Optimal Process Mining of Traces With Events and Transition Attributes With Application to Care Pathways of Cancer Patients

Zhihao Peng, Vincent Augusto, Lionel Perrier, and Xiaolan Xie, *Fellow, IEEE*

***Abstract*— Contrary to event traces considered in traditional process mining literature, this paper addresses the problem of optimal process mining of traces of events and attributes associated with transitions. The problem is formally defined with rigorous description of the input event logs, the output process model, the event game specifying the images of traces in the model, and a non standard quality metric termed relevance for both the model and all model components. A dynamic programming algorithm is proposed to determine the optimal event game of each trace for a given process model. A multi-start local optimization algorithm built on an original concept of marginal relevance measure is developed for process model optimization. The proposed algorithm is shown to outperform benchmark algorithms on 40 generated test instances and be able to produce near optimal process model with an optimality gap of less than 4.46%. Results of this paper are also applied to a real case study of the care pathways of sarcoma patients. The event log representation is shown to be able to describe accurately the impact of the health state on the care pathways with only minor model relevance degradation. The proposed approach is shown to be able to generate process model at various precision levels and to compare the care pathways of cancer patients. It is also shown to generate better process model than the widely used process mining tools Disco and DFvM on both our relevance and the traditional fitness quality metrics.**

***Note to Practitioners*—This paper is motivated by our collaboration with the French cancer centre (Centre Léon Bérard) on data-driven modeling of sarcoma patient care pathways. The primary goal is to investigate the impact of patient health state such as cancer progression on the care pathways. We achieve this by original representation of care pathways by traces of events interleaved by health states. The original concept of "relevance" clearly measures the importance of each element in the process model. The faithfulness of the process model and its complexity can be easily controlled by precision parameters including least significance level of each model element and the number of layers of the model. A case study of Sarcoma patients is presented to show the importance of our care pathway representation, the superiority of our process mining algorithm, the difference of care pathways of four different patient management strategies, and how the health condition intervenes in different strategies.**

***Index Terms*— Process mining, traces of events and transition attributes, optimal event game, process model optimization, cancer care pathways.**

I. Introduction

THIS paper addresses the problem of identifying the optimal process model underlying a given set of event traces, each characterized by a sequence of events and attributes associated with transitions. The joint consideration of events and transition attributes is a major departure of this paper from the existing literature of process mining which mainly address event logs with sequences of events. Another important departure is the rigorous characterization of the contribution of each process model component to the overall goodness score of the model. This allows us to define the optimal process model as the optimal process model composed only of meaningful enough components (nodes, arcs, attributes). We also recognize the role of event game, i.e. the way an event trace is played in a process model. To summarize, the problem considered in this paper consists in jointly determining the process model and the event game in order to maximize the goodness score subject to meaningfulness of each process model component.

The joint consideration of events and attributes is motivated by our healthcare application of care pathways of cancer patients, More precisely, the application focus on sarcomas, a heterogeneous group of rare malignant tumors [1]. Once the histological diagnosis is established, the care pathway of each patient is characterized by a sequence of medical treatments (sarcoma multidisciplinary tumor boards (RCP)) labelled NETSARC, surgery which remains the principal therapeutic modality, radiotherapy, adjuvant and/or neoadjuvant

Manuscript received 19 April 2023; revised 15 June 2023; accepted 5 July 2023. This article was recommended for publication by Associate Editor H. K. Lee and Editor L. Moench upon evaluation of the reviewers' comments. This work was supported in part by the French Ministry of Health through Programme de Recherche Médico Economique under Grant PRME-18-0162 and in part by the National Natural Science Foundation of China under Grant 72192822. *(Corresponding author: Xiaolan Xie.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the French National Committee on Informatics and Privacy (Commission Nationale de l'Informatique et des Libertés) under Approval No. DR-2021-035.

Zhihao Peng and Vincent Augusto are with the Centre CIS, CNRS, UMR 6158 LIMOS, Mines Saint-Étienne, Université Clermont Auvergne, 42023 Saint-Étienne, France.

Lionel Perrier is with the Léon Bérard Cancer Centre, GATE UMR 5824, University of Lyon, 69008 Lyon, France, and also with the Human and Social Science Department, Centre Léon Bérard, 69008 Lyon, France.

Xiaolan Xie is with the Centre CIS, CNRS, UMR 6158 LIMOS, Mines Saint-Étienne, Université Clermont Auvergne, 42023 Saint-Etienne, France, and also with the Antai College of Economics and Management, Shanghai Jiao Tong University, Shanghai 200000, China (e-mail: xie@emse.fr).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TASE.2023.3295947.

Digital Object Identifier 10.1109/TASE.2023.3295947

chemotherapy, etc.). Such care pathway should be adapted according to the evolution of the health state of the patient such as local recurrence, disease progression, etc. Trace of events interleaved by health states is a simple yet meaningful extension of tradition event traces to capture the impacts of health states on care pathways. We hope to clearly show the role of health states in the healthcare pathway models.

