The DIAG method is a turning point for the application of process mining; thanks to this method, the domain expert not only is able to extract the descriptive reference process model, but he/she would obtain an **automatic diagnosis** on top of the model as well.

6.7 Limitations

There is also a need to highlight the shortcomings of the presented research work. Similar to most process mining applications, the proposed methods here are also highly dependent on the quality, accuracy and reliability of the data.

For instance, in a location event log, some cases may have disruptions in their data. The presented methods in this dissertation filters those cases which do not have complete data. This could be considered as a limitation since sometimes the root of a problem in processes may be hidden within the information of such cases.

In addition, domain knowledge is an extremely important requirement. When initializing such analyses, a “warm-up” period should be considered in order to identify what activities may occur in each zone of the organization.

Moreover, the used indoor localization system here assumes that the existence of a tag in each of the zones must correlate with the execution of a certain activity. For instance, if a patient is inside the physician’s office, it means that the patient is being treated by a physician. However, in reality this may not be the case. As an example, a patient could be waiting for the physician until he or she comes to the room.

In this case, one could consider the cross-matching of the location data of the patient and the physician. Although this may be considered as a logical solution, there exists a significant social challenge, in that hospital staff are not eager to be monitored by the localization systems. These ethical challenges need to be addressed when tracking patients and hospital staff.

Additionally, in some cases several activities occur in the same zone. This could make it more challenging to interpret the events.

The stable heuristic miner algorithm needs to be improved so that it can discover the decision points from the location data. This is a major challenge due to the lack of sufficient information in location event logs. It could be achieved by integrating other information from the hospital information system.

6.8 Future perspective

Future works may include integrating hospital information systems with location data for extracting the decision points that could be a valid target for future research work. This would help to extract an executable semantic which could be used for the simulation of patients’ pathways as business process models.

This research work highly encourages researchers to investigate these issues:

1. DIAG meta-model is an evolving and a trustworthy base for the application of business process management in the healthcare sector. Researchers are mostly welcomed to participate for its development which is an on-going research in the Industrial Engineering Center of IMT Mines Albi.
2. Proposing an **overall evaluation approach for process discovery algorithms**. Conformance checking methods seem to be used in many works by considering *precision*, *generality*, *simplicity*, and *fitness* criteria. But as it has been seen in Augusto et al., **2019a** by using these criteria one is not able to indicate what could be a proper trade-off value among these criteria.
3. Another issue is the **extraction of decision points and the discovering of guards** by integrating location data and other supplementary information. This would help for simulation of models which are considered as “prognostic” actions within the “governance” state of DIAG methodology.
4. **Automatic business process diagnosing**, this issue has been completely neglected in the literature, whereas one can not improve a process without knowing the cause of inefficiencies.
5. Last but not least, **the social challenge** must be addressed in future works. The technological and scientific advancements for monitoring healthcare processes would become unavailing if the staff and patients are not eager to participate for similar experiments or to use the ILS technology.

Résumé étendu en français

Dans chaque organisation, les processus métier sont aujourd'hui incontournables. Cette thèse vise à développer une méthode pour les améliorer. Dans le domaine de la santé, les organisations hospitalières déploient beaucoup d'efforts pour mettre leurs processus sous contrôle, notamment à cause de la très faible marge d'erreur admise. Les parcours des patients au sein des structures de santé constituent l'application qui a été choisie pour démontrer les apports de cette méthode. Elle a pour originalité d'exploiter les données de géolocalisation des patients à l'intérieur de ces structures. Baptisée DIAG, elle améliore les parcours de soins grâce à plusieurs sous-fonctions :

(i) interpréter les données de géolocalisation pour la modélisation de processus, (ii) découvrir automatiquement les processus métier, (iii) évaluer la qualité et la performance des parcours et (iv) diagnostiquer automatiquement les problèmes de performance des processus. Cette thèse propose donc les contributions suivantes :

1. la méthode DIAG elle-même qui, grâce à quatre différents états, extrait les informations des données de géolocalisation ;
2. le méta-modèle DIAG qui a deux utilités : d'une part, interpréter les données de géolocalisation et donc passer des données brutes aux informations utilisables, et, d'autre part contribuer à vérifier l'alignement des données avec le domaine grâce à deux méthodes de diagnostic décrites plus bas ;
3. deux algorithmes de découverte de processus qui utilisent la stabilité statistique des logs d'événements ;
4. une nouvelle approche de process mining utilisant SPC (Statistical Process Control) pour l'amélioration ;
5. l'algorithme proDIST qui mesure les distances entre les modèles de processus ;
6. deux méthodes de diagnostic automatique de processus pour détecter les causes des déviations structurelles dans des cas individuels et pour des processus communs.

Le contexte de cette thèse confirme la nécessité de proposer de telles solutions. Une étude de cas dans le cadre de ce travail de recherche illustre l'applicabilité de la méthodologie DIAG et des fonctions et méthodes mentionnées.

Keywords: Process Mining, Systèmes de localisation en intérieur, Gestion des processus métiers, Diagnostic des processus métiers, Processus de soins

Contexte et problématique sociales pertinentes pour les parcours des patients

Challenge No. 1 - “Structure et organisation”

Chaque année au Royaume-Uni, environ 6,9 millions de rendez-vous à l'hôpital en ambulatoire sont manqués et chaque rendez-vous manqué coûte environ 108 £. Les nombreux articles soulignent le fait que les patients se perdent souvent dans les hôpitaux et ce problème peut entraîner de nombreuses complications, tels que des rendez-vous manqués.

Ce problème coûteux a également été mentionné dans d'autres travaux de recherche. Il a été mentionné précédemment que le système de santé américain est affecté par un coût annuel de 150 milliards \$ de rendez-vous manqués appelés “no-shows”. Ils ont mis en évidence plusieurs causes telles que **problèmes structurels ou organisationnels** ainsi que des **erreurs humaines**. Ce travail de recherche se concentre sur les *problèmes structurels* et non sur les erreurs humaines.

Fondamentalement, il considère que les rendez-vous manqués sont dus aux difficultés rencontrées par les patients pour se repérer dans les hôpitaux, avec des kilomètres de couloirs presque identiques menant d'une porte à l'autre avec des noms similaires. Il est clairement nécessaire de réduire ces coûts de défauts coûteux.

Des travaux de recherche antérieurs ont fait référence à ce problème en réalisant des rapports sur les placements de patients dans différents hôpitaux du monde: Royaume-Uni, Espagne, Nouvelle-Zélande et France.

Les déviations dans la structure des parcours des patients est un défi commun pour les organisations de santé. S'il était possible pour un hôpital d'acquérir une vue d'ensemble des **voies communes** qui montrent les **itinéraires habituels empruntés par les patients** lors de leurs déplacements dans l'hôpital et dans la conduite de leurs activités, les experts en soins de santé seraient en mesure de détecter les variations dans les processus des patients. Il suffirait de comparer le modèle actuel en l'état avec un modèle normatif défini par des experts qui souhaitent optimiser le fonctionnement des hôpitaux. Ainsi, une telle vision pourrait aider à éviter les comportements inefficaces dans ces processus, ou comme on les appelle ici, ces **parcours du patient**.

De nombreux travaux de recherche authentiques ont abordé le sujet des parcours des patients. Ce travail considère le cheminement du patient comme le chemin parcouru par un patient depuis le premier contact avec le personnel hospitalier à travers des tâches telles que la consultation sur ses problèmes de santé, à différentes actions diagnostiques et administratives à l'intérieur de l'établissement de santé jusqu'à ce qu'il quitte les lieux. Ces parcours de patients sont une sous-classe de parcours de soins et peuvent être considérés comme **processus métier**.

Pour approfondir cette affirmation, la définition des parcours de soins doit être revue, où les auteurs définissent les parcours de soins comme une “intervention complexe pour la prise de décision mutuelle et l'organisation des processus de soins pour un groupe bien défini de patients pendant une période bien définie”.

De plus, il a été démontré que le but de ces parcours de soins est d'améliorer la qualité des soins à travers le continuum en améliorant les résultats des patients, ajustés au risque, en promouvant la sécurité des patients, en augmentant la satisfaction des patients et en optimisant l'utilisation des ressources. Sur la base de ces définitions et

des méthodes d'analyse appliquées dans ce travail de recherche, on pourrait en déduire que les parcours des patients peuvent également être classés comme une sous-catégorie **parcours de soins**.

D'un autre côté, chaque cheminement du patient se compose de séquences d'*événements*, de plusieurs *étapes*, *points de décision*, *acteurs* et *activités* avec l'objectif de fournir des soins de santé au patient. Ce sont les éléments mutuels avec la définition de **processus métier** fournie par Dumas. Par conséquent, on pourrait concevoir ces voies comme des **modèles de processus métier**. Comme Dumas l'a mentionné, un processus opérationnel peut être considéré comme un moyen utilisé par les organisations pour fournir un service ou un produit aux clients.

Cela étant dit, pour répondre au problème structurel, trouver les voies communes à partir de toutes les voies du patient est une tâche principale. La voie commune permet de sensibiliser les organisations de santé aux processus et activités des patients. Ce travail de recherche présentera une nouvelle approche pour capturer un **modèle de processus de référence descriptif** qui pourrait servir de solution.

L'extraction de modèles descriptifs —parcours de patients —et les parcours de patients communs seront abordés dans ce travail de recherche en utilisant des données de localisation à l'intérieur et des activités d'exploration de processus.

Challenge No. 2 - “ Durée et distance ”

Imaginez ce scénario:

- Vous avez un rendez-vous dans une clinique médicale et par défaut, la clinique vous a demandé d'être là *20 minutes avant le rendez-vous*. Donc, vous respectez cette règle et arrivez à l'heure. Cependant, en raison de certains problèmes, vous devez attendre *40 minutes* avant d'être pris en charge.

Dans ce scénario, qu'est-ce qui vous dérange? Le problème pour vous n'est pas tellement le temps d'attente; vous avez respecté le délai de 20 minutes qu'ils ont demandé. En revanche, vous êtes agacé par le fait que vous ne savez pas à quoi vous attendre et que vous avez perdu 20 minutes de votre temps. Devez-vous respecter le délai de 20 minutes pour le deuxième rendez-vous ou non?

Par conséquent, dans ce cas, **l'incertitude des résultats du processus** rend le patient insatisfait et confus.

Compte tenu de cette problématique, ce travail de recherche vise à proposer différentes méthodes pour **surveiller et détecter** les variations et l'incertitude des résultats du processus. Une telle approche réfléchit à l'amélioration de la qualité des parcours des patients.

Par conséquent, la question sociale est:

Au regard de ces enjeux évoqués, ce travail de recherche est conçu pour répondre à cette question sociale suivante:

Comment améliorer les parcours des patients?

Objet et portée de ce travail

Afin de répondre à la question ci-dessus, les experts en soins de santé ¹ ont besoin d'un plan d'action pour surveiller les activités des patients.

Selon Dumas, les experts de la santé peuvent choisir l'une des approches suivantes:

1. Interview-Based: Comme son nom l'indique, cette approche est basée sur des entretiens avec les patients et le personnel pour capturer une image de la situation AS-IS des processus. L'objectif est de modéliser l'information et de voir où sont les problèmes cachés.

Comme on peut l'imaginer, les résultats de cette approche peuvent comporter plusieurs incertitudes, car chaque acteur du processus (patient ou personnel) peut percevoir le processus différemment. De plus, c'est une approche qui prend beaucoup de temps.

2. Workshop-Based: Dans cette approche, l'expert du domaine et l'expert en modélisation des processus métier travailleront ensemble pour concevoir les modèles de processus. La conception de ces modèles peut faire l'objet de plusieurs réunions et séances de discussion. Semblable à un entretien, il peut s'agir d'une approche lente.
3. Evidence-Based: Cette approche émerge par l'évolution rapide de la discipline de recherche en exploration de processus. En appliquant cette approche, les experts passent en revue les preuves existantes (par exemple les systèmes d'information) dans les organisations et essaient d'extraire et de concevoir des modèles de processus. La modélisation peut se faire manuellement ou informatiquement. Cependant, l'activité de découverte de processus de l'exploration de processus peut être une solution rapide et précise pour les experts pour cartographier l'état des processus AS-IS.

Dans ce contexte, le présent travail de recherche évoque l'idée d'associer les champs de recherche **Indoor Localisation Systems (ILS)** et **Process Mining**.

L'ILS permet de suivre les mouvements des patients à l'intérieur des locaux. Cette solution technologique offre plusieurs **avantages**; tels que: (a) *précision* dans le suivi des activités des patients. (b) *prise de conscience en temps réel* des situations des patients. (c) *amélioration de la sécurité* des patients et du personnel.

