Proceedings of the 2016 Winter Simulation Conference
T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, eds.

EVALUATION OF DISCOVERED CLINICAL PATHWAYS USING PROCESS MINING AND JOINT AGENT-BASED DISCRETE-EVENT SIMULATION

Vincent Augusto
Xiaolan Xie

Center for Biomedical and Healthcare Engineering
UMR CNRS 6158 LIMOS
École des Mines de Saint-Étienne
158 cours Fauriel
42023 Saint-Etienne cedex 2, FRANCE

Martin Prodel
Baptiste Jouaneton
Ludovic Lamarsalle

HEVA
186 Avenue Thiers
69465 Lyon cedex 06, FRANCE

ABSTRACT

The analysis of clinical pathways from event logs provides new insights about care processes. In this paper, we propose a new methodology to automatically perform simulation analysis of patients' clinical pathways based on a national hospital database. Process mining is used to build highly representative causal nets, which are then converted to state charts in order to be executed. A joint multi-agent discrete-event simulation approach is used to implement models. A practical case study on patients having cardiovascular diseases and eligible to receive an implantable defibrillator is provided. A design of experiments has been proposed to study the impact of medical decisions, such as implanting or not a defibrillator, on the relapse rate, the death rate and the cost. This approach has proven to be an innovative way to extract knowledge from an existing hospital database through simulation, allowing the design and test of new scenarios.

1 INTRODUCTION

In health-care, activities such as consultation, imaging examination or surgery are parts of a clinical pathway (CP). The design of a CP is a major challenge to better understand the impact of treatments on the whole journey of the patient. Health authorities intend to propose standardization of care processes for various operational purposes: organization of care activities, assignment of human resources, reducing practice variability, minimizing delays in treatments or decreasing costs while maintaining quality. The large amount of data collected from a CP by an Hospital Information Systems is valuable because it may reveal important patterns of the CP, allowing the creation of formal models that can be simulated. This paper provides a methodology in order to analyze and simulate such CPs using existing databases on the national level.

In a previous study (Prodel et al. 2015), we proposed a new approach to discover CPs from a national hospital database using Process Mining (PM) (Van der Aalst 2011) and Integer Linear Programming. The objective was to create the most representative process model of the event log under a constraint on the model's size. The metric for assessing representativeness was based on the frequency of events in the log and of direct transitions between events. Decision variables were the choices of events to keep in the nodes of the model and the arcs to keep between these nodes. In the literature, CP analysis of recorded data was mainly done using either Data Mining or Process Mining techniques. Such approaches receive an increasing attention in the field of Medical Informatics. The next step of this research consists in proposing a model that can be executed using simulation and in testing what-if scenarios. Scenarios can be related to various decisions, such as a change in the medical treatment of certain patients, the launch of new medical devices supposed to be more effective to cure certain diseases, or a change in hospital activities' financing.

In this paper, we propose a new methodology to automatically build a simulation model of patients' CP from a national hospital database using Process Mining techniques. Such methodology may be applied

978-1-5090-4486-3/16/$31.00 ©2016 IEEE

using any database as data input, and may be applied for any cohort of patients, which constitutes the main scientific contribution of this paper. Simulation of Clinical Pathways brings new knowledge and allows the evaluation of scenarios through design of experiments. Target users of our approach are numerous:

- **hospital managers**: predict the results of investments in new care services or management strategies;
- **health-care practitioners**: test the relevancy of new treatments at certain steps of the care pathway of the patients under study;
- **pharmaceutical firms**: extrapolate the impact of a new drug or a new medical device on the patient care pathway by taking into account the cost of hospital stays.

The remainder of this paper is organized as follows: a literature review related to joint use of PM and Simulation is given in Section 2. Basics of PM are described in Section 3. The methodology for building a CP simulation model from an hospital database using PM is presented in Section 4. The formal model based on a special class of state charts is given in Section 5. The case study and results are presented in Section 6. Finally, conclusions and perspectives are discussed in Section 8.

2 LITERATURE REVIEW

The goal of Process Mining is to extract new information about processes from event logs (Van der Aalst 2011). The field of Process Mining emerged in the early 2000s, when the bases were formalized (Van der Aalst 2004). Process Mining aims at providing an impartial view of a process based on what really happened, and not on a supposed organization. The use of Process Mining is motivated by two observations (Van der Aalst 2011): more and more information is stored in information systems, and techniques from Business Process Management have reached their limits. BPM techniques only study theoretical processes, unlike Process Mining which studies the actual behavior as it happened (Van der Aalst 2011). It can be applied in various systems and domains, such as health-care. The raw material of Process Mining is a specific type of data set, namely an event log. An event log is a set of traces, each trace being a sequence of ordered events (e.g., a patient is a trace and each hospitalization is an event). The only requirement is to have an event log such that (i) each event refers to an activity (i.e. a well-defined step in the process, e.g. an hospitalization to treat an heart issue), (ii) each event refers to a trace (e.g., a patient), and (iii) the events have a time stamp and are totally ordered (Van der Aalst 2011).

Here, we only focus on the Process Discovery step. It studies the process behavior, namely the activities in the process and their order of execution. It results in the creation of a process model, which is unknown beforehand and which reproduces the behavior of the recorded events. Examples of Process Discovery techniques include the Alpha-Algorithm (Van der Aalst 2004), the Fuzzy Miner (Gunther and Van der Aalst 2007) and the Heuristic Miner (Weijters et al. 2006). In our previous work (Prodel et al. 2015), we introduced a new process discovery algorithm which is more suitable to sizable and complex logs than existing approaches. It applies well to health-care data where there are almost as many different behaviors as there are patients. Our Process Mining approach allowed us to automatically build a generalized model where the most common clinical pathways stand out. The discovered model helps to describe the underlying process (care delays, treatment frequencies, deviations and death rate) as it really happened. Using this model as the starting point of a simulation model and *what if* scenarios analysis is not entirely new : Rozinat et al. (2009) used colored Petri nets to represent a simulation model found by Process Mining.