Process mining deals with a huge amount of data in order to discover underlying process, to check the adherence to CPGs and to predict the healthcare pathways. According to [2], traditional data-centric analysis techniques like machine learning and data mining are not suitable for discovering such process models since they focus on data and local decision making instead of end-to-end processes. However, the traditional process-centric tools often disconnect from actual event data. Process mining was thus proposed to bridge this gap by considering the two aspects.

Process mining of care pathways addressed in this paper presents some unique features and challenges. Most existing literature on process mining are motivated by business process and workflow management. The long term care pathway addressed in this paper describes the course of cares from cancer diagnostic to death. Whereas activities of the same type are traditionally represent by the same node (places/transitions), two care activities of the same type such surgeries at different phase of the pathway imply rather the degradation of patient's health state than simple activity repetition. As a result, the care pathway process mining should has the capability of revealing the forward evolution of the care pathways (represented by a layered acyclic process model here). Another important requirement is the ability to reveal the relationship between care pathways and patient's health state (represented by transition attributes here). A third requirement is the explanable meaningfulness of process model components. Whereas the traditional quality measures such as fitness, precision and structureness are not easily understandable by health professionals, we propose a non standard fitness-like measure called relevance to measure the amount of information of the event log represented by each process model component.

When applying process mining in model discovery, a good representation of the process model should firstly be chosen. In this paper, the process model is represented by a graph in which each node is an event and an arc shows the transition between two events. In the application context, same events may repeat several times in a healthcare pathway and cycles may be generated if the same node is used to represent the same events. The concept of layers [3] is used to avoid cycles. The repeated events can be represented by the same node but at different layers. As to the transition attribute representation, two possible options seems evident: 1) associate each event with a related attribute and create an event-attribute joint event; 2) treat the attribute as an unique event. However, these two representations may cause a loss of information since a node should capture an event and an attribute at the same time. In order to show evolution of attributes and maximize the captured information, each arc can be converted into an edited arc to capture the possible attributes between the related transition. The advantage of this representation will be illustrated in section VII.

The contributions of this paper are listed as follows:

- A formal process mining framework with original event log representation, original process model, event game and original relevance measures of all model components;
- Joint optimization of process models and event games;
- An exact dynamic programming algorithm for relevance maximization of event traces for a given process model;
- A multi-start process model local optimization algorithm built on an original marginal relevance of new nodes;
- First application of process mining to the care pathways of Sarcoma patients with the ability to show the impact of the health state on their care pathways.

The rest of this paper is organised as follows. Section two is dedicated to a brief review of relevant process mining literature. Section three presents the mathematical model of the studied problem. Section four shows the optimal event game of each trace for a given process model. Section five presents the proposed process model optimization algorithm. Section six gives the numeric results on generated test instances. Section seven shows the comparisons on the real cases. Section eight is a conclusion.

II. Literature Review

This section gives a brief literature review regarding the basic notions in process mining, with a focus on model discovery, and the existing gap in the literature.

A. Process Mining

Using an event log as an input, process mining aims at analyzing the underlying process. An event log contains a number of traces and each trace is a sequence of events marked with timestamps. A process model captures the most meaningful traces and gives an overall view on how the process is being executed in real life. Most meaningful can be either most frequent and least frequent and it depends on what the user investigates.

According to [4], we distinguish three disciplines of process mining: Discovery, Conformance checking, and Enhancement. The first one aims at discovering a process model from an event log in order to show representative process models. The second one captures the variations between the processes in an event log and an existing process model. The last one tries to enrich and improve an existing process model.

Since it was first introduced by Wil van der Aalst [5] 20 years ago, process mining has drawn a lot attentions from different domains [6]. In healthcare, recent literature reviews [7], [8], [9] illustrate the distinguishing characteristics of the healthcare domain as well as the challenges to be addressed, in particular the integration of additional medical information to process models. Another systematic review [10] has shown the wide application of process mining in healthcare, especially in oncology [11], [12], [13]. A literature review on process mining in oncology has been conducted by [14]. To our best knowledge, Sarcoma is still a type of

cancer to be exploited with process mining techniques due to in particular numerous histological subtypes and the complex disease management [15], [16].

B. Process Model Discovery in Process Mining

In process model discovery, various algorithms have been developed: [17] can be seen as the first paper to formalize process mining. The alpha-algorithm was proposed to extract a workflow model from a workflow log. Petri nets were used to represent the process model. The genetic miner [18] was developed in attempt to give the most appropriate Petri net model when the input event log is noised and has missing events. Reference [19] presented an heuristic miner to deal with noise and low frequent behaviour. A new process modelling so-called language "Causal Matrices" was also introduced in this paper. Reference [20] proposed the fuzzy miner in which they aggregated the lower significance classes in order to limit the number of nodes to display. The split miner is presented in [21]. The algorithm first generated a Directly-Follows Graph (DFG) and detect self-loops and short loops in the model. Then concurrency relations between tasks were analyzed and a pruned DFG was created. A filtering was then applied followed by the discovery of split and join gateways. A comparison with the state-of-the-art methods including Fodina Miner [22] and Inductive Miner [23] to show the superiority of their method. Optimization tools like Integer Linear Programming (ILP) [24], [25] also showed the interests of their application in process model discovery. Recently, novel discovery algorithms were developed to handle with more advanced process models. In [26], a Cross-department Collaborative Healthcare Process (CCHP) model was proposed to study the workflow collaboration between different medical departments. The authors proposed a discovery framework in which intra-department process models and collaboration patterns were mined and integrated into the global CCHP model. Hierarchical business processes were studied in [27] and were formalized in Hierarchical Petri Nets (HPNs). The proposed model considered the hierarchical structure between a subprocess and its original one. The link is described as the same activities from the two models and different abstraction levels for these activities are considered. The authors then gave a new algorithm for the discovery of HPNs model from event logs with lifecircle information.