Cependant, cet ILS peut fournir des ambiguïtés en raison de sa capacité à générer des ensembles de données volumineux. En fait, chaque balise d'emplacement peut émettre un événement par seconde. Il ne faut que quelques minutes pour rencontrer des milliers de lignes d'événements pour plusieurs patients.

Cela dit, l'extraction de processus est devenue une solution appropriée pour l'extraction d'informations significatives à partir des fichiers log d'événements de localisation.

Défis et questions scientifiques

Bien qu'une telle application semble intrigante, plusieurs **questions scientifiques** ont émergé au cours du développement de ce travail de recherche. Ce qui suit présentera ces problèmes.

¹Un domaine — expert en soins de santé — peut être une infirmière, un médecin, un directeur de service ou toute autre personne administrative chargée de fournir un ensemble d'informations principales sur l'environnement.

Quelle procédure orientée science des données convient pour extraire des données de localisation de formulaire de connaissances ?

Selon Provost, lorsqu'il s'agit d'un projet orienté données, il est extrêmement important d'identifier la **procédure d'exploration de données**.

Au tout début de ces travaux de recherche, le besoin d'une telle procédure formelle qui traitait de “ **extraction de connaissances** ” à partir des fichiers log d'événements de localisation.

Une telle procédure clarifie les étapes requises et leurs ordres pour extraire des connaissances des données. Par exemple, cela aide à comprendre quelle est la principale étape, le traitement des données ou la modélisation des processus.

La **méthodologie DIAG** a été introduite dans la deuxième partie de ce travail de recherche pour cibler cette exigence. Cette méthodologie sera exclusivement décrite dans la deuxième partie de cette thèse.

Comment interpréter les données de localisation ?

L'un des défis les plus importants était de fournir une base solide pour interpréter et enrichir les informations enregistrées dans les fichiers log d'événements de localisation. Par exemple, ces fichiers d'événements de localisation peuvent inclure des informations sur la température, l'humidité, le changement de position et les zones. Il est nécessaire d'interpréter ces données et de comprendre quel *événement représente une activité de processus*. Ce travail de recherche se réfère à cette exigence comme la **procédure de création de sens**.

Cette décision était principalement due au format de ces données. Le chapitre 3 expliquera ce problème.

Comment extraire un modèle de processus de référence descriptif ?

Comme cela sera discuté dans les chapitres 2 et 4, sur la base des méthodes précédentes dans la littérature de l'exploration de processus dans les soins de santé, il n'était pas possible d'extraire un modèle de processus qui indique la **voie commune** des patients.

Pour illustrer ce problème, imaginez le suivi de 100 patients à l'intérieur des locaux de l'hôpital. Ces patients pourraient avoir une voie unique s'ils suivent tous le modèle normatif. Cependant, en réalité, ce n'est pas le cas. Les patients peuvent suivre différentes voies et exécuter différentes activités.

En conséquence, il est difficile pour les experts du domaine de se référer à un certain processus —chemin du patient— comme chemin commun. La voie commune peut indiquer les activités qui sont normalement présentes dans n'importe quelle voie patient.

Les méthodes actuelles d'exploration de processus fournissent plusieurs modèles de processus et permettent aux experts de décider d'indiquer la voie commune. Par conséquent, cette décision et l'exploitation des processus sont complètement arbitraires.

Les méthodes présentées au chapitre 4 de cette thèse visent à surmonter ce défi.

Comment acquérir une approche analytique exhaustive pour les parcours des patients ?

D'après les publications sur l'analyse des processus métier, une approche analytique globale est construite sur deux piliers principaux: **analyses qualitatives** comme la visualisation des modèles de processus, et **analyses quantitatives**. Les analyses quantitatives doivent être approuvées par des méthodes mathématiques solides qui sont capables de détecter les **écarts** dans le résultat des processus et pas seulement de visualiser les méthodes statistiques de base.

Les méthodes quantitatives antérieures, telles que la visualisation de la durée des processus par des histogrammes, ne sont pas capables d'évaluer la performance des processus. Ces méthodes ne fournissent — encore une fois — que des mesures qualitatives. Cela est dû à leur incapacité à détecter les inefficacités et les écarts dans les processus.

Il s'agit d'une question cruciale qui sera étudiée dans le cinquième chapitre de ce travail de recherche.

Comment mesurer la distance d'un modèle descriptif par rapport au modèle normatif ?

Comme mentionné précédemment, la comparaison du modèle normatif avec le modèle descriptif peut être une approche utile pour évaluer les processus des patients. Pour ce faire, la distance d'un parcours patient extrait du parcours normatif est prise en compte.

Dans cette optique, ce travail de recherche introduit un algorithme capable de mesurer la distance entre les modèles de processus. L'algorithme ProDIST sera abordé au chapitre 5. De plus, il est utilisé pour effectuer un **diagnostic automatique de processus métier**.

Comment diagnostiquer automatiquement les causes des écarts ?

Dans les publications sur la gestion des processus métier, **BPA** (analyse des processus métier) traite de l'évaluation et de l'amélioration des processus; en outre, **activité d'amélioration** de l'exploration de processus se concentre également sur ce problème. Cependant, les méthodes actuelles ne parviennent pas à découvrir les causes des inefficacités.

Par conséquent, ce travail de recherche au chapitre 5 présente deux nouvelles méthodes pour effectuer un diagnostic automatique des processus métier.

Comment générer différents scénarios pour choisir la solution optimale ?

Après avoir détecté les causes de déviation, il est nécessaire de supprimer ces inefficacités pour améliorer le processus. En effet, une telle action embrasse la définition de l'amélioration continue et les objectifs de l'ingénierie qualité.

Pour ce faire, il faut s'assurer des effets de ces modifications sur les résultats du processus. Cela peut être défini comme **actions pronostiques**.

La simulation d'événements discrets est un candidat approprié pour accompagner les méthodes introduites de ce travail de recherche afin de proposer plusieurs solutions d'amélioration.

Comme mentionné par Van der Aalst, cela peut être " une rencontre au paradis ". Cette proposition est identifiée comme une perspective future de cette thèse, cependant, elle est définie comme la septième étape de la méthodologie DIAG.

Les contributions de la thèse

Conformément aux défis scientifiques mentionnés ci-dessus, cette section mentionne brièvement les solutions proposées pertinentes par cette thèse. Ces solutions sont définies comme contributions de ce travail. Ils sont répertoriés comme:

1. **DIAG methodology.**
2. **DIAG meta-model.**
3. **Les règles d'interprétation des données de localisation.**
4. **Stable Heuristic miner algorithms (V1 & V2).**
5. **Application du contrôle statistique des processus (SPC) pour l'activité d'amélioration de Process Mining.**
6. **ProDIST algorithm.**
7. **Miniscule Movements of Processes (MMP) method.**
8. **DIAG process diagnosing method.**

La première contribution est la méthodologie DIAG qui a un objectif principal; pour aider les experts du domaine dans leurs procédures de prise de décision basées sur les données (de localisation).

Cette contribution est motivée par cette question de recherche:

"Quelle procédure nous pourrions choisir pour extraire les connaissances des logs d'événements location?"

Cette méthodologie sera explorée dans la **deuxième partie** de cette thèse. Elle a quatre états principaux, **Data, Information, Awareness** et **Governance**. Chaque état comprend une ou plusieurs fonctions. Chaque fonction est définie pour répondre aux problèmes scientifiques mentionnés.

D'autres contributions énumérées de cette thèse sont définies dans les fonctions de cette méthodologie.

Par exemple, la première fonction présente le **méta-modèle DIAG** comme la deuxième contribution de la thèse. Les secondes fonctions concernent la collecte de données. La troisième fonction introduit les **règles d'interprétation des données de localisation**.

Ces contributions sont motivées par cette question de recherche:

"Comment interpréter les logs d'événements de localisation ?"

Le quatrième présente les deux **algorithmes de découverte de processus** pour extraire le modèle de processus de référence descriptif.

Ces contributions sont motivées par cette question de recherche:

"Comment extraire le comportement commun des patients ?"

La cinquième fonction se concentre sur **application de SPC** (Statistical Process Control) et l'application d'analyses quantitatives.

Cette contribution est motivée par cette question de recherche:

“Comment mesurer la distance d’un modèle descriptif par rapport au modèle normatif ?”

La sixième fonction introduit **deux méthodes (MMP ² et DIAG) pour détecter automatiquement les causes des écarts dans les modèles de processus.**

Ces contributions sont motivées par cette question de recherche:

“Comment diagnostiquer automatiquement les causes des écarts ?”

Enfin, la septième propose la simulation à événements discrets comme solution pour proposer de multiples scénarios d’amélioration.

Aperçu de la thèse

La **première partie** de cette thèse se concentre sur l’introduction du contexte et de l’état de l’art de ce travail de recherche.

Le deuxième chapitre présente l’état de l’art. Ce chapitre explore la littérature des disciplines de recherche les plus pertinentes pour ce travail: science des données, exploration de processus et application des données de localisation dans des projets d’exploration de données et de processus.

La **deuxième partie** de la thèse commence par introduire la **méthodologie DIAG**. Ensuite, le chapitre 3 se concentre sur la présentation des trois premières fonctions de la méthodologie DIAG. Elle montre principalement comment les données de localisation sont reçues et interprétées.

Le chapitre 4 présente des algorithmes de mineur heuristique stables (V1 et V2) pour extraire le modèle de processus de référence descriptif.

Le chapitre 5 traite de l’application de SPC, de l’algorithme ProDIST et des deux méthodes de diagnostic automatique des processus métier.

Enfin, dans la **troisième partie** de la thèse, le chapitre 6 présente une étude de cas, où les méthodes introduites de ce travail de recherche sont mises en œuvre.

La conclusion est l’élément final de la thèse.

Terminologie importante de cette thèse

Les explications suivantes clarifie clarifieront les termes utilisés dans cette thèse.

- Reference (or normative) model (RM):
Ce modèle est défini par le domaine — soins de santé — expert.
- Descriptive Model (DM):
Il peut être extrait des données par exploration de processus. Il représente une quantité aléatoire d’informations pour un modèle de type processus dans un fichier d’événements
- Descriptive Reference Process Model:
Ce modèle présente le processus commun et stable. Il est principalement utilisé dans le chapitre 4.
- Generated Model (GM):
Il s’agit d’un modèle généré en injectant des causes potentielles attribuables. Il sera traité au chapitre 5.

²MMP: mouvements minuscules des processus.

Revue internationale avec comité de sélection

En raison des travaux techniques exigeants au début de ma thèse, les principales contributions du travail sont apparues à la fin du projet. Un article est soumis à la revue Information Systems intitulé: "Stable Heuristic Miner- An algorithm to discover the common patient pathways". La deuxième version de l'algorithme a évolué après cette soumission et nous soumettons ce travail à la revue "Biomedical Informatics". Un algorithme pour diagnostiquer les causes des écarts dans les processus métier est préparé pour être soumis au numéro spécial "Conformance Checking" de la revue "Information Systems".

Articles soumis (ou en cours de préparation)

- Stable Heuristic Miner - An algorithm for discovering the common behaviors of patients from location event logs.
- Stable Heuristic Miner V2 algorithm for discovering statistically stable behaviors from event logs.
- Automatic business process diagnosing by the Minuscule Movements of Processes (MMP) method.

Les limites

Il est également nécessaire de souligner les lacunes des travaux de recherche présentés. Semblables à la plupart des applications d'extraction de processus, les méthodes proposées ici dépendent également fortement de la qualité, de l'exactitude et de la fiabilité des données.

Par exemple, dans un log des événements d'emplacement, certains cas peuvent avoir des perturbations dans leurs données. Les méthodes présentées dans cette thèse filtrent les cas qui ne disposent pas de données complètes. Cela pourrait être considéré comme une limitation car parfois la racine d'un problème dans les processus peut être cachée dans les informations de tels cas.

En outre, la connaissance du domaine est une exigence extrêmement importante. Lors de l'initialisation de ces analyses, une période de " warm-up " doit être envisagée afin d'identifier les activités susceptibles de se produire dans chaque zone de l'organisation.

De plus, le système de localisation en intérieur utilisé suppose ici que l'existence d'un tag dans chacune des zones doit être en corrélation avec l'exécution d'une certaine activité. Par exemple, si un patient se trouve à l'intérieur du cabinet du médecin, cela signifie que le patient est traité par un médecin. Cependant, en réalité, ce n'est peut-être pas le cas. Par exemple, un patient peut attendre le médecin jusqu'à ce qu'il ou elle vienne dans la chambre.

Dans ce cas, on pourrait envisager la mise en correspondance croisée des données de localisation du patient et du médecin. Bien que cela puisse être considéré comme une solution logique, il existe un défi social important, en ce que le personnel hospitalier n'a pas hâte d'être surveillé par les systèmes de localisation. Ces défis éthiques doivent être abordés lors du suivi des patients et du personnel hospitalier.