The main motivation for automating the creation of simulation model is that most simulation models are handmade models. Even if modelers are used to deal with such modeling bias, it is insufficient. Models are built using documentation, observations from the modeler and interviews of experts. This is a time consuming approach and a partial view of the processes. The perception of the actual process is influenced by the experience of the human studying it. Moreover, this approach is not easily reproducible as the model is built on a case by case basis. To avoid these biases, the idea of integrating various process mining results to automatically generate a complete simulation model was first done by Rozinat et al. (2009). Such idea of a complementary input for simulation models was then advocated by Martin et al. (2014a) and Martin et al. (2014b). In (Rozinat et al. 2009), the authors focused on the validation of a simulation model (whether generated or hand-made) since its quality is crucial for drawing conclusions from a simulation run. Finally, they highlighted challenges that are faced when discovering simulation models from event logs. It includes creating not too complex models to have usable results, including other available perspectives than the flow perspective (e.g., patients' feature or human resources), and adjusting the model for real-time simulation. They showed an example of their modeling methodology using Petri Net as the formal representation of their process models. It allows for a strong formalism of the modeling framework, but it lacks the capacity of dealing with very heterogeneous and eccentric behaviors.

Our work is driven by a health-care case study with hundreds of thousands of patients. For that reason, we need a flexible modeling framework, the Causal Net presented in the next section. The work of Zhou et al. (2014) also describes a case study of combining process mining and simulation in health-care. They used the fuzzy miner algorithm for process discovery (Gunther and Van der Aalst 2007). They specifically studied the pathway of patients during a single hospital stay, starting from the admission to the release. The process included the key steps of the process, such as *check in*, *medical consultation and diagnosis*, *waiting* and *check out*. Nevertheless, if a patient is readmitted later, he is considered as a new patient starting the process. Here, our approach intends to show the clinical pathway of patients over several months, or even years. A patient is followed over a long period of time and a national territory. It leads to a complete description of care pathways at a macroscopic view.

3 BASICS OF PROCESS MINING

This section introduces basic concepts of Process Mining used in this paper. It includes data driven concepts and the notion of a process model. The readers are referred to Van der Aalst (2011) for more details.

3.1 Event Logs

Process Mining is a data-driven approach. The goal is to extract useful information from existing data sources contained in a so-called event log. The following are formal definitions of relevant concepts including events, traces, and logs from Van der Aalst (2011) and Gunther (2009).

Definition 1 (Event) Let $A_{Event} = \{a_1, ..., a_p\}$ be a finite set of attributes (time-stamp, activity type, case ID, duration, ...), $p \in \mathbb{N}$. An event based on A_{Event} is a set of p values, one for each of the attributes. Each event is uniquely determined by the combination of all its attribute values.

Definition 2 (Trace) Let T be a set of events, a trace σ is an ordered sequence of T : $\sigma = \prec t_1, ..., t_{n_\sigma} \succ$, where $\forall i \in [\![1, n_\sigma]\!]$ (set of integer values between 1 and n_σ), $t_i \in T$. $n_\sigma \in \mathbb{N}$ is the trace's length and $\sigma(k)$ is the k^{th} element of σ's sequence. The set of all the traces over T is noted T^*.

Definition 3 (Log) Let T be a set of events, a log L over T is a non-empty set of traces over T : $L = \{\sigma_1, ..., \sigma_m\}$, $m \in \mathbb{N}$ and $\forall i \in [\![1, m]\!], \sigma_i \in T^*$. The events of a given log are defined on the same set of attributes (with different values).

Definition 4 (Event Class) Let A_{Event} be a set of attributes. An Event Class is a subset of the attribute vector space defined on A_{Event}. Let T be a set of events and C be the set of all the Event Classes. Alternatively, the function "*Class*" maps each event of T to an Event Class, $Class \in T \to C$. The set of event classes of T is $C(T) = \{Class(e) \mid e \in T\}$.

Example. Our previous work (Prodel et al. 2015) addressed a Process Mining analysis of a well-structured and exhaustive hospitalization database. It contains the record of each hospital stay for any patient in France from 2006 to 2014, that is about 15 millions of patients' stays. In the generic process mining lexicon, a patient is a trace, a stay is an event and the entire database is a log. The attributes are patients' features and medical diagnostic information. Event classes were defined by an attribute describing the medical reason of the stay: the diagnosis. This data field is filled using the 10th International Classification of Diseases. Any event was assumed unique as two stays could not happen at the same time, for the same patient and the same medical reason. However, two stays were said to be similar if they have the same class. For example, an appendicitis operation is the medical reason of the stay, but two stays may be different if they last 2 or 5 days.

Events are unique. Two events related to the execution of the same activity are not identical events. The concept of *Event Class* is introduced to describe the relations among events. Events of the same class are considered similar. This notion of class is extremely important in the search of important trace patterns. Due to the uniqueness of the events, each trace is also unique at the event level. The only way to model precisely the underlying process would be to represent each of the traces entirely, which is impracticable for systems with a huge number of traces. The concept of class will allow us to identify commonality of traces. In the remaining, events are assumed to have at least 3 attributes: a time-stamp, a trace ID and a class. After mapping each event to its class, the order of events in a trace still holds as the time-stamps remain unchanged. Hence, for a given log, a class is said to be directly followed by another if there exists at least one trace in the log for which the two classes are following each other. It is formalized as follows:

Definition 5 (Direct following relation) Let T be a set of events, L a log over T and $C(L)$ the set of event classes of L. The direct following relation among classes of L is defined as follows. Let C_1, $C_2 \in C(L)$, then C_1 is directly followed by $C_2 \iff (C_1 \Rightarrow C_2) \iff (\exists\, \sigma \in L,\, k \in [\![1,(n_\sigma - 1)]\!] \mid \sigma(k) = C_1 \wedge \sigma(k+1) = C_2)$. Then (C_1, C_2) is called a *transition* over L.