C. Closely Related Works

The most related papers to our work are: [28], and [3]. Reference [28] proposed a compact care pathway model by grouping the related events into a general one. However, the event repetition was not taken into account leading to cyclic process models. This aspect was improved in [3] by introducing the notion of "layer". In both, the quality of the process model is evaluated by a replayability score with some given rule for replayability of traces in the process model. Process models are determined under some complexity constraints given by the maximal number of nodes and arcs. These papers does not meet the requirements of our case study for several aspects: transition attributes (health state) not taken into account, lack of the meaningfulness measures of process model components, heuristic replayability rule.

A combination of split miner and deep learning was proposed in [29], [30], and [31] to discover the care pathway models of ICU patients and to predict the next event such as readmission and death. The resulting process model contains cycles and does not allow the forward evolution modelling needed for our study. Further, the resulting Petri net model suffers from similar weakness of [3] and [28]. Especially, the Petri net model makes the explanability of the meaningfulness measures of the model components to practitioners. Finally, the objective differs and is prediction in their papers and understanding care pathway model here.

This paper differs from the existing ones in several aspects: richer event log with transition attributes, layered DFG-like (Directly-Follows Graph) process model, unique quality measure termed relevance for all model components, process model with meaningful enough components only.

The choice of multi-layer DFG process model is motivated by the gaps of existing models such as Petri nets with respect to the key requirements of our study: modelling the forward evolution of care pathways, consideration of transition attributes, explanable meaningfulness meaures of all model components. The multi-layer DFG process model and the optimal event game resolve the limitations of DFG identified in [32]: (i) the Spaghetti-like DFGs with loops, (ii) "invisible gaps", and (iii) misleading performance information. Issue (i) is naturally solved with our multi-layer DFG. Issue (iii) is not relevant as the performance information is not considered here. Issue (ii) remains but should be clear to practitioners as only meaningful enough components including arcs are given in the model. Of course, the extension to explanable meaningfulness measures of components of more complex mult-layer process models such as Petri nets is an interesting future research beyond the scope of this paper.

Compared to the traditional quality dimensions such as fitness, precision and structure, relevance is used to measure the meaningfulness of all model components. Relevance can be considered as a new non standard variant of the traditional fitness adapted to our event log and process model. It measures not only the fitness of the process model and the event log but also the meaningfulness of all model components. It makes possible the efficient optimization of event games and the process model optimization subject the minimal meaningfulness requirement. Numerical experiments presented at the end compare our approach with some existing algorithms and show the alignment of both fitness and relevance. The precision dimension is partly taken into account by the minimal meaningfulness requirement of different model components that has better explanability in practice. The structure dimension is a qualitative measure on readability of the process model by practitioners and taken into account by the simplicity of the multi-layer DGF models and the relevance measures associated with all model components.

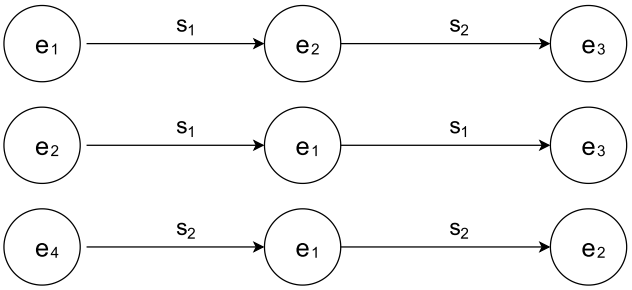

Fig. 1. An event log of three traces.

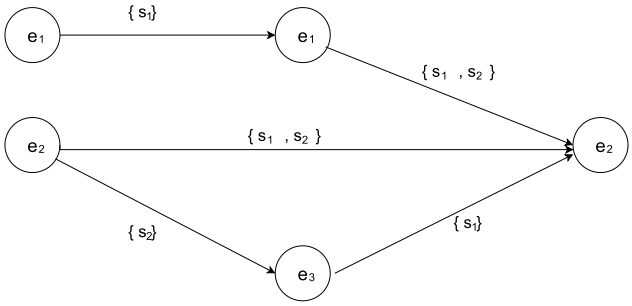

Fig. 2. A process model.

III. Problem Setting

A. Input Data Representation

This subsection provides formal definitions of the input data used for the process mining. Roughly speaking, the input data is an event log (Fig. 1) consisting of a set of traces. Each trace is an order sequence of labels called events and a transition attribute between any two consecutive events. Transition attribute is sometime called attribute for short.

Formally speaking, the event log is built upon two alphabet sets:

- E: a finite set of event labels;
- S: a finite set of transition attributes.