De plus, dans certains cas, plusieurs activités se produisent dans la même zone. Cela pourrait rendre plus difficile l'interprétation des événements.

L'algorithme de stable heuristic miner doit être amélioré afin qu'il puisse découvrir les points de décision à partir des données de localisation. Il s'agit d'un défi majeur en raison du manque d'informations suffisantes dans les logs d'événements de localisation. Cela pourrait être réalisé en intégrant d'autres informations du système d'information de l'hôpital.

Perspectives d'avenir

Les travaux futurs pourraient inclure l'intégration des systèmes d'information hospitaliers avec les données de localisation pour extraire les points de décision qui pourraient être une cible valable pour de futurs travaux de recherche. Cela aiderait à extraire une sémantique exécutable qui pourrait être utilisée pour la simulation des parcours des patients en tant que modèles de processus.

Ce travail de recherche encourage fortement les chercheurs à étudier ces questions:

- Le méta-modèle DIAG est une base évolutive et fiable pour l'application de la gestion des processus dans le secteur de la santé. Les chercheurs sont invités à participer à son développement qui est une recherche en cours au Centre de génie industriel de IMT Mines Albi.
- Proposition d'une approche globale d'évaluation des algorithmes de découverte de processus. Les méthodes de vérification de la conformité semblent être utilisées dans de nombreux travaux en considérant des critères de précision, de généralité, de simplicité et de fitness. Mais, comme on l'a vu dans ce travail, en utilisant ces critères, on n'est pas en mesure d'indiquer ce qui pourrait être une valeur de compromis appropriée entre ces critères.
- Un autre problème est l'extraction des points de décision et la découverte de gardes en intégrant des données de localisation et d'autres informations supplémentaires. Cela aiderait à la simulation de modèles qui sont considérés comme des actions «pronostiques» dans l'état de «gouvernance» de la méthodologie DIAG.
- Diagnostic automatique des processus métiers, ce problème a été complètement négligé dans la littérature, alors qu'on ne peut pas améliorer un processus sans connaître la cause des inefficacités.
- Enfin, le défi social doit être abordé dans les travaux futurs. Les progrès technologiques et scientifiques pour la surveillance des processus de soins de santé deviendraient inutiles si le personnel et les patients ne sont pas désireux de participer à des expériences similaires ou d'utiliser la technologie ILS.

List of Figures

1.1	Positioning the definition of patient pathways in accordance with previous research works such as the one in (W. Yang et al., 2014).	4
1.2	A representation of the normative, descriptive, and common pathways. .	6
1.3	Associating process mining activities with the indoor localization solution.	8
1.4	An overall view of this dissertation.	12
2.1	The structure by which two main keywords of this research work have been explored. Several decisions are taken based on the gaps found in the literature. The DIAG methodology and its functions are devised toward addressing these gaps.	14
2.2	The relationship among data science, data engineering, and data-driven decision making in business (Provost et al., 2013).	15
2.3	The Cross Industry Standard for Data Mining proposed in (Shearer, 2000); cited from (Provost et al., 2013).	16
2.4	The relationship among data science, business process science, and process mining defined in (W. M. P. v. d. Aalst, 2016)	20
2.5	The approach for analyzing the application of process mining activities in the healthcare sector.	24
2.6	This chart presents the focus of research works during last 5 years for the application process mining in healthcare.	25
2.7	The approach for analyzing research works related to applying localization techniques for a better organization of processes and activities in the industry of the society in general.	32
2.8	Analysis of the literature of process mining, indoor localization systems, and data mining relevant to the approaches in which the location data were used.	33
2.9	Analysis of the literature relevant to three criteria; discovery of patterns, conformance checking, and enhancement of performance.	34
2.10	Analyzing the literature of the application of process mining on location data during a 5 year-period. Accordingly, in 2016 the enhancement and performance evaluation did not receive enough attention.	36

2.11	The DIAG methodology.	44
2.12	The overall picture of the dissertation. A map for the readers.	47
3.1	Illustrating the need to provide a sense-making procedure for the recorded information in location event logs.	50
3.2	Representing the approach for detecting the existing concepts in the location event logs.	51
3.3	Structure of chapter 3.	52
3.4	The first function of data state, configuring the environments and systems	53
3.5	Process package of DIAG meta-model	53
3.6	Function and healthcare function packages within DIAG meta-model . .	54
3.7	Three of main packages in DIAG meta-model; Organization, Resources, and Objectives.	56
3.8	DIAG meta-model first version without development of the support for the business process diagnostic.	58
3.9	R.IO-DIAG framework.	59
3.10	A screen-shot of R.IO-DA which represents the modeling of the organization and its resources.	61
3.11	A screen-shot of R.IO-DA which represents the modeling of the functions. .	62
3.12	A screen-shot of R.IO-DA which represents the modeling of the objectives which is related to each case profile.	63
3.13	The second function of data state, location data gathering.	65
3.14	An illustration to present how movements of objects are tracked by the localization system.	65
3.15	Each entry to a zone generates a "update.zone" event.	66
3.16	Data interpreting function.	69
3.17	Association of the two sets of CEP interpretation rules.	70
3.18	A screen shot of the application loading the event logs.	74
3.19	A screen shot of the application extracting primary nodes without any added knowledge.	75
3.20	A process model representing the activities of an individual case by OPC modeling language.	79
3.21	OPC modeling language elements and their relations with the healthcare functions package and value-class concept in DIAG meta-model.	80
3.22	An example of patient pathways represented in R.IO-DIAG application by OPC modeling language.	81
3.23	An exemplary framework to present the relationship among data, information, and knowledge, (Benaben et al., 2019).	82
4.1	A schema to present the Information state and the integrated activities within.	87

4.2	A view on how a process discovery algorithm should address the noise problems cited in (W. M. P. v. d. Aalst, 2016).	90
4.3	Basic steps in heuristic mining algorithms (W. M. P. v. d. Aalst, 2016). . .	90
4.4	An illustration of two different systems with emergent properties.	93
4.5	An illustration of the hypothesis, which is to find a stable state between observed behaviors and modeled behaviors.	95
4.6	The sequence of actions for applying stable heuristic miner.	95
4.7	The sequence of applying calculations for the stable heuristic miner algorithm.	97
4.8	An illustrative example of the algorithm outcome. The descriptive reference process model representing stable behavior of the example event log (L). Red activities correspond to high variation in behaviors. Two activities are removed with lower significance level for the general behavior.	102
4.9	Order of tasks to extract the stable state by the stable heuristic miner V2	109
4.10	The result of applying stable heuristic miner $V2$. Statistical stability is applied for both edges and activities behaviors.	110
4.11	The process model extracted by the classic heuristic miner approach with a manual thresholds set at 20% . Accordingly, the model presents activities that have a dependency measure higher than 20%. Only 67% of the recorded behaviors respect this threshold.	115
4.12	The process model extracted by the classic heuristic miner approach with a manual thresholds set at 50% . Accordingly, the model presents activities that have a dependency measure higher than 50% . Only 61% of the recorded behaviors respect this threshold.	115
4.13	The process model extracted by the classic heuristic miner approach with a manual thresholds set at 80% . Accordingly, the model presents activities that have a dependency measure higher than 80%. Only 45% of the recorded behaviors respect this threshold.	115
4.14	The process model extracted by the classic heuristic miner approach with a manual thresholds set at 90% . Accordingly, the model presents activities that have a dependency measure higher than 90%. Only 25% of the recorded behaviors respect this threshold.	116
4.15	The process model extracted by the classic heuristic miner approach with manual configuration of thresholds. The value for threshold is set at 0 , therefore, the model shows all the registered behaviors.	116
4.16	The process model extracted by stable heuristic miner V1 , automatic detection of thresholds for activities. Red activities have shown high level of variations in their behavior according to the algorithm.	117
4.17	The model extracted by the stable heuristic miner V2 , automatic detection of thresholds for both activities and edges. Red edges and activities have shown high level of variations in their behavior according to the algorithm.	117
4.18	A general summary of what has been presented in this chapter.	118
5.1	An overall view on top of the third state of the DIAG methodology, Awareness.	123

5.2	This figure shows how this chapter is structured.	124
5.3	The defined steps to prepare the data.	125
5.4	This figure provides an illustration of the process capability ratio analysis.	131
5.5	The modeled data when $C_p = 1$	132
5.6	The modeled data when $C_p > 1$	132
5.7	The modeled data when $C_p < 1$	132
5.8	Analyzing the variations of pathways' length	132
5.9	The \bar{x} -chart to monitoring the stability of patients' pathways	133
5.10	Evaluation of the process capability by the length of patients' pathways. Note that the average(μ) is not well-centered.	134
5.11	Showing the solar system and the approximated and already measured distances among different elements.	138
5.12	What happens when the previous defined distance range changes? A hy- pothesis is made that a significant massive object is appeared between the two already discovered objects.	138
5.13	It shows the orientation of the descriptive process models (DM) around a normative or a reference model. Note that the DM's are extracted from existing information and it is possible for them to deviate from the reference behavior due to existence of the unknown causes.	140
5.14	It illustrates the hypothesis of MMP method. This method tries to inject some potential assignable causes (PAC) into the reference behavior to generate several processes (GM). If the minimum distance between a GM and a DM is found, it is possible to indicate they are deviated by the same cause.	140
5.15	This illustrates how both domain knowledge and data from information systems can be used by the MMP method to diagnose the distance between the modeled and observed behavior.	141
5.16	The newly added context package for representing the effect of PAC on the healthcare functions.	142
5.17	An illustration of the second diagnostic approach devised in this research work. Diagnosing the deviations by applying both the stable heuristic miner algorithms and the corresponding domain knowledge.	152
5.18	This figure shows the adjustment of the stable heuristic miner v2 algorithm by injecting potential assignable causes as the domain knowledge.	153
5.19	The discovered model which shows the causes of deviations. The activities and edges that cause higher variations –than the normal value– are shown in red . The deviating activities are shown in green . The dashed edges show the unexpected connections among activities. The cause of these deviations are shown on each behavior. If a deviation does not correspond to the domain knowledge, it gets a 0 value.	156
5.20	The second version of DIAG meta-model to support automatic diagnosing of patient pathways.	157
6.1	Structure of chapter six.	165

6.2	Modeling existing resources in the hospital.	167
6.3	Modeling provisioned human functions in the hospital.	167
6.5	An illustration of replaying the event logs to set off the interpretation rules. The chart presents the detection of in/out of zone events. The map corresponds to the movement of the tagged patient within the facility.	168
6.6	An illustration of the interpretation procedure.	169
6.7	An illustration of the process discovery procedure.	169
6.8	A comparison between the classic heuristic miner and stable heuristic miner algorithms results.	172
6.9	The discovered model of the hospital by applying the classic heuristic miner approach. In this model, the threshold value is set at 0percent This model is hardly analyzable.	173
6.10	The discovered model of the hospital by applying the classic heuristic miner approach. In this model, the threshold value is set at 20%.	174
6.11	The discovered model of the hospital by applying the classic heuristic miner approach. In this model, the threshold value is set at 50%.	175
6.12	The discovered model of the hospital by applying the classic heuristic miner approach. In this model, the threshold value is set at 80%.	176
6.13	The discovered model by the stable heuristic miner V1 . Note that, the statistical stability is not considered for edges. This has been addressed by the second version of the algorithm.	177
6.14	The discovered model by the stable heuristic miner V2. This model is related to the behavior of all of the 7 departments. Red activities and edges correspond to behaviors which represent where are the high variation behavior in the process.	179
6.15	The discovered model of the urology department by the classic heuristic miner.	180
6.16	The discovered model of the urology department by the stable heuristic miner V1. The Statistical stability method is applied only on activities behavior.	180
6.17	The discovered model of the urology department by the stable heuristic miner V2. The Statistical stability methods are applied for both activities and edges behavior.	181
6.18	This figure shows the duration of patient pathways for each individual case. The size of each observation changes based on the number of events that were registered for that particular process.	183
6.19	This figure shows the distance of patient pathways for each individual case. In addition, it provides a density analysis as well.	184
6.20	Visualizing patients pathways behaviors by considering the distance and duration of each case. y axis is related to the process duration and z axis is presenting the process distance.	185
6.21	The R -chart monitoring the in/outOf-control status of the process according to the range in which duration of processes change.	185
6.22	The \bar{x} -chart monitoring the in/outOf-control status of the process according to the average duration of each sample.	186

6.23	The R -chart monitoring the in/outOf-control status of the process according to the range in which distance of processes change.	186
6.24	The \bar{x} -chart monitoring the in/outOf-control status of the process according to the average distance of each sample.	186
6.25	Evaluating the performance of the process by analyzing duration of processes. This evaluation is based on the specification limits and the behavior of the modeled data according to these specification limits.	187
6.26	Evaluating the performance of the process by analyzing distance of processes.	187
6.27	The discovered model of the urology department which indicates the causes of deviations. The activities and edges with high variations are shown by a red color. The dashed edges show the unexpected and deviating behaviors among activities of the patient pathways. The causes of these deviations are indicated on each edge. The activities that are shown in green are unexpected activities that do not correspond to the common behavior of patients.	192
6.28	A summary of the gaps found in the literature (presented in chapter2) and the proposed solution in this research work.	195