Definition 6 (Transition set) Let T be a set of events, L a log over T and $C(L)$ the set of event classes of L. The set of transitions of L is $E_{max} = \{(C_1,C_2) \in C(L) \times C(L) \mid \; C_1 \Rightarrow C_2\}$.

The direct following relation between event classes is the starting point of process discovery for most process mining algorithms (Van der Aalst 2011). For instance, the Alpha miner algorithm (Van der Aalst 2011) builds a Petri Net with all existing direct relations whereas the Heuristic miner algorithm (Weijters, Van der Aalst, and de Medeiros 2006) only considers the most frequent transitions. In this paper, we use the following relation to define evaluation metrics of process models. The previous definitions lay formal foundations of the data concept. It allows us to introduce the abstract concept of Process Model.

3.2 Process Model

A Process Model (PsM) is an abstracted and simplified way to represent a real process, i.e. an event log. It is useful if the model is representative of the log data (Van der Aalst 2011). A model is always created using a notation formalism. Several notations are available (Petri Nets, Business Process Model and Notation, Markov chain, Flowchart, Program Evaluation and Review Technique, ...). Petri nets are often used in the context of Process Mining. They are used for process discovery by the Alpha-algorithm and the region-based techniques (Van der Werf et al. 2008), and also for conformance checking (Rozinat and Van der Aalst 2008). In this paper, we choose the Causal Net of Gunther (2009).

Definition 7 (Process Model) A process model *PsM* is composed of a set N of nodes (event classes), and a set E of arcs (transitions). Let T be a set of events, L a log over T, $PsM = (N,E) = (\{n_1,...,n_x\},\{e_1,...,e_y\})$ where $\forall i \in [\![1,x]\!]$, $n_i \in C(L)$, and $\forall j \in [\![1,y]\!]$, $e_j \in E_{max}$.

Example. Let $T = \{A,B,C,D,E\}$ be a set of events and $L = \{ABCD, ABB, ABCB\}$ be a log over T containing 3 traces. A trace represents a patient's sequence of hospitalizations (e.g., A is “thorax radiography after a heart failure”, B is “lung surgery due to cancer”, etc.). Then, $PsM(L) = (\{A,B,C\},\; \{(A,B),(B,C),(B,B),(C,B)\})$ is a process model of L with 3 nodes and 4 arcs. Figure 1 gives a graphical representation of $PsM(L)$. $PsM(L)$ does not include a node labeled D, whereas this event exists in the log, because the model was constraint to have only 3 nodes and D was the least representative event in the traces (1 occurrence for 3 traces).

An advantage of this notation is to be simple to represent and straightforward to interpret. Nodes represent the tasks in the process. Arcs, connecting the nodes, represent ordering relations upon the tasks. No theoretical knowledge is required to read a model, unlike Petri Nets and BPMN. Here, all the incoming

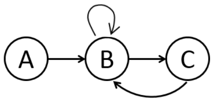


Figure 1: Example of a process model with 3 nodes and 4 arcs.

joins and the outgoing splits of the nodes are exclusive disjunction (XOR) : exactly one path is chosen in the flow. There is no need of complex structures to handle combinations of XOR/XAND splits and joins.

After defining a process model, we now present a way to evaluate its quality. Our objective is to build a Process Model metric referring to one or several of the three following dimensions : the model must be **highly representative**, it must be as **detailed** as possible, and it must have a **low complexity**. The reader is referred to Prodel et al. (2015) for an extensive definition of the quality evaluation of process models.

4 METHODOLOGY

The global methodology of the study is illustrated in Figure 2. The base material for applying the study is an hospital stays database related to a certain cohort of patients. To produce significant process models of clinical pathways, we suppose that all entities of the database are related to a common pathology (e.g. breast cancer, cardiovascular disease...). **Step (1)** consists in extracting data from the raw database in order to have a formatted data set that can be used as an input for our process mining approach. Such data set is called an event log, as described in Section 3. In **Step (2)**, a Process Mining algorithm dedicated to health-care event logs is executed to obtain a causal net (Prodel et al. 2015). Any Process Mining discovery algorithm producing causal net could be used, but a detailed description of these algorithms is out of the scope of this article. The discovered causal net can be used for a preliminary analysis. It describes the most significant clinical pathways of the cohort as it really happened. The number of nodes and arcs in the model is a parameter of the discovery algorithm. Thus, we propose several models of different sizes in output of Step (2): (i) models for analysis and communication with health practitioners with a limited number of nodes and arcs (rather small models, with less than 20 nodes), and (ii) models to be used for simulation, having a higher number of nodes and arcs (rather large models). The latest have too many nodes to be efficiently shown and used as descriptive tools and to identify unwanted medical practices. However, they can be converted into a complex simulation engine in which the final output will be key performance indicators (KPIs) (Section 5). In both cases, the node of the discovered causal nets are related to a medical diagnosis (health state of the patient) that requires a specific care process (medical decision).

Before converting the causal net into a simulation model, we perform statistical and data mining analysis in **Step (3)** to get additional information on the process (probability of following each path, patients' features impacting these probabilities, length of stays distribution). The resulting statistical distributions describe all the parameters of the model to be implemented. In **Step (4)**, a procedure takes as input the causal net and produces a state chart that will be used in simulation. The state chart describes the behavior of each patient of our cohort. The conversion procedure and the setting of the state diagram are described in Section 5. The resulting state chart is finally implemented under the shape of a joint multi-agent discrete-event simulation model in **Step (5)**. We propose a behavioral multi-agent simulation model to describe the crucial steps of the disease evolution for each patient, whereas a synchronized discrete-event simulation model is used to time the creation of agents, and to describe the clinical pathway of patients depending on the current diagnosis. Such a model is executed throughout a design of experiments and produces reports with KPIs. The simulation model implementation, design of experiment and KPIs are described in Section 6.