Definition 1 (Trace): A trace denoted t is defined by its length $m \in N$, a sequence of events $\{e_1, \ldots, e_m\}$ with $e_i \in E$ and a transition attribute $s_i \in S$ associated with any two consecutive events e_i and e_{i+1}. The notation $t = e_1(s_1)e_2(s_2)\ldots e_m$ will also be used. To each trace are associated the following notation and functions:

- $\|t\|$: the number of events in trace t, i.e. its length m;
- $\pi(t, e)$: the position of event e in trace t, i.e. $\pi(t, e_i) = i$;
- $\varepsilon(t, i)$: the i-th event of trace t, i.e. e_i. It will be called the event function;
- $\sigma(t, i, i+1)$: the attribute associated with transition (e_i, e_{i+1}), i.e. s_i. It will be called the attribute function. The attribute function will also be extended to nonconsecutive events with $\sigma(t, i, j) = \{s_i, \ldots, s_{j-1}\}$.

Definition 2 (Event Log): An event log L is a set of traces $L = \{t_1, \ldots, t_{card(L)}\}$. It consists of the input data of our process mining problem.

For the event log of Figure 1, $E = \{e_1, e_2, e_3, e_4\}$, $S = \{s_1, s_2\}$, $L = \{t_1, t_2, t_3\}$ with $t_1 = e_1(s_1)e_2(s_2)e_3$, $t_2 = e_2(s_1)e_1(s_1)e_3$, and $t_3 = e_4(s_2)e_1(s_2)e_2$.

Remark 1: In practice, each event e is associated with its occurrence time called time stamps denoted as $time(e)$. As a result, for any trace t with event sequence $\{e_1, \ldots, e_m\}$, $time(e_1) \leq time(e_2) \leq \ldots \leq time(e_m)$.

Remark 2: Our event log model differs from the existing ones by introducing the attribute between any two events. As will be proved by case studies, it allows to explicit show the impact of patient health state on the care pathways and hence provides richer information than traditional event log models.

Remark 3: As will be seen in the case study part, traditional event log models can also be used by appropriate representation of the events such as event-health state couples or event-health state-position triplets. Unfortunately, the resulting process models are significantly more complex and hence do not clearly show the impact of states on patient pathways.

B. Process Model Representation

This subsection describes the solution of our process mining problem. It is a multi-layer network model with each node associated with an event label, arcs connecting lower layer nodes to higher layer nodes, and each arc associated with a subset of attributes. Fig. 2 presents an example of a process model with 3 layers, 3 events and 2 attributes.

Definition 3 (Process Model): A process model denoted by PsM is a four-uplet $(N, A, \varepsilon, \sigma)$ where:

- $N = N_1 \cup N_2 \cup \ldots \cup N_K$ with N_k being the set of nodes of layer k and K the number of layers. The notation N_k is extended to $N_{[\![k,k']\!]}$ to indicate the set of nodes of layers k to $k' > k$;
- $A \subset N \times N$ being the set of arcs such that $(n, n') \in A$ with $n \in N_k, n' \in N_{k'}$ implies $k < k'$, i.e. arcs connecting lower layer nodes to upper layer nodes;
- $\varepsilon(PsM, n) \in E$ associates with each node n an event label such that, for all k, $\varepsilon(PsM, n) \neq \varepsilon(PsM, n'), \forall n \neq n' \in N_k$, i.e. nodes of the same layer have different event labels. ε is called the event function and $\varepsilon(PsM, N_k) \subset E$;
- $\sigma(PsM, n, n') \subset S$ and $\sigma(PsM, n, n') \neq \varnothing$ associates with each arc $(n, n') \in A$ a nonempty set of attributes and σ is called the attribute function.

Given the above, each node can be either denoted by its node ID n or by its layer and event label (k, e). Note that the same notation ε and σ is used for event (attribute) function for both the traces and the process model. It will create no confusion and allows clear link between the process model and the traces.

Remark 4: In the above definition, the set of attributes of each arc is nonempty. Further, each node is associated with a single event label and the process model is acyclic. The extensions to nodes associated with a subset of event labels and process models with cycles are interesting future research directions beyond the scope of this paper.

C. Event Game of a Trace in a Process Model

The fundamental assumption of this paper is that all traces in an event log cannot be completely and exactly captured by any process model of interest. As a result, we need to determine which events and attributes of a trace can be represented by a given process model.

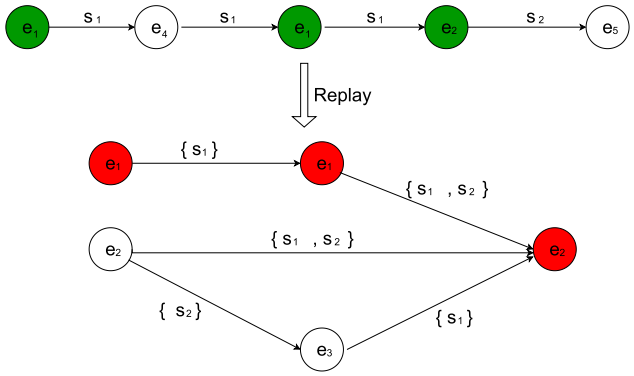

Fig. 3. Event game with footprint in green and image in red.