List of Tables

1.1	Scientific questions and the proposed solutions.	10
1.2	This table clarifies some of the important terms that will be seen by the readers. These terms are related to the processes (patient pathways) that are either designed by the experts or discovered from event logs.	12
2.1	Cited research works that addressed the knowledge extraction from location data and those which addressed interpretation of these data.	34
2.2	Cited works addressing process mining activities.	35
2.3	A general overview of the scientific and technical challenges and the corresponding solutions applied by the DIAG methodology.	46
4.1	The approach of the chapter in covering the state of the art.	93
4.2	The calculation of \bar{x} , C_{4m} , and σ for each sample and the grand average of the population.	99
4.3	Presenting the behavior of the activities in the event log.	103
4.4	The table presenting the result of performing the sequence of calculation for the stable edges miner algorithm. In the "edge value" column, the red cells are related to the hot edges, the violet ones are representing "dirt roads", and the other cells represent the stable edges.	111
5.1	An example to illustrate the first extracted data table (dt1).	126
5.2	An example to show the sorted data table 1 (dt1) by ID of each case.	126
5.3	dt1 with the added duration of activities.	127
5.4	dt2 representing the data table for sampling the duration of processes.	127
5.5	An example of what dt3 could look like.	128
5.6	This table provides certain information about the key definitions within MMP method.	141
5.7	This table shows an example of how the relationship among PAC's and functions for patient pathways can be defined. It should be noted, these rules can be modified by the domain experts.	143

5.8	An example to illustrate how user and the system should interact. User based on the type of activity and the potential assignable cause can define which impact action should be triggered.	144
5.9	This table presents the generated models for the mentioned example. Note that the facts and their impacts are addressed each time by using one activity and one PAC.	146
5.10	Creating a matrix of processes activities and their orders.	148
5.11	It shows how to continue and evolve the comparison between two strings.	149
5.12	Final results of the example: the distance between two process models is 2. Two actions are required to convert one model to the other one. By replacing two activities	149
5.13	This table compares the distance of a descriptive model extracted from an event log and the generated models.	151
5.14	This table shows an illustration of how the domain knowledge will be recognized by the application.	155
6.1	A table corresponding to the added domain knowledge by the experts. This will be used as an input for the MMP method to diagnose the deviations of patient pathways.	170
6.2	A comparison between the number of observed behaviors in the event log (of all of the departments) and the modeled behaviors in the descriptive reference process model, thanks to stable heuristic miner <i>V1</i>	177
6.3	Comparing the number of observed behavior in the event log, with the result of stable heuristic miner <i>V1</i>	180
6.4	This table presents the total number of activities seen in the event log of the UROLOGY department. Also, it presents the result of applying stable heuristic miner algorithm V1 for evaluating the statistical stability of activities.	181
6.5	The used information for generating fake process models from the normative model (RM).	189
6.6	The generated models by injecting potential assignable causes into certain activities of the normative model (RM).	189
6.7	This table presents the final result of the MMP method where a process was detected with minimum distance from the descriptive model (DM_{240}). Since there were more than 3000 generated models, not all the processes are presented. This table shows the descriptive model has minimum distance from GM36. This could indicate the existence of same causes on DM_{240}	190
6.8	The knowledge set provided by the domain expert to be used by DIAG method.	191

Bibliography

- (Aa et al., 2019) H. van der Aa, C. Di Ciccio, H. Leopold, and H. A. Reijers. “[Extracting Declarative Process Models from Natural Language](#).” en. In: *Lecture Notes in Computer Science* (2019). Ed. by P. Giorgini and B. Weber, pp. 365–382 (cit. on p. 22).
- (W. M. P. v. d. Aalst, 2013a) W. M. P. van der Aalst. “[Decomposing Petri nets for process mining: A generic approach](#).” In: 31.4 (2013), pp. 471–507 (cit. on p. 21).
- (W. M. P. v. d. Aalst, 2013b) W. M. P. van der Aalst. “[Mediating between modeled and observed behavior: The quest for the “right” process: Keynote](#).” In: *IEEE 7th International Conference on Research Challenges in Information Science (RCIS)*. May 2013, pp. 1–12 (cit. on p. 114).
- (W. M. P. v. d. Aalst, 2016) W. M. P. van der Aalst. [Process Mining: Data Science in Action](#). en. Google-Books-ID: hUEGDAAAQBAJ. Springer, Apr. 2016 (cit. on pp. 19, 20, 89–91, 96).
- (W. M. P. v. d. Aalst et al., 2003) W. M. P. van der Aalst, B. F. van Dongen, J. Herbst, L. Maruster, G. Schimm, and A. J. M. M. Weijters. “[Workflow mining: A survey of issues and approaches](#).” In: *Data & Knowledge Engineering* 47.2 (Nov. 2003), pp. 237–267 (cit. on p. 20).
- (W. v. d. Aalst et al., 2004) W. v. d. Aalst, T. Weijters, and L. Maruster. “[Workflow mining: discovering process models from event logs](#).” In: *IEEE Transactions on Knowledge and Data Engineering* 16.9 (Sept. 2004), pp. 1128–1142 (cit. on p. 19).
- (W. v. d. Aalst et al., 2012) W. van der Aalst et al. “[Process Mining Manifesto](#).” In: *Business Process Management Workshops*. Ed. by F. Daniel, K. Barkaoui, and S. Dustdar. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 169–194 (cit. on p. 45).
- (W. M. P. v. d. Aalst, 2018) W. M. P. van der Aalst. “[Process Mining and Simulation: A Match Made in Heaven!](#)” In: *Proceedings of the 50th Computer Simulation Conference*. SummerSim ’18. event-place: Bordeaux, France. San Diego, CA, USA: Society for Computer Simulation International, 2018, 4:1–4:12 (cit. on pp. 10, 28).
- (Adame et al., 2018) T. Adame, A. Bel, A. Carreras, J. Melia-Segui, M. Oliver, and R. Pous. “CUIDATS: An RFID–WSN hybrid monitoring system for smart health care environments.” In: 78 (2018), pp. 602–615 (cit. on p. 30).
- (Agrawal et al., 1998) R. Agrawal, D. Gunopulos, and F. Leymann. “[Mining process models from workflow logs](#).” en. In: *Advances in Database Technology — EDBT’98*. Ed. by H.-J. Schek, G. Alonso, F. Saltor, and I. Ramos. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 1998, pp. 467–483 (cit. on pp. 19, 20).

-
- (Aguilar-Savén, 2004) R. S. Aguilar-Savén. [“Business process modelling: Review and framework.”](#) In: *International Journal of Production Economics*. Production Planning and Control 90.2 (July 2004), pp. 129–149 (cit. on p. 20).
- (Al Nuaimi et al., 2011) K. Al Nuaimi and H. Kamel. [“A survey of indoor positioning systems and algorithms.”](#) In: *2011 international conference on innovations in information technology*. IEEE. 2011, pp. 185–190 (cit. on pp. 30, 31).
- (Alameda et al., 2009) C. Alameda and C. Suárez. [“Clinical outcomes in medical outliers admitted to hospital with heart failure.”](#) In: *European Journal of Internal Medicine* 20.8 (Dec. 2009), pp. 764–767 (cit. on p. 4).
- (Amundson et al., 2009) I. Amundson and X. D. Koutsoukos. [“A Survey on Localization for Mobile Wireless Sensor Networks.”](#) In: *Mobile Entity Localization and Tracking in GPS-less Environments*. Ed. by R. Fuller and X. D. Koutsoukos. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 235–254 (cit. on pp. 30, 31).
- (Araghi et al., 2018a) S. N. Araghi, F. Fontanili, E. Lamine, N. Salatge, J. Lesbegueries, S. R. Pouyade, L. Tancerel, and F. Benaben. [“A Conceptual Framework to Support Discovering of Patients’ Pathways as Operational Process Charts.”](#) In: *2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA)*. Oct. 2018, pp. 1–6 (cit. on pp. 34, 35, 51, 54, 67, 79).
- (Araghi et al., 2019) S. N. Araghi, F. Fontanili, E. Lamine, N. Salatge, J. Lesbegueries, S. R. Pouyade, and F. Benaben. [“Evaluating the Process Capability Ratio of Patients’ Pathways by the Application of Process Mining, SPC and RTLS.”](#) In: Sept. 2019, pp. 302–309 (cit. on pp. 34, 35, 130).
- (Araghi et al., 2018b) S. N. Araghi, F. Fontanili, E. Lamine, L. Tancerel, and F. Benaben. [“Monitoring and analyzing patients’ pathways by the application of Process Mining, SPC, and I-RTLS.”](#) In: *IFAC-PapersOnLine*. 16th IFAC Symposium on Information Control Problems in Manufacturing INCOM 2018 51.11 (Jan. 2018), pp. 980–985 (cit. on pp. 34, 35, 130, 133, 194).
- (Aryal et al., 2017) A. M. Aryal and Sujing Wang. [“Discovery of patterns in spatio-temporal data using clustering techniques.”](#) In: *2017 2nd International Conference on Image, Vision and Computing (ICIVC)*. June 2017, pp. 990–995 (cit. on pp. 34, 35, 38).
- (Ashdown et al., 2003) D. Ashdown, D. Williams, K. Davenport, and R. Kirby. “The impact of medical outliers on elective surgical lists.” In: *Bulletin of the Royal College of Surgeons of England* 85.2 (2003), pp. 46–47 (cit. on p. 4).
- (Augusto et al., 2019a) A. Augusto, R. Conforti, M. Dumas, M. L. Rosa, F. M. Maggi, A. Marrella, M. Mecella, and A. Soo. [“Automated Discovery of Process Models from Event Logs: Review and Benchmark.”](#) In: *IEEE Transactions on Knowledge and Data Engineering* 31.4 (Apr. 2019), pp. 686–705 (cit. on pp. 20, 114, 197).
- (Augusto et al., 2019b) A. Augusto, R. Conforti, M. Dumas, M. La Rosa, and A. Polyvyanyy. [“Split miner: automated discovery of accurate and simple business process models from event logs.”](#) In: *Knowledge and Information Systems* 59.2 (May 2019), pp. 251–284 (cit. on p. 22).
- (Bao et al., 2017) X. Bao and L. Wang. [“Discovering Interesting Co-location Patterns Interactively Using Ontologies.”](#) In: *Database Systems for Advanced Applications*. Ed. by Z. Bao, G. Trajcevski, L. Chang, and W. Hua. Cham: Springer International Publishing, 2017, pp. 75–89 (cit. on pp. 34, 35, 38).
- (Bárány et al., 2007) I. Bárány and V. Vu. [“Central limit theorems for Gaussian polytopes.”](#) EN. In: *The Annals of Probability* 35.4 (July 2007), pp. 1593–1621 (cit. on p. 96).