The whole procedure is intended to be automatic: from the original hospital database, it is possible to produce an executable simulation model that can be used to test *what-if* scenarios related to the care pathway of the considered cohort of patients. The validation of the model is guaranteed during the process mining step with the optimization of the significance value of the causal net, whereas numerical parameters are directly generated from the original database through statistical distribution fitting. For more details

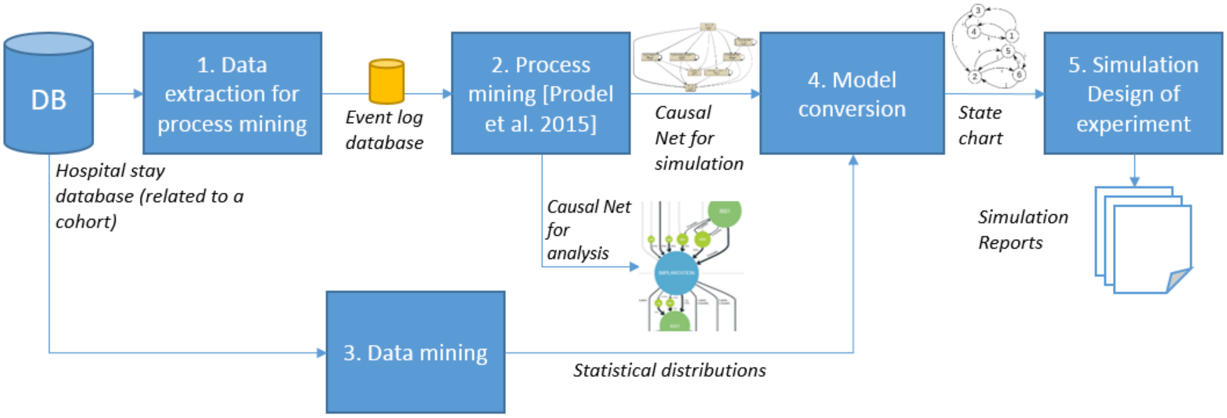

DB
1. Data extraction for process mining
Event log database
2. Process mining [Prodel et al. 2015]
Causal Net for simulation
4. Model conversion
State chart
5. Simulation Design of experiment
Hospital stay database (related to a cohort)
Causal Net for analysis
Simulation Reports
3. Data mining
Statistical distributions

Figure 2: Global methodology.

about the data extraction and the Process Mining approach, see (Prodel et al. 2015). The conversion procedure used to produce a state chart for simulation is detailed in the following section.

5 FORMAL MODELING USING STATE CHARTS FOR CLINICAL PATHWAY SIMULATION

In this section, we propose a step by step conversion procedure in order to produce a simulation model using as input a causal net process model (Step 4 of Figure 2).

5.1 State Chart Structure

In order to simulate the clinical care pathway of patients, we define the following sub-class of state charts:

Definition 8 (State Chart) A state chart (SC) is a 4-tuple $M = (S, V, \zeta, \tau)$ where $S = \{s_1, s_2, \dots, s_n\}$ is a finite set of states, $V \subseteq (S \times S)$ is a finite set of transitions, $\zeta : V \rightarrow [0,1]$ is a the probability of activating a transition, and $\tau : S \rightarrow \mathbb{N}$ is the length of stay in a state.

A state chart is used to model the clinical pathway of a patient. A patient is modeled using the concept of entity (or agent). Each entity is defined by a list of parameters and an active state.

Definition 9 (Entity) An entity is a 2-tuple $u = (A, s)$, with A a list of attributes and s its current state.

Two types of states are defined to distinguish states related to a stay in a hospital and states related to a waiting period between two hospital stays.

Definition 10 (Care-state) A Care-state is a 2-tuple $p^s = (l, B)$ where $l \in \mathbb{L}$ is a unique label and $B = \{(p_1, v_1), \dots, (p_n, v_n)\}$ is the list of parameters to be updated in this state. It includes a state-related cost that will be used as a performance indicator.

A Care-state is related to a change in the patient's health condition and requires a medical response process. During this process, the entity's attributes may change according to set B.

Definition 11 (Wait-state) A Wait-state is a singleton $p^w = (l)$ where $l \in \mathbb{L}$ is a unique label.

Finally, we propose a new subclass of state chart to describe the clinical care pathway of patients, denoted Clinical Pathway State Chart.

Definition 12 (CP State Chart) A Clinical Pathway State Chart (CPSC) is a state chart M whose underlying graph is a bigraph (bipartite graph) having the 3 following properties:

- The two types of states of the bipartite graph are Care-states and Wait-states. By definition, two states of the same type cannot be linked by a transition. E.g., after a Care-state "surgery during an hospitalization for heart failure" there could be 3 wait states, one for "die", one for "recover partially" and one for "recover fully".

- Each Wait-state has exactly one input transition and one output transition: the probability of activating the output transition of a Wait-state is always equal to 1.
- The sum of all probabilities of output transitions of a Care-state is equal to 1.

According to Definition 12, a Care-state may be followed by $\{0;n\}$ Wait-state(s), meaning one of n options will be realized according to a probability distribution. A Wait-state is always followed by exactly one Care-state: the probability of the transition between a Wait-state and a Care-state is equal to 1.

5.2 Process Model to CP State Chart Conversion Procedure

The conversion procedure described below is used to create the structure of the state chart.

Input: A process model PsM composed of a set N of nodes and a set E of arcs.
Output: A state chart SC.