We introduce the concept of event game to represent how traces are represented in a given process model. Event games are subject to the following obvious constraints:

- each event can only be represented by a node of the same event label;
- events of a trace are represented by nodes in increasing order of layers.

Definition 4 (Event Game): An event game denoted by γ is a mapping from events of traces t to nodes of the process model such that, for the i-th event of t, $\gamma(t,i)$ is either undefined denoted as $\gamma(t,i) \uparrow$ or $\gamma(t,i) \in N$ and, for all well-defined mapping $\gamma(t,i)$ and $\gamma(t,j)$ such that $i < j$, $\gamma(t,j)$ belongs to higher layer than $\gamma(t,i)$.

Definition 5 (Footprint and Image): The set of event positions of a trace t represented by an event game in a process model is called its footprint and denoted as $\{[1],[2],\ldots,[\|\gamma(t)\|]\}$ where $\|\gamma(t)\|$ is the number of events represented and $[k]$ is the k-th position of trace t represented, i.e. $\gamma(t,[k]) \in N$. The image of t denoted by $IM(\gamma,t)$ is the set of corresponding nodes, i.e. $IM(\gamma,t) = \{\gamma(t,[1]),\gamma(t,[2]),\ldots\}$.

For Fig. 3, the footprint is $\{1, 3, 4\}$ and $IM = \{(1,e_1),(2,e_1),(3,e_2)\}$ implying that the 1st, 3rd and 4th events are mapped to nodes $(1,e_1),(2,e_1),(3,e_2)$. Further, $\gamma(t,2)$ and $\gamma(t,5)$ are undefined, i.e. events e_4 and e_5 are not replayed.

D. Goodness Measures of a Process Model and an Event Game

This subsection proposes goodness measures of a process model controlled by an event game that we call relevance. Both local relevance with respect to a given trace and relevance with respect to the whole event log are considered. Besides the model-wide goodness measures, we also measure the importance of each component of the model, i.e. nodes, arcs and attributes associated to arcs.

Definition 6 (Local Relevance With Respect to a Trace): For a given process model $PsM = (N,A,\varepsilon,\sigma)$, an event game γ and a trace $t = e_1(s_1)e_2(s_2)\ldots e_m$, let $\{n_1,n_2,\ldots,n_J\}$ be the image of t. The local relevance with respect to trace t is defined as follows:

- $f^{node}(\gamma,t,n)$ local node relevance of node n with $f^{node}(\gamma,t,n) = 1$ if n belongs to the image of t and 0 otherwise;
- $f^{arc}(\gamma,t,n,n')$ local arc relevance of arc (n,n') with $f^{arc}(\gamma,t,n,n') = (1-\lambda) + \lambda\frac{1}{[j+1]-[j]}\sum_{s\in\sigma(PsM,n,n')}\#(\sigma(t,[j],[j+1]),s)$ if $n = n_j$ and $n' = n_{j+1}$ for some j and 0 otherwise, where the term following λ denotes the percentage of attributes $s_{[j]},\ldots s_{[j+1]-1}$ that are associated to arc (n,n');
- $f^{attribute}(\gamma,t,n,n',s)$ local attribute relevance of an attribute s associated with an arc (n,n') with $f^{attribute}(\gamma,t,n,n',s) = \frac{\#(\sigma(t,[j],[j+1]),s)}{\sum_{s'\in\sigma(PsM,n,n')}\#(\sigma(t,[j],[j+1]),s')} f^{arc}(\gamma,t,n,n')$ if $n = n_j$ and $n' = n_{j+1}$ for some j and $s \in \sigma(PsM,n,n')$,and 0 otherwise;
- $f^{model}(\gamma,t) = \sum_{n\in N} f^{node}(\gamma,t,n) + \alpha\sum_{(n,n')\in A} f^{arc}(\gamma,t,n,n')$ local model relevance.

where $\#(\sigma(t,[j],[j+1]),s)$ denotes the number of occurrences of s in $\sigma(t,[j],[j+1])$, $\lambda \in [0,1)$ is the relative weight of attribute with respect to arc, $\alpha > 0$ is the weight of arcs with respect to nodes.

For the trace and process model of Fig. 3, $f^{node}(\gamma,t,(1,e_1)) = 1$, $f^{node}(\gamma,t,(1,e_2)) = 0$, $f^{arc}(\gamma,t,(1,e_1),(2,e_1)) = 1$, $f^{arc}(\gamma,t,(1,e_2),(2,e_3)) = 0$, $f^{attribute}(\gamma,t,(2,e_1),(3,e_1),s_1) = 1$, $f^{model}(\gamma,t) = 3+2\alpha$.

Remark 5: Whereas the node relevance is natural, some discussions are needed for arc relevance and the relevance of attributes naturally follows. Clearly only arcs (n,n') corresponding to transitions $(e_{[j]},e_{[j+1]})$ of the trace are relevant. The arc relevance also depends on the percentage of attributes $\{s_{[j]},\ldots,s_{[j+1]-1}\}$ that are associated to the arc. $f^{arc}(\gamma,t,n,n') = 1$ implies arc (n,n') corresponds to some $(e_{[j]},e_{[j+1]})$ and all attributes $\{s_{[j]},\ldots,s_{[j+1]-1}\}$ are associated to the arc. Further, if a trace is perfectly represented, then $f^{model}(\gamma,t) = m + \alpha(m-1)$.