- (Batista et al., 2019) E. Batista, A. Martínez-Ballesté, M. Peña, X. Singla, and A. Solanas. "HGIS: A Healthcare-Oriented Approach to Geographic Information Systems." In: *Applications in Electronics Pervading Industry, Environment and Society*. Ed. by S. Saponara and A. De Gloria. Cham: Springer International Publishing, 2019, pp. 59–65 (cit. on p. 27).
- (Benaben et al., 2008) F. Benaben, C. Hanachi, M. Lauras, P. Couget, and V. Chapurlat. "A metamodel and its ontology to guide crisis characterization and its collaborative management." In: *Proceedings of the 5th International Conference on Information Systems for Crisis Response and Management (ISCRAM), Washington, DC, USA, May. 2008*, pp. 4–7 (cit. on p. 59).
- (Benaben et al., 2019) F. Benaben, J. Li, I. Koura, B. Montreuil, M. Lauras, W. Mu, and J. Gou. "A Tentative Framework for Risk and Opportunity Detection in A Collaborative Environment Based on Data Interpretation." eng. In: Jan. 2019 (cit. on p. 82).
- (Bénaben et al., 2016) F. Bénaben, M. Lauras, S. Truptil, and N. Salatgé. "A Metamodel for Knowledge Management in Crisis Management." In: *2016 49th Hawaii International Conference on System Sciences (HICSS)*. Jan. 2016, pp. 126–135 (cit. on p. 59).
- (Bernardi et al., 2014) M. L. Bernardi, M. Cimitile, and F. M. Maggi. "Discovering cross-organizational business rules from the cloud." In: *2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*. Dec. 2014, pp. 389–396 (cit. on p. 22).
- (Bezerra et al., 2013) F. Bezerra and J. Wainer. "Algorithms for Anomaly Detection of Traces in Logs of Process Aware Information Systems." In: 38.1 (Mar. 2013), pp. 33–44 (cit. on p. 88).
- (Blank et al., 2016) P. Blank, M. Maurer, M. Siebenhofer, A. Rogge-Solti, and S. Schonig. "Location-Aware Path Alignment in Process Mining." In: *2016 IEEE 20th International Enterprise Distributed Object Computing Workshop (EDOCW)*. Sept. 2016, pp. 1–8 (cit. on pp. 34, 35, 38).
- (Boersma et al., 2019) H. J. Boersma, T. I. Leung, R. Vanwersch, E. Heeren, and G. G. van Merode. "Optimizing Care Processes with Operational Excellence & Process Mining." In: *Fundamentals of Clinical Data Science*. Ed. by P. Kubben, M. Dumontier, and A. Dekker. Cham: Springer International Publishing, 2019, pp. 181–192 (cit. on p. 29).
- (Booker et al., 2001) J. D. Booker, M. Raines, and K. G. Swift. *Designing capable and reliable products*. Butterworth-Heinemann, 2001 (cit. on p. 131).
- (Boyles, 1991) R. A. Boyles. "The Taguchi Capability Index." In: *Journal of Quality Technology* 23.1 (Jan. 1991), pp. 17–26 (cit. on p. 130).
- (Bozkaya et al., 2009) M. Bozkaya, J. Gabriels, and J. M. v. d. Werf. "Process Diagnostics: A Method Based on Process Mining." In: *2009 International Conference on Information, Process, and Knowledge Management*. Feb. 2009, pp. 22–27 (cit. on p. 43).
- (Breuker et al., 2016) D. Breuker, M. Matzner, P. Delfmann, and J. Becker. "Comprehensive Predictive Models for Business Processes." In: *MIS Q.* 40.4 (Dec. 2016), pp. 1009–1034 (cit. on p. 21).
- (Buijs et al., 2014) J. C. a. M. Buijs, B. F. van Dongen, and W. M. P. van der Aalst. "Quality Dimensions in Process Discovery: The Importance of Fitness, Precision, Generalization and Simplicity." In: *International Journal of Cooperative Information Systems* 23.01 (Mar. 2014), p. 1440001 (cit. on p. 21).
- (Campbell et al., 1998) H. Campbell, R. Hotchkiss, N. Bradshaw, and M. Porteous. "Integrated care pathways." en. In: *BMJ* 316.7125 (Jan. 1998), pp. 133–137 (cit. on p. 5).

-
- (Cho et al., 2019) M. Cho, M. Song, S. Yoo, and H. A. Reijers. [“An Evidence-Based Decision Support Framework for Clinician Medical Scheduling.”](#) In: *IEEE Access* 7 (2019), pp. 15239–15249 (cit. on p. 27).
- (Cho et al., 2014) M. Cho, M. Song, and S. Yoo. [“A Systematic Methodology for Outpatient Process Analysis Based on Process Mining.”](#) In: *Asia Pacific Business Process Management*. Ed. by C. Ouyang and J.-Y. Jung. Cham: Springer International Publishing, 2014, pp. 31–42 (cit. on p. 45).
- (Commission et al., 2000) A. Commission et al. “Inpatient admissions and bed management in NHS acute hospitals.” In: *The Stationery Office. London* (2000) (cit. on p. 4).
- (Conforti et al., 2014) R. Conforti, M. Dumas, L. García-Bañuelos, and M. La Rosa. [“Beyond Tasks and Gateways: Discovering BPMN Models with Subprocesses, Boundary Events and Activity Markers.”](#) In: *Business Process Management*. Ed. by S. Sadiq, P. Soffer, and H. Völzer. Cham: Springer International Publishing, 2014, pp. 101–117 (cit. on p. 22).
- (Conforti et al., 2016) R. Conforti, M. Dumas, L. García-Bañuelos, and M. La Rosa. [“BPMN Miner: Automated discovery of BPMN process models with hierarchical structure.”](#) In: *Information Systems* 56 (Mar. 2016), pp. 284–303 (cit. on p. 22).
- (Cook et al., 1998) J. E. Cook and A. L. Wolf. [“Discovering Models of Software Processes from Event-based Data.”](#) In: 7.3 (July 1998), pp. 215–249 (cit. on p. 19).
- (Cotera et al., 2016) P. Cotera, M. Velazquez, D. Cruz, L. Medina, and M. Bandala. [“Indoor Robot Positioning Using an Enhanced Trilateration Algorithm.”](#) en. In: *International Journal of Advanced Robotic Systems* 13.3 (May 2016), p. 110 (cit. on p. 31).
- (Creamer et al., 2010) G. L. Creamer, A. Dahl, D. Perumal, G. Tan, and J. B. Koea. [“Anatomy of the ward round: the time spent in different activities.”](#) eng. In: *ANZ journal of surgery* 80.12 (Dec. 2010), pp. 930–932 (cit. on p. 4).
- (Curran et al., 2011) K. Curran, E. Furey, T. Lunney, J. Santos, D. Woods, and A. McCaughey. [“An evaluation of indoor location determination technologies.”](#) In: *Journal of Location Based Services* 5.2 (June 2011), pp. 61–78 (cit. on p. 31).
- (Dahlin et al., 2019) S. Dahlin, H. Eriksson, and H. Raharjo. [“Process Mining for Quality Improvement: Propositions for Practice and Research.”](#) eng. In: *Quality Management in Health Care* 28.1 (Mar. 2019), pp. 8–14 (cit. on p. 29).
- (De Cnudde et al., 2014) S. De Cnudde, J. Claes, and G. Poels. [“Improving the Quality of the Heuristics Miner in ProM 6.2.”](#) In: *Expert Syst. Appl.* 41.17 (Dec. 2014), pp. 7678–7690 (cit. on pp. 91, 92, 114, 118, 171, 194).
- (Dixit et al., 2018) P. M. Dixit, J. C. A. M. Buijs, and W. M. P. v. d. Aalst. [“ProDiGy : Human-in-the-loop process discovery.”](#) In: *2018 12th International Conference on Research Challenges in Information Science (RCIS)*. May 2018, pp. 1–12 (cit. on p. 25).
- (Dogan et al., 2019a) O. Dogan, J.-L. Bayo-Monton, C. Fernandez-Llatas, and B. Oztaysi. [“Analyzing of Gender Behaviors from Paths Using Process Mining: A Shopping Mall Application.”](#) eng. In: *Sensors (Basel, Switzerland)* 19.3 (Jan. 2019) (cit. on pp. 34, 35, 39).
- (Dogan et al., 2019b) O. Dogan, A. Martinez-Millana, E. Rojas, M. Sepúlveda, J. Munoz-Gama, V. Traver, and C. Fernandez-Llatas. [“Individual Behavior Modeling with Sensors Using Process Mining.”](#) en. In: *Electronics* 8.7 (July 2019), p. 766 (cit. on p. 7).
- (Duma, 2019) D. Duma. [“Online optimization methods applied to the management of health services.”](#) In: *4OR* (July 2019) (cit. on pp. 28, 30).

- (Duma et al., 2018) D. Duma and R. Aringhieri. [“An ad hoc process mining approach to discover patient paths of an Emergency Department.”](#) In: *Flexible Services and Manufacturing Journal* (Dec. 2018) (cit. on p. 26).
- (Dumas et al., 2018) M. Dumas, M. L. Rosa, J. Mendling, and H. Reijers. [Fundamentals of Business Process Management](#). en. 2nd ed. Berlin Heidelberg: Springer-Verlag, 2018 (cit. on pp. 5, 6, 10, 54).
- (M. L. v. Eck et al., 2017) M. L. v. Eck, N. Sidorova, and W. M. P. v. d. Aalst. [“Guided Interaction Exploration in Artifact-centric Process Models.”](#) In: *2017 IEEE 19th Conference on Business Informatics (CBI)*. Vol. 01. July 2017, pp. 109–118 (cit. on p. 21).
- (M. L. v. Eck et al., 2015) M. L. van Eck, X. Lu, S. J. J. Leemans, and W. M. P. van der Aalst. [“PM²: A Process Mining Project Methodology.”](#) en. In: *Advanced Information Systems Engineering*. Ed. by J. Zdravkovic, M. Kirikova, and P. Johannesson. Lecture Notes in Computer Science. Springer International Publishing, 2015, pp. 297–313 (cit. on p. 45).
- (M. L. v. Eck et al., 2016) M. L. van Eck, N. Sidorova, and W. M. P. van der Aalst. [“Discovering and Exploring State-Based Models for Multi-perspective Processes.”](#) In: *Business Process Management*. Ed. by M. La Rosa, P. Loos, and O. Pastor. Cham: Springer International Publishing, 2016, pp. 142–157 (cit. on p. 21).
- (Eisenthal et al., 1983) S. Eisenthal, C. Koopman, and A. Lazare. [“Process Analysis of Two Dimensions of the Negotiated Approach in Relation to Satisfaction in the Initial Interview.”](#) ENGLISH. In: *The Journal of Nervous and Mental Disease* 171.1 (Jan. 1983), pp. 49–54 (cit. on p. 193).
- (Ertek et al., 2017) G. Ertek, X. Chi, and A. N. Zhang. [“A Framework for Mining RFID Data From Schedule-Based Systems.”](#) In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 47.11 (Nov. 2017), pp. 2967–2984 (cit. on pp. 34, 35, 37).
- (Feng Ling et al., 2016) Feng Ling, Tianyue Sun, Xinning Zhu, Qingqing Chen, Xiaosheng Tang, and Xin Ke. [“Mining travel behaviors of tourists with mobile phone data: A case study in Hainan.”](#) In: *2016 2nd IEEE International Conference on Computer and Communications (ICCC)*. Oct. 2016, pp. 1524–1529 (cit. on pp. 34, 35, 38).
- (Ferilli, 2014) S. Ferilli. [“WoMan: Logic-Based Workflow Learning and Management.”](#) In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 44.6 (June 2014), pp. 744–756 (cit. on p. 21).
- (Fernandez-Llatas et al., 2015) C. Fernandez-Llatas, A. Lizondo, E. Monton, J.-M. Benedi, and V. Traver. [“Process Mining Methodology for Health Process Tracking Using Real-Time Indoor Location Systems.”](#) en. In: *Sensors* 15.12 (Nov. 2015), pp. 29821–29840 (cit. on pp. 7, 34, 35, 38, 45, 89).
- (Garaeva et al., 2017) A. Garaeva, F. Makhmutova, I. Anikin, and K. Sattler. [“A framework for co-location patterns mining in big spatial data.”](#) In: *2017 XX IEEE International Conference on Soft Computing and Measurements (SCM)*. May 2017, pp. 477–480 (cit. on pp. 34, 35, 38).
- (Garg et al., 2009) L. Garg, S. McClean, B. Meenan, and P. Millard. [“Non-homogeneous Markov models for sequential pattern mining of healthcare data.”](#) en. In: *IMA Journal of Management Mathematics* 20.4 (Oct. 2009), pp. 327–344 (cit. on p. 28).
- (Gatta et al., 2018) R. Gatta, M. Vallati, J. Lenkiewicz, C. Casà, F. Cellini, A. Damiani, and V. Valentini. [“A Framework for Event Log Generation and Knowledge Representation for Process Mining in Healthcare.”](#) In: *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)*. Nov. 2018, pp. 647–654 (cit. on p. 30).