1. Initialization: Let S be the set of states and V be the set of transitions. S and V are empty.
2. For each node $n \in N$, add a Care-state p_n^c to set S.
3. For each arc $e \in E$ having $n \in N$ (resp. $m \in N$) as origin node (resp. destination node): (i) add a Wait-state p_e^w having the Care-state p_n^c (resp. p_m^c) related to node n (resp. m) as predecessor (resp. successor) to set S; (ii) add transitions $\{(p_n^c, p_e^w), (p_e^w, p_m^c)\}$ to set V.

Using such an algorithm, the resulting state chart SC is a CPSC as formally defined in Definition 12. In order to execute a CPSC in a multi-agent simulation model, we need to define functions ζ (transition probabilities) and τ (length of stay in each state of the CPSC).

5.3 Clinical Pathway State Chart Setting

The CPSC setting is related to the functions ζ and τ. As Care-state i of CPSC is followed by $\{0 \ldots j \ldots n\}$ Wait-states, $\zeta(i, j)$ is the set of n probabilities of firing the transition to each of the n Wait-states. If $n > 0$, then the sum of the probabilities of firing each transition is equal to 1: $\sum_{j=1}^{n} \zeta(i, j) = 1$. $\tau(i)$ represents the random distribution of length duration in Care-state i and is obtained using the best fitting distribution on the data. Both ζ and τ functions are determined using data history from the hospital database described in the case study. Let $Care_i$ be a Care-state and $Wait_j$ be a Wait-state. After listing all the possible Wait-states after $Care_i$ as observed in the data, we can compute $\zeta(i, j)$: the probability of firing the transition between $Care_i$ and $Wait_j$ is equal to the number of patients that had the transition $Care_i \Rightarrow Wait_j$ ($Care_i$ is directly followed by $Wait_j$) divided by the total number of patients that had $Care_i$.

5.4 Example

We consider the process model given in Figure 3. This process model is formally defined as a causal net by the sets $N = \{s_1, \ldots, s_6\}$ and $E = \{e_1, \ldots, e_{10}\}$ (6 nodes and 10 edges). The conversion procedure produces the CP State Chart $\{S, V, \zeta, \tau\}$ with:

- $S = \{s_1^c, \ldots, s_5^c, s_1^w, \ldots, s_{10}^w\}$ where state s_i^c is a Care-state related to node s_i and state s_i^w is a Wait-state related to edge e_i. Care-states refer to hospital stays and Wait-states to waiting between two stays.
- $V = \{(s_1^c, s_1^w), (s_1^w, s_2^c), (s_1^c, s_2^w), (s_2^w, s_2^c), \ldots, (s_{10}^w, s_4^c)\}$

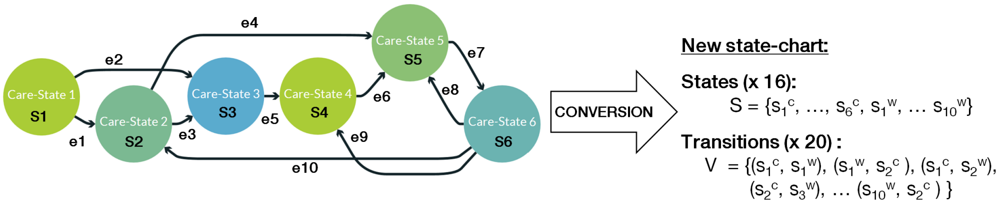


Figure 3: Process model example : a causal net is made of nodes s_i and of edges e_j.

6 CASE STUDY

This section presents the application of our previously defined methodology on a practical health-care case study. First, we introduce the medical context of cardiovascular diseases. Then, we describe the data set used to build a causal net and its conversion into a Clinical Pathway State Chart. Finally, we design a simulation experiment to assess the impact of medical leverages on the care performance.

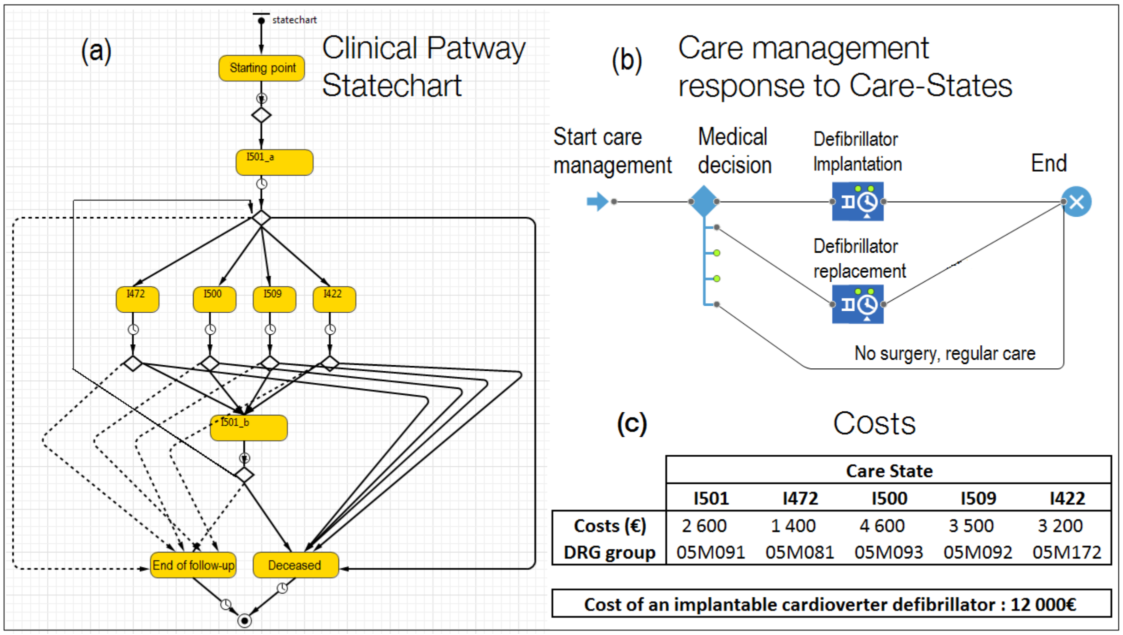
6.1 Cardiovascular Diseases and Implantable Defibrillators