Definition 7 (Relevance With Respect to an Event Log): For a given process model $PsM = (N,A,\varepsilon,\sigma)$, an event game γ and an event log L, the relevance is defined as follows:

- $F^{node}(\gamma,n) = \sum_{t\in L} f^{node}(\gamma,t,n)$ node relevance of node n;
- $F^{attribute}(\gamma,n,n',s) = \sum_{t\in L} f^{attribute}(\gamma,t,n,n',s)$ attribute relevance of an attribute s associated with an arc (n,n');
- $F^{arc}(\gamma,n,n') = \sum_{t\in L} f^{arc}(\gamma,t,n,n')$ arc relevance of arc (n,n');
- $F^{model}(\gamma) = \sum_{t\in L} f^{model}(\gamma,t)$ model relevance.

By definition, we also have:

$$F^{model}(\gamma) = \sum_{n\in N} F^{node}(\gamma,n) + \alpha\sum_{(n,n)\in A} F^{arc}(\gamma,n,n')$$

$$F^{model}(\gamma) = \sum_{t\in L} f^{model}(\gamma,t)$$

Remark 6: Although the relevance offers a consistent measure of both the quality of the model and the significance of model components, the rigorous extension of the traditional

fitness metric to our event log with richer information and the optimization of the exact fitness metric is an interesting future research beyond the scope of the paper.

E. Process Model Optimization Formulation

This subsection gives the formal definition of the process model optimization problem. In other words, its consists in determining a process model and an event game in order to maximize the model relevance subject to some minimal relevance constraints. Note that constraints are needed to avoid spaghetti-like messy and over-complicated model.

Formally speaking, the process model optimization problem is as follows:

$$\max_{PsM,\gamma} F^{model}(\gamma) = \sum_{n\in N} F^{node}(\gamma, n) + \alpha \sum_{(n,n)\in A} F^{arc}(\gamma, n, n') \tag{1}$$

$$\text{subject to: } PsM = (N, A, \varepsilon, \sigma) \tag{2}$$

$$N = N_1 \cup N_2 \cup \ldots \cup N_K \ with\ \varepsilon(PsM, N_k) \subset E \tag{3}$$

$$\sigma(PsM, n, n') \neq \varnothing, \forall (n, n') \in A \tag{4}$$

$$F^{node}(\gamma, n) \geq LB^{node} \tag{5}$$

$$F^{arc}(\gamma, n, n') \geq LB^{arc} \tag{6}$$

$$F^{attribute}(\gamma, n, n', s) \geq LB^{attribute} \tag{7}$$

where $LB^{node}, LB^{arc}, LB^{attribute}$ are minimal relevance for nodes, arcs and attributes in the model. Constraint (3) defines the maximum number of layers, (4) restricts to arcs of nonempty attribute sets, (5)-(7) are used to build a process model with only meaningful enough components, i.e. components that represents significant volume of information of the event log.

Remark 7: The event game γ defines how each trace is replayed in the model. Finding the optimal event game is an optimization problem. References [3] and [28] proposed a fixed event game to define how traces are replayed. To the best of our knowledge, the optimal event game has never been studied in the literature. In this paper, a dynamic programming algorithm is proposed to address the problem and will be illustrated in Section IV.

Remark 8: The model precision level is controlled via parameters LB^{node}, LB^{arc}, $LB^{attribute}$ and K. A more (less) complicated model can accommodate more (less) information. This will be demonstrated in Section VII-B.

Theorem 1: There exists a feasible process model with at least one node if and only if $LB^{node} \leq U$ where $U \equiv \max_{e\in E} u_e$ where u_e is the total number of traces containing event e. Further $F^{node}(\gamma, n) \leq U, \forall n \in N$.

Proof: We first prove $F^{node}(\gamma, n) \leq U, \forall n \in N$. Let e be the event label of n. By definition, $f^{node}(\gamma, t, n) = 0$ for all trace t not containing e and hence $F^{node}(\gamma, n) \leq u_e \leq U, \forall n \in N$. As a result, there is no feasible process model with at least one node if $LB^{node} > U$. Otherwise, the process model with a single node of event e^* with $e^* \equiv \operatorname{argmax}_{e\in E} u_e$ is a feasible solution. Q.E.D. □

To avoid trivial cases, the assumption $LB^{node} \leq U$ is assumed throughout the paper.