-
- (Ghionna et al., 2008) L. Ghionna, G. Greco, A. Guzzo, and L. Pontieri. [“Outlier Detection Techniques for Process Mining Applications.”](#) In: *Proceedings of the 17th International Conference on Foundations of Intelligent Systems*. ISMIS’08. Springer-Verlag, 2008, pp. 150–159 (cit. on p. 88).
- (Gilligan et al., 2008) S. Gilligan and M. Walters. [“Quality improvements in hospital flow may lead to a reduction in mortality.”](#) en. In: *Clinical Governance: An International Journal* (Jan. 2008) (cit. on p. 4).
- (Gorban, 2014) I. I. Gorban. [“Phenomenon of statistical stability.”](#) In: *Technical Physics* 59.3 (Mar. 2014), pp. 333–340 (cit. on p. 94).
- (Gorban, 2017) I. I. Gorban. [The Statistical Stability Phenomenon](#). en. Mathematical Engineering. Springer International Publishing, 2017 (cit. on pp. 92, 93).
- (Goulding et al., 2012) L. Goulding, J. Adamson, I. Watt, and J. Wright. [“Patient safety in patients who occupy beds on clinically inappropriate wards: a qualitative interview study with NHS staff.”](#) en. In: *BMJ Quality & Safety* 21.3 (Mar. 2012), pp. 218–224 (cit. on p. 4).
- (Greco et al., 2015) G. Greco, A. Guzzo, F. Lupia, and L. Pontieri. [“Process Discovery Under Precedence Constraints.”](#) In: *ACM Trans. Knowl. Discov. Data* 9.4 (June 2015), 32:1–32:39 (cit. on p. 21).
- (Greco et al., 2012) G. Greco, A. Guzzo, and L. Pontieri. [“Process Discovery via Precedence Constraints.”](#) In: *Proceedings of the 20th European Conference on Artificial Intelligence*. ECAI’12. event-place: Montpellier, France. Amsterdam, The Netherlands, The Netherlands: IOS Press, 2012, pp. 366–371 (cit. on p. 21).
- (Gunther et al., 2007) C. W. Gunther and W. M. P. van der Aalst. [“Fuzzy Mining – Adaptive Process Simplification Based on Multi-perspective Metrics.”](#) In: *Business Process Management*. Ed. by G. Alonso, P. Dadam, and M. Rosemann. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 328–343 (cit. on p. 172).
- (Günther et al., 2007) C. W. Günther and W. M. P. van der Aalst. [“Fuzzy Mining – Adaptive Process Simplification Based on Multi-perspective Metrics.”](#) en. In: *Business Process Management*. Ed. by G. Alonso, P. Dadam, and M. Rosemann. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2007, pp. 328–343 (cit. on p. 23).
- (Gurgen Erdogan et al., 2018) T. Gurgen Erdogan and A. Tarhan. [“A Goal-Driven Evaluation Method Based On Process Mining for Healthcare Processes.”](#) en. In: *Applied Sciences* 8.6 (June 2018), p. 894 (cit. on p. 25).
- (Hall, 2013) R. Hall, ed. [Patient Flow: Reducing Delay in Healthcare Delivery](#). en. 2nd ed. International Series in Operations Research & Management Science. Springer US, 2013 (cit. on p. 5).
- (Harel et al., 2016) Z. Harel, S. A. Silver, R. F. McQuillan, A. V. Weizman, A. Thomas, G. M. Chertow, G. Nesrallah, C. T. Chan, and C. M. Bell. [“How to Diagnose Solutions to a Quality of Care Problem.”](#) In: *Clinical Journal of the American Society of Nephrology : CJASN* 11.5 (May 2016), pp. 901–907 (cit. on p. 143).
- (He et al., 2019) Z. He, Y. Du, L. Qi, and H. Du. [“A Model Repair Approach Based on Petri Nets by Constructing Free-loop Structures.”](#) In: (2019) (cit. on p. 21).
- (Herland et al., 2014) M. Herland, T. M. Khoshgoftaar, and R. Wald. [“A review of data mining using big data in health informatics.”](#) In: *Journal Of Big Data* 1.1 (June 2014), p. 2 (cit. on pp. 16, 17).
- (Hess et al., 2012) L. M. Hess, F. B. Stehman, M. W. Method, T. D. Weathers, P. Gupta, and J. M. Schilder. [“Identification of the optimal pathway to reach an accurate diagnosis in the absence of an early detection strategy for ovarian cancer.”](#) eng. In: *Gynecologic Oncology* 127.3 (Dec. 2012), pp. 564–568 (cit. on p. 17).

- (Huang et al., 2011) Z. Huang and A. Kumar. "A Study of Quality and Accuracy Trade-offs in Process Mining." In: *INFORMS Journal on Computing* 24.2 (Mar. 2011), pp. 311–327 (cit. on p. 21).
- (Hwang et al., 2017) I. Hwang and Y. J. Jang. "Process Mining to Discover Shoppers' Pathways at a Fashion Retail Store Using a WiFi-Base Indoor Positioning System." In: *IEEE Transactions on Automation Science and Engineering* 14.4 (Oct. 2017), pp. 1786–1792 (cit. on pp. 34, 35, 37).
- (Ibanez-Sanchez et al., 2019) G. Ibanez-Sanchez, C. Fernandez-Llatas, A. Martinez-Millana, A. Celda, J. Mandingorra, L. Aparici-Tortajada, Z. Valero-Ramon, J. Munoz-Gama, M. Sepúlveda, E. Rojas, V. Gálvez, D. Capurro, and V. Traver. "Toward Value-Based Healthcare through Interactive Process Mining in Emergency Rooms: The Stroke Case." In: *International Journal of Environmental Research and Public Health* 16.10 (May 2019) (cit. on p. 27).
- (Ijcsis et al., 2019) J. o. C. S. Ijcsis and G. Battineni. "Process mining case study approach: Extraction of unconventional event logs to improve performance in Hospital Information Systems (HIS) Corresponding Author." en. In: *IJCSIS Vol 17 No 4 April Issue* () (cit. on p. 27).
- (Institute of Medicine (US) Committee on Assuring the Health of the Public in the 21st Century, 2002) Institute of Medicine (US) Committee on Assuring the Health of the Public in the 21st Century. *The Future of the Public's Health in the 21st Century*. eng. Washington (DC): National Academies Press (US), 2002 (cit. on p. 18).
- (Ishak et al., 2005) I. S. Ishak and R. A. Alias. *Designing a strategic information system planning methodology For Malaysian institutes of higher learning (ISP-IPTA)*. Universiti Teknologi Malaysia, 2005 (cit. on p. 43).
- (Jalali et al., 2010) H. Jalali and A. Baraani. "Genetic-based anomaly detection in logs of process aware systems." In: 64 (2010), pp. 304–309 (cit. on p. 88).
- (Jin et al., 2015) P. Jin, J. Du, C. Huang, S. Wan, and L. Yue. "Detecting hotspots from trajectory data in indoor spaces." In: *International Conference on Database Systems for Advanced Applications*. Springer. 2015, pp. 209–225 (cit. on pp. 34, 35).
- (Johnson et al., 2019) O. A. Johnson, T. Ba Dhafari, A. Kurniati, F. Fox, and E. Rojas. "The ClearPath Method for Care Pathway Process Mining and Simulation." In: *Business Process Management Workshops*. Ed. by F. Daniel, Q. Z. Sheng, and H. Motahari. Cham: Springer International Publishing, 2019, pp. 239–250 (cit. on pp. 26, 45).
- (Jothi et al., 2015) N. Jothi, N. A. Rashid, and W. Husain. "Data Mining in Healthcare – A Review." In: *Procedia Computer Science*. The Third Information Systems International Conference 2015 72 (Jan. 2015), pp. 306–313 (cit. on p. 16).
- (Kalenkova et al., 2017) A. A. Kalenkova, W. M. P. van der Aalst, I. A. Lomazova, and V. A. Rubin. "Process mining using BPMN: relating event logs and process models." In: *Software & Systems Modeling* 16.4 (Oct. 2017), pp. 1019–1048 (cit. on p. 22).
- (Kamel Boulos et al., 2012) M. N. Kamel Boulos and G. Berry. "Real-time locating systems (RTLS) in healthcare: a condensed primer." eng. In: *International Journal of Health Geographics* 11 (June 2012), p. 25 (cit. on p. 89).
- (Kim et al., 2019) H. S. Kim, H. K. Kim, K. O. Kang, and Y. S. Kim. "Determinants of health-related quality of life among outpatients with acute coronary artery disease after percutaneous coronary intervention." en. In: *Japan Journal of Nursing Science* 16.1 (2019), pp. 3–16 (cit. on p. 29).

-
- (Kouylekov, 2006) M. O. Kouylekov. [“Recognizing Textual Entailment with Tree Edit Distance: Application to Question Answering and Information Extraction.”](#) eng. In: 2006 (cit. on p. 147).
- (Kuei, 2004) C.-H. Kuei. [Statistical Methods for Six Sigma in R&D and Manufacturing.](#) en. June 2004 (cit. on p. 130).
- (Lamr et al., 2016) M. Lamr and J. Skrbek. [Traffic Data and Possibilities of Their Utilization for Safer Traffic.](#) English. Ed. by J. Skrbek, D. Nejedlova, and T. Semerádova. WOS:000404420200007. Technical Univ Liberec, Faculty Economics, 2016 (cit. on pp. 34, 35, 38).
- (M. Leemans et al., 2017) M. Leemans and W. M. P. van der Aalst. [“Modeling and Discovering Cancellation Behavior.”](#) en. In: *On the Move to Meaningful Internet Systems. OTM 2017 Conferences.* Ed. by H. Panetto, C. Debruyne, W. Gaaloul, M. Papazoglou, A. Paschke, C. A. Ardagna, and R. Meersman. Lecture Notes in Computer Science. Springer International Publishing, 2017, pp. 93–113 (cit. on p. 21).
- (S. J. J. Leemans et al., 2014a) S. J. J. Leemans, D. Fahland, and W. M. P. van der Aalst. [“Process and deviation exploration with inductive visual miner.”](#) en. In: 1295 (2014). Ed. by R. Schmidt, W. Guédria, I. Bider, and S. Guerreiro, pp. 46–50 (cit. on p. 21).
- (S. J. J. Leemans et al., 2013) S. J. J. Leemans, D. Fahland, and W. M. P. van der Aalst. [“Discovering Block-Structured Process Models from Event Logs - A Constructive Approach.”](#) In: *Application and Theory of Petri Nets and Concurrency.* Ed. by J.-M. Colom and J. Desel. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 311–329 (cit. on p. 21).
- (S. J. J. Leemans et al., 2014b) S. J. J. Leemans, D. Fahland, and W. M. P. van der Aalst. [“Discovering Block-Structured Process Models from Incomplete Event Logs.”](#) In: *Application and Theory of Petri Nets and Concurrency.* Ed. by G. Ciardo and E. Kindler. Cham: Springer International Publishing, 2014, pp. 91–110 (cit. on p. 21).
- (S. J. J. Leemans et al., 2015) S. J. J. Leemans, D. Fahland, and W. M. P. van der Aalst. [“Exploring Processes and Deviations.”](#) In: *Business Process Management Workshops.* Ed. by F. Fournier and J. Mendling. Cham: Springer International Publishing, 2015, pp. 304–316 (cit. on p. 21).
- (Lepage et al., 2009) B. Lepage, R. Robert, M. Lebeau, C. Aubeneau, C. Silvain, and V. Migeot. [“Use of a risk analysis method to improve care management for outlying inpatients in a university hospital.”](#) eng. In: *Quality & Safety in Health Care* 18.6 (Dec. 2009), pp. 441–445 (cit. on p. 4).
- (Li, G. et al., 2017) Li, G., van der Aalst, W.M.P., and Information Systems WSK&I. [“A framework for detecting deviations in complex event logs.”](#) en. In: *Intelligent Data Analysis* 21.4 (Aug. 2017), pp. 759–779 (cit. on p. 88).
- (Liao et al., 2015) J. Liao, Z. Wang, L. Wan, Q. C. Cao, and H. Qi. [“Smart Diary: A Smartphone-Based Framework for Sensing, Inferring, and Logging Users’ Daily Life.”](#) In: *IEEE Sensors Journal* 15.5 (May 2015), pp. 2761–2773 (cit. on pp. 34, 35).
- (Lira et al., 2019) R. Lira, J. Salas-Morales, R. de la Fuente, R. Fuentes, M. Sepúlveda, M. Arias, V. Herskovic, and J. Munoz-Gama. [“Tailored Process Feedback Through Process Mining for Surgical Procedures in Medical Training: The Central Venous Catheter Case.”](#) In: *Business Process Management Workshops.* Ed. by F. Daniel, Q. Z. Sheng, and H. Motahari. Cham: Springer International Publishing, 2019, pp. 163–174 (cit. on p. 26).