Heart disease is one of the major health problems today. It was ranked as the first leading cause of death in the world in 2012 by the WHO. More specifically, cardiac arrhythmia is the most important cause of sudden cardiac death, affecting about 40,000 people per year in France and 300,300 in the USA. Implantable Cardioverter Defibrillators (ICDs) are medical devices that are indicated in prevention of cardiac arrest due to a ventricular tachycardia (or after a survived episode). According to the French PMSI database, the use of ICDs has increased in France during the last decade, from 5,300 new patients implanted in 2006 to 9,200 in 2013. Cardiologists know the main steps of these patients' care pathway : a severe heart failure, the device implantation, potential postoperative complications, the device replacement, another heart failure and sometimes death (high mortality is observed). These steps are indeed observed in the discovered process model. Here, we study all the patients who had a severe heart failure in 2008, followed or not by a defibrillator implantation. Our goal is to simulate scenarios in which a broader part of patients with heart condition are chosen for implantation. We will assess the impact on the global cost and the quality of care.

6.2 Data Collection

Data used here were obtained from a single database : the database of all hospitalizations in France (Data provided by the French Technical Agency for Hospitalization Information [ATIH], accreditation number 2015-111111-56-18, database number M14N056, M14L056). This database contains the record of all the hospital stays from 2006 to 2014 included, both in public and private sectors. It represents 27 million hospital stays and 11 million unique patients annually.

We first extracted all the hospital stays caused by a heart failure in 2008. It was done using the 10th International Classification of Disease. The medical diagnosis of heart failure is coded 'I50'. We obtained 152,393 patients. The patients were followed from this first 2008 stay until the end of 2014. We extracted all their hospital stays, all causes included (heart related or not). The total number of stays was 997,648. Using a label mapping based on the medical diagnosis, these events were grouped into 6,912 different classes. So many classes makes the Process Mining analysis computationally unfeasible. Further insights in raw data showed that a small number of classes represented a major part of the stays. Here, as a first approach in converting a Process Mining model into a simulation model, we decided to drastically filter the data to focus only on the main behaviors. Only the 8 most frequent classes of stays were kept in the final refined data set, which maps 31% of stays. We intend to include more classes in future works.



	Care State				
	I501	I472	I500	I509	I422
Costs (€)	2 600	1 400	4 600	3 500	3 200
DRG group	05M091	05M081	05M093	05M092	05M172

Cost of an implantable cardioverter defibrillator : 12 000€

Figure 4: (a) State Chart used in the heart failure case study (Anylogic software screenshots), (b) Care process triggered by certain Care-state, (c) Table of the costs (based on http://www.aideaucodage.fr/ghm).

6.3 State Chart Creation

After the data extraction, we created a Causal Net using a Process Mining approach, as described in (Prodel et al. 2015). The State Chart obtained using our conversion algorithm was implemented in Anylogic 7.2.0 software and is shown on Figure 4. Figure 4-a shows the CPSC. Its 8 Care-states are depicted by yellow boxes (first yellow box is excluded as it is the common entry point) : (1) I501a for the *first left ventricular failure*, (2) I472 for *ventricular tachycardia*, (3) I500 for *congestive heart failure*, (4) I509 for *unspecified heart failure*, (5) I422 for *other hypertrophic cardiomyopathy*, (6) I501b for a *relapse of left ventricular failure*, (7) *deceased* and (8) *end of follow-up*. A financial cost is assigned to each Care-States. It includes human, material and facility costs to take care of the patient during his/her hospitalization. The 20 Wait-states are depicted by arrows (solid and dotted lines). On Figure 4-b, we show a simple model of the medical decision triggered when a patient is in one of 4 specific Care-states (1,2,3,4). A physician decides to implant or not a cardioverter defibrillator to the patient to prevent a cardiac arrest. If the patient was previously implanted, the physician may decide to replace the device and to implant a new one (lifetime of a defibrillator is 5 to 8 years). The third possible decision is not to implant the patient but to hospitalize him/her and to provide regular care (nursing, monitoring and drugs). Decisions of implantation and replacement are based on probabilities observed from data history. These 2 probabilities will be studied as variable inputs in the simulation experiments.

The last components of CPSC to define are the functions ζ and τ. Transition probabilities represent the risk for a patient to switch from his current state to another state. It is the risk of being readmitted at hospital later for another issue. In terms of patient's health condition, lower probabilities are always better. Only the transition probability toward Care-state number (8)-*end of follow-up* shall be high for a better outcome : the patient was cured and will have no more adverse event. Implanted Patients do not have significantly different length of stays compared to never implanted patients, so τ is assumed the same for all patients. However, major differences in transition probabilities were observed between the 3 groups of patients: those never implanted, those implanted once and those implanted and replaced. Data history shows that implanted patients have lower risk of readmission compared to not implanted patients. It tends to show that implantable defibrillators have positive effects on the patients' health condition. Similarly, replaced patients have a slightly lower risk of readmission than implanted patients. The underlying reasons might be correlated to factors out of the scope of the current study (age of first implantation, device lifetime or type of technology). We used different transition probabilities for the 3 groups. Distributions used for τ were found using a best fit tool based on the mean squared error. Distributions for each of the 8 Care-states are (unit is days): (1) $371 \times Beta(1.2, 46.4)$, (2) $213 \times Beta(0.688, 24.1)$, (3) $Weibull(11.6, 1.18)$, (4) $-0.5 + Weibull(12.4, 1.28)$, (5) $-0.5 + LogN(5.86, 8.22)$, (6) $-0.5 + Erlang(4.77, 2)$, Care-states (7) and (8) have no distribution by definition.