IV. Optimal Event Game of a Trace in a Given Model

This section addresses the problem of optimal event game of a given trace in a given model and proposes a dynamic programming method. More specifically, consider a given process model $PsM = (N, A, \varepsilon, \sigma)$ and a trace $t = e_1(s_1)e_2(s_2)\ldots e_m$. The problem of optimal event game consists in determining an event game γ in order to maximize the local model relevance, i.e.

$$f^{model}(\gamma^*, t) = \max_{\gamma} f^{model}(\gamma, t)$$

Let us first define useful notation:

- $N^-(n) = N_{[\![1,k-1]\!]}, \forall n \in N_k$: set of lower layer nodes of node n;
- $N(e_i) = \{n \in N : \varepsilon(PsM, n) = e_i\}$: set of nodes that can represent e_i;
- $N(t) = \bigcup_{i=1}^{m} N(e_i)$: set of nodes that can represent an event of t;
- $l(n, n', i, j)$: local arc relevance of any node couple (n, n') representing events e_i and e_j with $j > i$ defined in (8), as shown at the bottom of the next page.

The dynamic programming method relies on the following optimal event game for all partial traces:

- $g_i(n)$: the optimal model relevance of trace $e_1(s_1)e_2(s_2)\ldots e_i$ with event e_i represented by node n.

By definition, $g_i(n) = 0$ for all $n \notin N(e_i)$ and hence we focus on nodes $n \in N(e_i)$.

The partial solutions $g_i(n)$ can be determined recursively from $i = 1$ to $i = m$ as defined in (9), as shown at the bottom of the next page, for all $n \in N(e_i)$ where $PRE(i, n) = \{(i', n') : i' \in [\![1, i-1]\!], n' \in N(e_{i'}) \cap N^-(n)\}$ denotes the set of footprints and images that could be predecessor of footprint and image (i, n). The optimal image and footprint are determined by equation (10), as shown at the bottom of the next page, where $event_i^-(n)$ and $node_i^-(n)$ denote the previous event represented and its image. By convention, $(event_i^-(n), node_i^-(n)) = (0, 0)$ implies that no preceding event is represented.

The optimal event game of trace t can then be determined by the following:

$$f^{model}(\gamma^*, t) = \begin{cases} 0 & \text{if } N(t) = \varnothing \\ \max\limits_{t\in[\![1,m]\!], n\in N(e_i)} g_i(n) & \text{otherwise} \end{cases}$$

V. Proposed Algorithm for Process Model Optimization

This section addresses the process model optimization problem, i.e. determines a process model and its event game that maximizes the model relevance subject to minimal relevance constraints of nodes, arcs and attributes. The problem is extremely complex and for this reason, this section proposes a multi-start local optimization approach. The key of

Algorithm 1 Optimal Event Game

Input: a process model $PsM = (N, A, \varepsilon, \sigma)$ and a trace $t = e_1(s_1)e_2(s_2)\dots e_m$;
Output: event game that maximizes the local model relevance;
Step 1. Determine sets $N(e_i)$ and $N(t)$ of nodes that are relevant to t;
Step 2. If $N(t) = \varnothing$, then $f^{model}(\gamma^*, t) = 0$ and no event is represented;
Step 3. For $i = 1$ to m, determine the partial optimum event game $(g_i(n), event_i^-(n), node_i^-(n)), \forall n \in N(e_i)$
Step 4. Determine the optimal event game
4.1 Determine $(i^*, n^*) = \underset{i\in[\![1,m]\!],n\in N(e_i)}{\operatorname{argmax}} g_i(n)$;
4.2 Set $f^{model}(\gamma^*, t) = g_{i^*}(n^*)$, add i^* and n^* to lists footprint and image;
4.3 While $event_{i^*}^- \neq 0$ {
$\bullet (t^*, n^*) \leftarrow (event_{i^*}^-, node_{i^*}^-)$;
• Add i^* and n^* to the head of Footprint and Image;
} endwhile.

local search is the marginal model relevance of adding a new node. All neighbor solutions are repaired for feasibility. Local optimal event game of Section IV is used for process model evaluation. Different components of the proposed algorithm and the overall algorithm are presented hereafter. This section ends with the presentation of different benchmarking algorithms.

A. Marginal Model Relevance of a New Node

This subsection considers the fundamental problem of the benefit evaluation of adding a new node n' to a given process model PsM. A natural measure of this benefit is the following marginal relevance increase:

$$F^{model}(PsM', \gamma') - F^{model}(PsM, \gamma)$$

where the dependence of the model relevance to the process model is introduced explicitly, PsM' is the optimal feasible model obtained by adding n' to PsM, γ and γ' are optimal event game of PsM and PsM' respectively. Whereas this marginal relevance seems attractive, it requires significant computational effort and is not suited for any local search optimization algorithms.

Instead, we define the benefit measure of adding a node of event e to layer k $\Delta(PsM, \gamma, e, k)$ as the number of new events that can be represented without alternating the existing event game γ. It is determined by means of $\delta(\gamma, t, e, k)$ a binary number equal to 1 if a new event of trace t can be represented by node (e, k). More specifically, $\delta(\gamma, t, e, k)$ equals to 1 if there exists an event $e_i = e$ of t that is not represented, all preceding events are either not represented or have images in lower layers, all following events are either not represented or have images in higher layers. As a result,