- (Litchfield et al., 2018) I. Litchfield, C. Hoyer, D. Shukla, R. Backman, A. Turner, M. Lee, and P. Weber. “Can process mining automatically describe care pathways of patients with long-term conditions in UK primary care? A study protocol.” eng. In: *BMJ open* 8.12 (2018), e019947 (cit. on p. 193).
- (Maggi et al., 2011) F. M. Maggi, A. J. Mooij, and W. M. P. v. d. Aalst. “User-guided discovery of declarative process models.” In: *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*. Apr. 2011, pp. 192–199 (cit. on p. 22).
- (Maggi et al., 2013) F. M. Maggi, R. P. J. C. Bose, and W. M. P. van der Aalst. “A Knowledge-Based Integrated Approach for Discovering and Repairing Declare Maps.” In: *Advanced Information Systems Engineering*. Ed. by C. Salinesi, M. C. Norrie, and Ó. Pastor. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 433–448 (cit. on p. 22).
- (Mans et al., 2013) R. S. Mans, W. M. P. van der Aalst, R. J. B. Vanwersch, and A. J. Moleman. “Process Mining in Healthcare: Data Challenges When Answering Frequently Posed Questions.” In: *Process Support and Knowledge Representation in Health Care*. Ed. by R. Lenz, S. Miksch, M. Peleg, M. Reichert, D. Riaño, and A. ten Teije. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 140–153 (cit. on p. 22).
- (Mans et al., 2015) R. S. Mans, W. v. d. Aalst, and R. J. B. Vanwersch. *Process Mining in Healthcare: Evaluating and Exploiting Operational Healthcare Processes*. en. SpringerBriefs in Business Process Management. Springer International Publishing, 2015 (cit. on p. 17).
- (Martin, 2019) N. Martin. “Using Indoor Location System Data to Enhance the Quality of Healthcare Event Logs: Opportunities and Challenges.” en. In: *Business Process Management Workshops*. Ed. by F. Daniel, Q. Z. Sheng, and H. Motahari. Lecture Notes in Business Information Processing. Springer International Publishing, 2019, pp. 226–238 (cit. on p. 39).
- (Martinez-Millana et al., 2019) A. Martinez-Millana, A. Lizondo, R. Gatta, S. Vera, V. T. Salcedo, and C. Fernandez-Llatas. “Process Mining Dashboard in Operating Rooms: Analysis of Staff Expectations with Analytic Hierarchy Process.” In: *International Journal of Environmental Research and Public Health* 16.2 (Jan. 2019) (cit. on pp. 34, 35, 38).
- (Mărușter et al., 2009) L. Mărușter and N. R. T. P. van Beest. “Redesigning business processes: a methodology based on simulation and process mining techniques.” In: *Knowledge and Information Systems* 21.3 (June 2009), p. 267 (cit. on p. 43).
- (Mazimpaka et al., 2016) J. D. Mazimpaka and S. Timpf. “Trajectory data mining: A review of methods and applications.” In: *Journal of Spatial Information Science* 2016.13 (Dec. 2016), pp. 61–99 (cit. on pp. 34, 35, 37).
- (Mendling et al., 2019) J. Mendling, H. Leopold, L. H. Thom, R. G. do Sul, and H. van der Aa. “Natural Language Processing with Process Models (NLP4RE Report Paper).” In: (2019) (cit. on p. 22).
- (Miclo et al., 2015) R. Miclo, F. Fontanili, G. Marquès, P. Bomert, and M. Luras. “RTLS-based Process Mining: Towards an automatic process diagnosis in healthcare.” In: *2015 IEEE International Conference on Automation Science and Engineering (CASE)*. Aug. 2015, pp. 1397–1402 (cit. on pp. 34, 35, 38).
- (Miranda et al., 2019) M. A. Miranda, S. Salvatierra, I. Rodríguez, M. J. Álvarez, and V. Rodríguez. “Characterization of the flow of patients in a hospital from complex networks.” In: *Health Care Management Science* (Jan. 2019) (cit. on p. 29).

-
- (Molka et al., 2015) T. Molka, D. Redlich, M. Drobek, X.-J. Zeng, and W. Gilani. “[Diversity Guided Evolutionary Mining of Hierarchical Process Models](#).” In: *Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation*. GECCO ’15. event-place: Madrid, Spain. New York, NY, USA: ACM, 2015, pp. 1247–1254 (cit. on p. 22).
- (Montgomery, 2019) D. C. Montgomery. *[Introduction to Statistical Quality Control, 8th Edition / Industrial Engineering / Manufacturing / General & Introductory Industrial Engineering / Subjects / Wiley](#)*. en-us (cit. on pp. 10, 56, 79, 95, 96, 98, 100, 105, 107, 122, 129, 134).
- (Muzammal et al., 2018) M. Muzammal, M. Gohar, A. U. Rahman, Q. Qu, A. Ahmad, and G. Jeon. “[Trajectory Mining Using Uncertain Sensor Data](#).” In: *IEEE Access* 6 (2018), pp. 4895–4903 (cit. on pp. 33–36).
- (Namaki Araghi et al., 2018) S. Namaki Araghi, F. Fontanili., E. Lamine., L. Tancerel., and F. Benaben. “[Applying Process Mining and RTLS for Modeling, and Analyzing Patients’ Pathways](#).” In: *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 5: HEALTHINF, INSTICC*. SciTePress, 2018, pp. 540–547 (cit. on pp. 34, 35, 57, 194).
- (Nambiar et al., 2013) R. Nambiar, R. Bhardwaj, A. Sethi, and R. Vargheese. “[A look at challenges and opportunities of Big Data analytics in healthcare](#).” In: *2013 IEEE International Conference on Big Data*. Oct. 2013, pp. 17–22 (cit. on p. 18).
- (Orellana et al., 2018) A. Orellana, L. Castañeda, and A. Valladares. “[Analysis of Hospital Processes from the Time Perspective Using Process Mining](#).” In: *IEEE Latin America Transactions* 16.6 (June 2018), pp. 1741–1748 (cit. on p. 28).
- (Partington et al., 2015) A. Partington, M. Wynn, S. Suriadi, C. Ouyang, and J. Karnon. “[Process Mining for Clinical Processes: A Comparative Analysis of Four Australian Hospitals](#).” In: *ACM Trans. Manage. Inf. Syst.* 5.4 (Jan. 2015), 19:1–19:18 (cit. on p. 29).
- (Pettersson et al., 2013) E. Pettersson, B. Megyesi, and J. Nivre. “[Normalisation of Historical Text Using Context-Sensitive Weighted Levenshtein Distance and Compound Splitting](#).” In: *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*. Oslo, Norway: Linköping University Electronic Press, Sweden, May 2013, pp. 163–179 (cit. on p. 147).
- (Pinchin, 2015) J. Pinchin. “[Getting lost in hospitals costs the NHS and patients](#).” en-GB. In: (Mar. 2015) (cit. on p. 3).
- (Probst et al., 2012) H. B. Probst, Z. B. Hussain, and O. Andersen. “[Cancer patient pathways in Denmark as a joint effort between bureaucrats, health professionals and politicians—A national Danish project](#).” In: *Health Policy* 105.1 (2012), pp. 65–70 (cit. on p. 5).
- (Prodel, 2017) M. Prodel. “[Modélisation automatique et simulation de parcours de soins à partir de bases de données de santé](#).” thesis. Lyon, Apr. 2017 (cit. on pp. 16, 17).
- (Provost et al., 2013) F. Provost and T. Fawcett. *[Data Science for Business: What You Need to Know About Data Mining and Data-analytic Thinking](#)*. 1st. O’Reilly Media, Inc., 2013 (cit. on pp. 7, 15, 16).

- (Ramos et al., 2017) J. Ramos, A. César, J. Neves, and P. Novais. “[Adapting the User Path Through Trajectory Data Mining](#).” en. In: *Ambient Intelligence– Software and Applications – 8th International Symposium on Ambient Intelligence (ISAmI 2017)*. Ed. by J. F. De Paz, V. Julián, G. Villarrubia, G. Marreiros, and P. Novais. Advances in Intelligent Systems and Computing. Springer International Publishing, 2017, pp. 195–202 (cit. on pp. 34, 35, 38).
- (Rebuge et al., 2012) Á. Rebuge and D. R. Ferreira. “[Business process analysis in healthcare environments: A methodology based on process mining](#).” In: *Information Systems. Management and Engineering of Process-Aware Information Systems 37.2* (Apr. 2012), pp. 99–116 (cit. on pp. 19, 88).
- (Rojas et al., 2019) E. Rojas and D. Capurro. “[Characterization of Drug Use Patterns Using Process Mining and Temporal Abstraction Digital Phenotyping](#).” In: *Business Process Management Workshops*. Ed. by F. Daniel, Q. Z. Sheng, and H. Motahari. Cham: Springer International Publishing, 2019, pp. 187–198 (cit. on p. 26).
- (Rojas et al., 2017a) E. Rojas, C. Fernández-Llatas, V. Traver, J. Munoz-Gama, M. Sepúlveda, V. Herskovic, and D. Capurro. “PALIA-ER: Bringing Question-Driven Process Mining Closer to the Emergency Room.” In: *BPM (Demos)*. 2017 (cit. on pp. 34, 35, 38).
- (Rojas et al., 2016) E. Rojas, J. Munoz-Gama, M. Sepúlveda, and D. Capurro. “[Process mining in healthcare: A literature review](#).” In: *Journal of Biomedical Informatics* 61 (June 2016), pp. 224–236 (cit. on pp. 22, 23, 91, 118, 194).
- (Rojas et al., 2017b) E. Rojas, M. Sepúlveda, J. Munoz-Gama, D. Capurro, V. Traver, and C. Fernandez-Llatas. “[Question-Driven Methodology for Analyzing Emergency Room Processes Using Process Mining](#).” en. In: *Applied Sciences* 7.3 (Mar. 2017), p. 302 (cit. on p. 45).
- (Rozinat et al., 2008) A. Rozinat and W. van der Aalst. “[Conformance checking of processes based on monitoring real behavior](#).” In: *Information Systems* 33.1 (2008), pp. 64–95 (cit. on p. 151).
- (Rubin et al., 2007) V. Rubin, C. W. Günther, W. M. P. van der Aalst, E. Kindler, B. F. van Dongen, and W. Schäfer. “[Process Mining Framework for Software Processes](#).” In: *Software Process Dynamics and Agility*. Ed. by Q. Wang, D. Pfahl, and D. M. Raffo. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 169–181 (cit. on p. 21).
- (Salimifard et al., 2001) K. Salimifard and M. Wright. “[Petri net-based modelling of workflow systems: An overview](#).” In: *European Journal of Operational Research* 134.3 (Nov. 2001), pp. 664–676 (cit. on p. 125).
- (2019). “[ScholarWorks: A Data Analysis Methodology for Process Diagnosis and Redesign in Healthcare](#).” In: () (cit. on p. 29).
- (Senderovich et al., 2016) A. Senderovich, A. Rogge-Solti, A. Gal, J. Mendling, and A. Mandelbaum. “[The ROAD from Sensor Data to Process Instances via Interaction Mining](#).” In: *Advanced Information Systems Engineering*. Ed. by S. Nurcan, P. Soffer, M. Bajec, and J. Eder. Cham: Springer International Publishing, 2016, pp. 257–273 (cit. on pp. 33–35).
- (Shearer, 2000) C. Shearer. “The CRISP-DM Model: The New Blueprint for Data Mining.” In: *Journal of Data Warehousing* 5.4 (2000) (cit. on pp. 15, 16).
- (Song et al., 2009) M. Song, C. W. Günther, and W. M. P. van der Aalst. “[Trace Clustering in Process Mining](#).” In: *Business Process Management Workshops*. Ed. by D. Ardagna, M. Mecella, and J. Yang. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 109–120 (cit. on p. 23).

-
- (Straus et al., 2019) S. E. Straus, a. Glasziou Pau, a. Richardson W. Scot, and a. Haynes R. Bria. *Evidence-based medicine : how to practice and teach EBM*. English. Fifth edition. Edinburgh : Elsevier, 2019 (cit. on p. 22).
- (Szttyler et al., 2016) T. Szttyler, J. Carmona, J. Völker, and H. Stuckenschmidt. *Self-tracking reloaded: applying process mining to personalized health care from labeled sensor data*. 2016 (cit. on pp. 34, 35, 37).
- (Taner et al., 2007) M. T. Taner, B. Sezen, and J. Antony. "An overview of six sigma applications in healthcare industry." en. In: *International Journal of Health Care Quality Assurance* (June 2007) (cit. on p. 143).
- (Tang et al., 2015) L.-A. Tang, X. Yu, Q. Gu, J. Han, G. Jiang, A. Leung, and T. L. Porta. "A Framework of Mining Trajectories from Untrustworthy Data in Cyber-Physical System." In: *ACM Trans. Knowl. Discov. Data* 9.3 (Feb. 2015), 16:1–16:35 (cit. on pp. 34, 35, 39).
- (Tanuja et al., 2016) V. Tanuja and P. Govindarajulu. "Application of trajectory data mining techniques in CRM using movement based community clustering." In: 16.11 (2016), p. 20 (cit. on pp. 34, 35).
- (Tanuja et al., 2017) V. Tanuja and P. Govindarajulu. "A Novel Framework for Geo-clustering of User Movements based on Trajectory Data." In: 17.3 (2017), p. 212 (cit. on pp. 34, 35, 38).
- (Terragni et al., 2019) A. Terragni and M. Hassani. "Optimizing Customer Journey Using Process Mining and Sequence-aware Recommendation." In: *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*. SAC '19. event-place: Limassol, Cyprus. New York, NY, USA: ACM, 2019, pp. 57–65 (cit. on p. 27).
- (Thor et al., 2007) J. Thor, J. Lundberg, J. Ask, J. Olsson, C. Carli, K. P. Härenstam, and M. Brommels. "Application of statistical process control in healthcare improvement: systematic review." en. In: *BMJ Quality & Safety* 16.5 (Oct. 2007), pp. 387–399 (cit. on p. 129).
- (Traxler et al., 2018) B. Traxler, E. Helm, O. Krauss, A. Schuler, and J. Kueng. "Towards Semantic Interoperability in Health Data Management Facilitating Process Mining." en. In: *International Journal of Privacy and Health Information Management (IJPHIM)* 6.2 (July 2018), pp. 1–12 (cit. on p. 26).
- (Ullah et al., 2019) S. Ullah and S. Ullah. "Why do Patients Miss their Appointments at Primary Care Clinics?" en-US. In: () (cit. on p. 4).
- (Vanhaecht et al., 2010) K. Vanhaecht, W. Sermeus, J. Peers, C. Lodewijckx, S. Deneckere, F. Leigheb, M. Decramer, M. Panella, and EQCP Study Group. "The impact of care pathways for exacerbation of Chronic Obstructive Pulmonary Disease: rationale and design of a cluster randomized controlled trial." eng. In: *Trials* 11 (Nov. 2010), p. 111 (cit. on p. 5).
- (Vázquez-Barreiros et al., 2014) B. Vázquez-Barreiros, M. Mucientes, and M. Lama. "A Genetic Algorithm for Process Discovery Guided by Completeness, Precision and Simplicity." en. In: *Business Process Management*. Ed. by S. Sadiq, P. Soffer, and H. Völzer. Lecture Notes in Computer Science. Springer International Publishing, 2014, pp. 118–133 (cit. on p. 21).
- (Vázquez-Barreiros et al., 2015) B. Vázquez-Barreiros, M. Mucientes, and M. Lama. "ProDiGen: Mining complete, precise and minimal structure process models with a genetic algorithm." In: *Information Sciences*. Innovative Applications of Artificial Neural Networks in Engineering 294 (Feb. 2015), pp. 315–333 (cit. on p. 21).