6.4 Experimentation and Results

The previously defined Clinical Pathway State Chart models the evolution of patients' health condition in heart failure. It also models the medical decision to implant or not a patient with a defibrillator. Such decision impacts on the probability of adverse events. An implantable cardioverter defibrillator costs between 10,000 to 16,000 euros, whereas hospitalizations for a heart issue cost between 1,000 to 4,000 euros depending on the severity (See Figure 4-c). We now show how we used our simulation model to study several care management scenarios that balance costs and care quality. The performance of scenarios is assessed by evaluating 3 key performance indicators (KPIs): (1) total cost incurred by all patients, (2) death rate and (3) proportion of patients who had a heart failure relapse. All three KPIs are measured after a fixed simulation time of 5 years.

We specifically studied the variation of two parameters : **first implantation probability** (i.e. the medical decision to implant the device when in Care-States (2)(3)(4)(5)) and **replacement probability** (i.e. the medical decision to replace the device in the same Care-states). Both parameters varied between 0 and 0.5 with a 0.1 step. In order to ensure the statistical validity of our results, we performed several replications of each parameters setting. The number of replications was chosen large enough to ensure a 95% confidence interval on the 3 KPIs. It is set to 20. Results of the simulation runs are shown in Figure 5.

Scenario	Implantation Probability	Replacement probability	Total cost (€)	Death rate (%)	Hear failure relapse
1	0,3	0,05	335244700	33,41%	23890
2	0,5	0,05	374559180	33,41%	23888
3	0,5	0,05	365288560	33,41%	23888
4	0,1	0,1	412749240	33,40%	23886
5	0,2	0,1	383460280	33,41%	23888
6	0,2	0,1	367166560	33,40%	23870
7	0,5	0,1	356303940	33,40%	23876
8	0,10	0,15	346010720	33,40%	23880
9	0,20	0,15	394767820	33,39%	23879
10	0,30	0,15	385351760	33,39%	23876
11	0,40	0,15	406123420	33,39%	23865
12	0,50	0,15	435296100	33,38%	23862
13	0,1	0,2	387164280	33,39%	23862
14	0,1	0,2	365774400	33,38%	23886
15	0,2	0,2	307058400	33,38%	23898
16	0,3	0,2	356332580	33,39%	23884
17	0,4	0,2	409677080	33,40%	23837
18	0,4	0,2	429235600	33,41%	23828
19	0,4	0,2	438516340	33,40%	23842
20	0,5	0,2	398310720	33,40%	23860

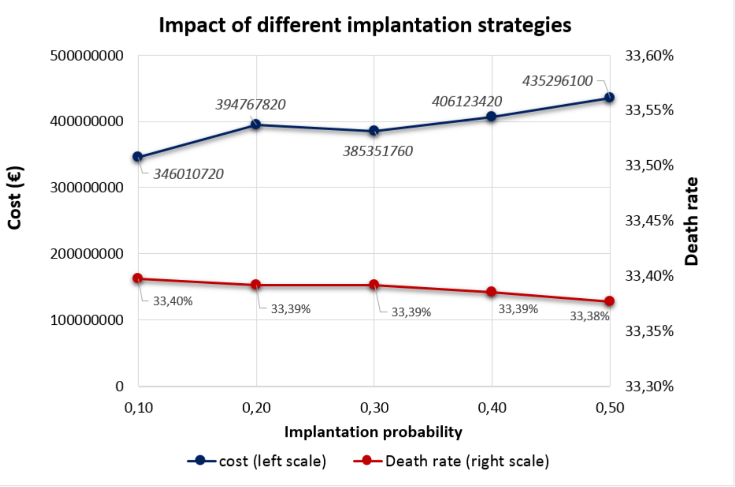


Figure 5: Simulation results : measure of 3 KPI (cost, death rate and heart failure relapse) for different values of implantation and replacement probabilities. Each simulation was done with 40,000 patients.

Numerical results validate our modeling approach and the balance mechanism between costs and care quality. When the implantation rate increases, the total cost follows because of the device's cost. It also slightly decreases the death rate (significant difference only between extreme scenarios). No significant reduction of heart failure relapse was observed. The decreasing trend in death rate is slow compared to the increase in cost. It shows that the current model reaches its limits and is not rich and complex enough to capture all the mechanisms at work. These results are preliminaries for deeper experimentation. Our goal here was to validate the concept of an innovative way of converting Process Mining results into a simulation model. This objective was reached and this work is a good starting point for further investigations.

7 DISCUSSION

Our methodology relies on a simulation model which has several advantages compared to optimization methods. It brings flexibility for decision makers to test what-if scenarios. It helps identifying the impact of different policies prior to implementation in clinical practice without negatively impacting patient outcomes. It is also suitable to computationally simulate large cohorts of patients (few hundreds of thousands of patients). It also captures the stochastic aspects of a new simulated patient's clinical pathways thanks to probabilistic distributions of duration and of the choice of a pathway. Our simulation model also has limitations. First, we limited the number of even classes to 8, which is extremely restrictive. We plan to include tens of classes in future works. In the State Chart structure of the case study, it was assumed that $\zeta(i,j)$ is independent of $\tau(i)$. It is a strong assumption since in practice state transition probability from Care-state i to Wait-state (i,j) may depend on the length of stay in Care-state i. We also assumed that τ was not patient-dependent. Patients with the same disease condition may however have different stays in the given care-state. According to us, these 2 assumptions smooth specific behaviors and tend to model a global cohort's behavior. More advanced models, with clinical use ambitions, shall capture such information.