- (H. M. W. Verbeek et al., 2017) H. M. W. Verbeek, W. M. P. van der Aalst, and J. Munoz-Gama. [“Divide and Conquer: A Tool Framework for Supporting Decomposed Discovery in Process Mining.”](#) en. In: *The Computer Journal* 60.11 (Nov. 2017), pp. 1649–1674 (cit. on p. 21).
- (H. M. W. (Verbeek et al., 2013) H. M. W. (Verbeek and W. M. P. van der Aalst. [“An Experimental Evaluation of Passage-Based Process Discovery.”](#) en. In: *Business Process Management Workshops*. Ed. by M. La Rosa and P. Soffer. Lecture Notes in Business Information Processing. Springer Berlin Heidelberg, 2013, pp. 205–210 (cit. on p. 21).
- (Vergidis et al., 2008) K. Vergidis, A. Tiwari, and B. Majeed. [“Business Process Analysis and Optimization: Beyond Reengineering.”](#) In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 38.1 (Jan. 2008), pp. 69–82 (cit. on pp. 9, 10).
- (Wan et al., 2017) N. Wan, G. L. Kan, and G. Wilson. [“Addressing location uncertainties in GPS-based activity monitoring: A methodological framework.”](#) en. In: *Transactions in GIS* 21.4 (2017), pp. 764–781 (cit. on pp. 33–36).
- (Weber et al., 2018) P. Weber, R. Backman, I. Litchfield, and M. Lee. [“A Process Mining and Text Analysis Approach to Analyse the Extent of Polypharmacy in Medical Prescribing.”](#) In: *2018 IEEE International Conference on Healthcare Informatics (ICHI)*. June 2018, pp. 1–11 (cit. on p. 25).
- (A. J. M. M. Weijters et al., 2003) A. J. M. M. Weijters and W. M. P. van der Aalst. [“Rediscovering workflow models from event-based data using little thumb.”](#) en. In: *Integrated Computer-Aided Engineering* 10.2 (Jan. 2003), pp. 151–162 (cit. on p. 91).
- (A. J. M. M. Weijters et al., 2011) A. J. M. M. Weijters and J. T. S. Ribeiro. [“Flexible Heuristics Miner \(FHM\).”](#) In: *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*. Apr. 2011, pp. 310–317 (cit. on p. 91).
- (A. Weijters et al., 2006) A. Weijters, W. M. P. van der Aalst, and A. A. De Medeiros. [“Process mining with the heuristics miner-algorithm.”](#) In: 166 (2006), pp. 1–34 (cit. on pp. 22, 91).
- (T. Weijters et al., 2004) T. Weijters, L. Maruster, and W. M. P. van der Aalst. [“Workflow Mining: Discovering Process Models from Event Logs.”](#) In: 16.9 (2004), pp. 1128–1142 (cit. on p. 91).
- (Williams et al., 2019) R. Williams, D. M. Ashcroft, B. Brown, E. Rojas, N. Peek, and O. Johnson. [“Process Mining in Primary Care: Avoiding Adverse Events Due to Hazardous Prescribing.”](#) eng. In: vol. 264. Aug. 2019, pp. 447–451 (cit. on p. 27).
- (Wolstenholme et al., 2004) E. Wolstenholme, D. McKelvie, G. Smith, and D. Monk. [“Using system dynamics in modelling health and social care commissioning in the UK.”](#) In: *Proceedings of the 2004 International System Dynamics Conference, Oxford, England.(CD-ROM)*. 2004 (cit. on p. 4).
- (Yahya et al., 2013) B. N. Yahya, H. Bae, S.-o. Sul, and J.-Z. Wu. [“Process Discovery by Synthesizing Activity Proximity and User’s Domain Knowledge.”](#) In: *Asia Pacific Business Process Management*. Ed. by M. Song, M. T. Wynn, and J. Liu. Cham: Springer International Publishing, 2013, pp. 92–105 (cit. on p. 21).
- (Yahya et al., 2016) B. N. Yahya, M. Song, H. Bae, S.-o. Sul, and J.-Z. Wu. [“Domain-driven actionable process model discovery.”](#) In: *Computers & Industrial Engineering* 99 (Sept. 2016), pp. 382–400 (cit. on p. 21).
- (S. Yang et al., 2018a) S. Yang, W. Ni, X. Dong, S. Chen, R. A. Farneth, A. Sarcevic, I. Marsic, and R. S. Burd. [“Intention Mining in Medical Process: A Case Study in Trauma Resuscitation.”](#) eng. In: vol. 2018. June 2018, pp. 36–43 (cit. on p. 25).

-
- (S. Yang et al., 2018b) S. Yang, F. Tao, J. Li, D. Wang, S. Chen, I. Marsic, O. Z. Ahmed, and R. S. Burd. "[Process Mining the Trauma Resuscitation Patient Cohorts.](#)" eng. In: vol. 2018. June 2018, pp. 29–35 (cit. on p. 25).
- (W. Yang et al., 2014) W. Yang and Q. Su. "[Process mining for clinical pathway: Literature review and future directions.](#)" In: *2014 11th International Conference on Service Systems and Service Management (ICSSSM)*. June 2014, pp. 1–5 (cit. on pp. 4, 28).
- (Zafari et al., 2017) F. Zafari, A. Gkelias, and K. K. Leung. "[A Survey of Indoor Localization Systems and Technologies.](#)" In: *CoRR* abs/1709.01015 (2017). arXiv: [1709.01015](#) (cit. on p. 31).
- (Zahoor et al., 2019) E. Zahoor, K. Munir, O. Perrin, and C. Godart. "[Verification of Service-Based Declarative Business Processes: A Satisfiability Solving-Based Formal Approach.](#)" en. In: 2019, pp. 155–193 (cit. on p. 22).
- (W. Zheng et al., 2019) W. Zheng, Y. Du, L. Qi, and L. Wang. "[A Method for Repairing Process Models Containing a Choice With Concurrency Structure by Using Logic Petri Nets.](#)" In: *IEEE Access* 7 (2019), pp. 13106–13120 (cit. on p. 21).
- (Y. Zheng, 2015a) Y. Zheng. "[Trajectory Data Mining: An Overview.](#)" en-US. In: *ACM Transaction on Intelligent Systems and Technology* (Sept. 2015) (cit. on pp. 34, 35).
- (Y. Zheng, 2015b) Y. Zheng. "[Trajectory Data Mining: An Overview.](#)" In: *ACM Trans. Intell. Syst. Technol.* 6.3 (May 2015), 29:1–29:41 (cit. on pp. 34, 35, 38).
- (Zhenjiang et al., 2017) D. Zhenjiang, J. Deng, J. Xiaohui, and W. Yongli. "[RTMatch: Real-time location prediction based on trajectory pattern matching.](#)" English (US). In: *Database Systems for Advanced Applications - DASFAA 2017 International Workshops: BDMS, BDQM, SeCoP, and DMMOOC, Proceedings*. Springer Verlag, Jan. 2017, pp. 103–117 (cit. on pp. 34, 35, 38).
- (Zolfaghar et al., 2013) K. Zolfaghar, N. Meadem, A. Teredesai, S. B. Roy, S. Chin, and B. Muckian. "[Big data solutions for predicting risk-of-readmission for congestive heart failure patients.](#)" In: *2013 IEEE International Conference on Big Data*. Oct. 2013, pp. 64–71 (cit. on p. 17).

Résumé

Une méthodologie de découverte et de diagnostic des processus métier basée sur les données de localisation intérieures : application à l'amélioration du parcours patients

Dans chaque organisation, les processus métier sont aujourd'hui incontournables. Cette thèse vise à développer une méthode pour les améliorer. Dans le domaine de la santé, les organisations hospitalières déploient beaucoup d'efforts pour mettre leurs processus sous contrôle, notamment à cause de la très faible marge d'erreur admise. Les parcours des patients au sein des structures de santé constituent l'application qui a été choisie pour démontrer les apports de cette méthode. Elle a pour originalité d'exploiter les données de géolocalisation des patients à l'intérieur de ces structures. Baptisée DIAG, elle améliore les parcours de soins grâce à plusieurs sous-fonctions : (i) interpréter les données de géolocalisation pour la modélisation de processus, (ii) découvrir automatiquement les processus métier, (iii) évaluer la qualité et la performance des parcours et (iv) diagnostiquer automatiquement les problèmes de performance des processus. Cette thèse propose donc les contributions suivantes : la méthode DIAG elle-même qui, grâce à quatre différents états, extrait les informations des données de géolocalisation ; le méta-modèle DIAG qui a deux utilités : d'une part, interpréter les données de géolocalisation et donc passer des données brutes aux informations utilisables, et, d'autre part contribuer à vérifier l'alignement des données avec le domaine grâce à deux méthodes de diagnostic décrites plus bas ; deux algorithmes de découverte de processus qui utilisent la stabilité statistique des logs d'événements ; une nouvelle approche de process mining utilisant SPC (Statistical Process Control) pour l'amélioration ; l'algorithme proDIST qui mesure les distances entre les modèles de processus ; deux méthodes de diagnostic automatique de processus pour détecter les causes des déviations structurelles dans des cas individuels et pour des processus communs. Le contexte de cette thèse confirme la nécessité de proposer de telles solutions. Une étude de cas dans le cadre de ce travail de recherche illustre l'applicabilité de la méthodologie DIAG et des fonctions et méthodes mentionnées.

Mot-clés : Process Mining, Systèmes de localisation en intérieur, Gestion des processus métiers, Diagnostic des processus métiers, Processus de soins.

Abstract

A methodology for business process discovery and diagnosis based on indoor location data: Application to patient pathways improvement

Business processes are everywhere and, as such, we must acknowledge them. Among all of them, hospital processes are of vital importance. Healthcare organizations invest huge amount of efforts into keeping these processes under control, as the allowed margin of error is so slight. This research work seeks to develop a methodology to endorse improvement of patient pathways inside healthcare organizations. It does so by using the indoor location data of patients. This methodology is called DIAG (Data state, Information state, Awareness, Governance). It is constructed of several different functions. The most important ones are as follows: (i) location data interpreting, (ii) automatic discovery of business process models, (iii) business process analyzing for evaluating the performance and quality of processes, and finally, (iv) automatic diagnosing of business processes. Along the former functions, the contribution of this thesis are: The DIAG methodology which, through four different states, extracts knowledge from location data; the DIAG meta-model which supports both the interpretation of location data (from raw data to usable information) and the alignment of the domain knowledge (which are used for the diagnosing methods); two process discovery algorithms which explore statistical stability in event logs. application of Statistical Process Control (SPC) for the "enhancement notation" of Process Mining; the ProDIST algorithm for measuring the distance between process models; two automatic process diagnosing methods to detect causes of structural deviations in individual cases and common processes. The state of the art in this dissertation endorses the necessity for proposing such solutions. A case study within this research work illustrates the applicability of the DIAG methodology and its mentioned functions and methods.

Keywords: Process Mining, Indoor Localization Systems, Business Process Management, Business Process Diagnosis, Healthcare Processes