8 CONCLUSIONS AND PERSPECTIVES

In this paper we proposed a new methodology to automatically build a simulation model of patient clinical pathway from a raw national hospital database using process mining techniques. The process mining approach produces a causal net which is converted into a formal state chart used to model the health state evolution of a patient and the associated care pathways. We defined a new sub-class of state chart called Clinical Pathway State Chart (CPSC). The model is implemented in a joint agent-based discrete-event simulation model: in the agent-based simulation model, each patient is modeled as a behavioral agent with a CPSC as behavior model; activation of a new state triggers a discrete-event simulation model which describes the care pathway of the patient in the hospital and associated medical decisions that may update the health state (i.e. parameters of the agent). The simulation model has been tested on a real case study related to the care pathway of patients with cardiac diseases. A design of experiment was conducted to study the impact of a variation in the medical decision process on the whole cohort. For future works we intend to update the model with the following elements: (i) add more parameters to describe the health status of a patient and more decisions related to the care process of the patient; (ii) replace fixed probabilities in the state chart by dynamic decision trees built upon the database; (iii) include data from consultations with general practitioners and specialists outside of the hospital. Finally, the methodology should be implemented in a standalone decision tool that may be used for any type of cohort.

REFERENCES

Gunther, C. W. 2009. *Process Mining in flexible environments*. Ph. D. thesis, Eindhoven University of Technology.

Gunther, C. W., and W. Van der Aalst. 2007. "Fuzzy Mining: Adaptive process simplification based on multi perspective metrics". In *Business Process Management*, Volume 4714, 328–343.

Martin, N., B. Depaire, and A. Caris. 2014a, Aug. "Event log knowledge as a complementary simulation model construction input". In *Simulation and Modeling Methodologies, Technologies and Applications (SIMULTECH), 2014 International Conference on*, 456–462.

Martin, N., B. Depaire, and A. Caris. 2014b, Dec. "The use of Process Mining in a business process simulation context: Overview and challenges". In *Computational Intelligence and Data Mining (CIDM), 2014 IEEE Symposium on*, 381–388.

Prodel, M., V. Augusto, X. Xie, B. Jouaneton, and L. Lamarsalle. 2015, Aug. "Discovery of patient pathways from a national hospital database using Process Mining and integer linear programming". In *Automation Science and Engineering (CASE), 2015 IEEE International Conference on*, 1409–1414.

Rozinat, A., R. S. Mans, M. Song, and W. Van der Aalst. 2009, May. "Discovering simulation models". *Inf. Syst.* 34 (3): 305–327.

Rozinat, A., and W. Van der Aalst. 2008. "Conformance checking of processes based on monitoring real behavior". *Information Systems* 33 (1): 64 – 95.

Van der Aalst, W. 2004. "Workflow mining: Discovering process models from event logs". *Computers in industry* 16:1128–1142.

Van der Aalst, W. 2011. *Process Mining: Discovery, conformance and enhancement of business processes*. 1st ed. Springer Publishing Company, Incorporated.

Van der Werf, J., B. Van Dongen, C. Hurkens, and A. Serebrenik. 2008. "Process Discovery using integer linear programming". In *Applications and Theory of Petri Nets*, Volume 5062 of *Lecture Notes in Computer Science*, 368–387. Springer Berlin Heidelberg.

Weijters, A., W. Van der Aalst, and A. de Medeiros. 2006. "Process Mining with the Heuristics Miner-algorithm". *BETA Working Paper Series, Eindhoven University of Technology* WP 166:1–34.

Zhou, Z., Y. Wang, and L. Li. 2014, April. "Process Mining based modeling and analysis of workflows in clinical care - A case study in a chicago outpatient clinic". In *Networking, Sensing and Control (ICNSC), 2014 IEEE 11th International Conference on*, 590–595.

AUTHOR BIOGRAPHIES

MARTIN PRODEL is currently a Ph.D. Student at the company HEVA and the Center for Health Engineering at Mines Saint-Etienne (ENSMSE), France. He received both his engineering degree and his Msc degree in Industrial Engineering from the ENSMSE in 2013. His email address is martin.prodel@emse.fr.

VINCENT AUGUSTO received his Ph.D. degree from the École Nationale Superieure des Mines de Saint-Étienne (EMSE), France, in 2008 and his Habilitation à Diriger des Recherches degree from the Jean Monnet University, in 2016. Currently, he is a professor of industrial engineering in the Center for Health Engineering and in the IEOR team of CNRS UMR 6158 LIMOS, EMSE. His research interests include modeling, simulation, optimization of health care systems and their supply chains. His e-mail and web addresses are augusto@emse.fr and http://www.emse.fr/~augusto, respectively.

XIAOLAN XIE received his Ph.D degree from the University of Nancy I, Nancy, France, in 1989, and the Habilitation à Diriger des Recherches degree from the University of Metz, France, in 1995. Currently, he is a distinguished professor of industrial engineering, the head of the department of Healthcare Engineering of the Center for Health Engineering and the head of IEOR team of CNRS UMR 6158 LIMOS, Mines Saint-Etienne, France. He is also a chair professor and director of the Center for Healthcare Engineering at the Shanghai Jiao Tong University, China. His research interests include design, planning and scheduling, supply chain optimization, and performance evaluation of health-care and manufacturing systems. His email address is xie@emse.fr.

BAPTISTE JOUANETON received a MSc in Computer Science in 2006 from the Université Lumière Lyon-II, France. He works as a Data Manager at the company HEVA since 10 years. He has deep knowledge about healthcare data management. His email address is bjouaneton@hevaweb.com.

LUDOVIC LAMARSALLE is both a Pharmacist (Pharm.D) and a Health economist (Msc). He created the company HEVA in 2005. HEVA is a company specialized in statistical analysis and visualization of health-care data (http://www.hevaweb.com). His email address is llamarsalle@hevaweb.com.