$$\Delta(PsM, \gamma, e, k) = \sum_{t\in L} \delta(\gamma, t, e, k)$$

Algorithm 2 Marginal Relevance of a New Node

Input: a process model $PsM = (N, A, \varepsilon, \sigma)$ and an event game γ;
Output: Benefit measure matrix $\Delta(PsM, \gamma, e, k)$;
Step 1. Initialization: $\Delta(PsM, \gamma, e, k) \leftarrow 0$;
Step 2. For all $t \in L$ and for all $j = 0$ to $[\|\gamma(t)\|]$,
$\Delta(PsM, \gamma, e, k) \leftarrow \Delta(PsM, \gamma, e, k) + 1$,
$\forall e \in \{\varepsilon(t, [j] + 1), \dots, \varepsilon(t, [j + 1] - 1)\}$ and $\forall k \in [\![layer(\gamma(t, [j])) + 1, layer(\gamma(t, [j + 1])) - 1]\!]$ where $[1], \dots, [\|\gamma(t)\|]$ is the footprint of t and $\{\gamma(t, [1]), \dots\}$ its image. By convention $[0] = 0, [\|\gamma(t) + 1\|] = m + 1$, $\gamma(t, [j])$ belongs to layer 0 and $\gamma(t, m + 1)$ layer $K + 1$, $layer(n)$ denotes the layer of n.

Remark 9: $\Delta(PsM, \gamma, e, k) + F^{model}(PsM, \gamma)$ is the model relevance of process model PsM' derived from PsM by adding node (e, k) without adding any arc and without alternating the images of traces on nodes of PsM. Of course, PsM' derived this ways needs not to be feasible.

B. Solution Repair

This subsection addresses the repair of an infeasible process model PsM, i.e. with the violation of at least one minimal relevance constraint of model components. The basic idea is to derive a feasible process model PsM' by removing infeasible attributes, arcs and nodes with corresponding relevance less than $LB^{attribute}$, LB^{arc} and LB^{node} respectively. To account

$$l(n, n', i, j) = \begin{cases} 0 & if (n, n') \notin A \\ 1 - \lambda + \lambda \dfrac{1}{j - i} \displaystyle\sum_{i'=i}^{j-1} 1(s_{i'} \in \sigma(PsM, n, n')) & otherwise \end{cases} \tag{8}$$

$$g_i(n) = \begin{cases} 1, & if\ PRE(i, n) = \varnothing \\ \max\limits_{(i',n')\in PRE(i,n)} 1 + \alpha l(n', n, i', i) + g_{i'}(n'), & otherwise \end{cases} \tag{9}$$

$$(event_i^-(n), node_i^-(n)) = \begin{cases} (0, 0), & \text{if } PRE(i, n) = \varnothing \\ \operatorname{argmax}\limits_{(i',n')\in PRE(i,n)} 1 + \alpha l(n', n, i', i) + g_{i'}(n'), & \text{otherwise} \end{cases} \tag{10}$$

for the change of event game, the infeasible model components are removed one at a time in the order of attributes, arcs and then nodes and in the non-decreasing order of relevance for each type of model components.

Algorithm 3 Process Model Repair

Input: a given process model PsM;
Output: a feasible process model PsM';
Step 1. Set $PsM' \leftarrow PsM$ and evaluate PsM' with algorithm 1.
Step 2. While PsM' is infeasible do
2.1 Determine the minimal relevance attribute (n_1^s, n_2^s, s^s), the minimal relevance arc (n_1^a, n_2^a), and the minimal relevance node n^n;
2.2 Modify PsM' according to the following three cases:
• if (n_1^s, n_2^s, s^s) is infeasible, then remove the attribute and its arc if it is the only attribute;
• else if (n_1^a, n_2^a) is infeasible, then remove the arc;
• else, then remove the node n^n.
2.3 Evaluate PsM' with algorithm 1;
endwhile.

Note that the above algorithm terminates as the empty process model without any node is a feasible solution. Further, preliminary numerical experiments show that the above progressive repair gives better results than removing all infeasible components at a time. This is due to the fact that removing some components might make feasible other infeasible ones.

C. Initial Solution

This subsection proposes a random generation of feasible solutions. It starts by random generation of event labels of different layers, then derives a process model by connecting all nodes of different layers and assigning to each arc the complete set of attributes, and then repairs the solution by algorithm 3. More specifically,

Algorithm 4 Random Feasible Process Model Generation

Step 1. Generate the set of nodes: add node (e, k) with probability 0.5 for all event e and layer k;
Step 2. Generate the set of arcs: add an arc (n, n') for all nodes $n = (e, k)$, $n' = (e', k')$ and $k < k'$;
Step 3. Generate the attributes: $\sigma(PsM, n, n') = S$ for all arcs (n, n');
Step 4. Repair the process model PsM by algorithm 3;
Step 5. Repeat steps 1-4 till a non-empty process model is obtained.

Note that, by condition of Theorem 1, the above algorithm terminates.

D. The Proposed Algorithm

This subsection proposes a multi-start local optimization heuristic. It starts with a randomly generated initial solution PsM, improves PsM by adding nodes of positive marginal relevance and solution repair, and restart when the current solution cannot be improved.

We first define the node insertion operator. Let $Insert(PsM, (e, k))$ be a new process model derived from PsM by adding a new node (e, k), by connecting all lower layer nodes to (e, k) and connecting (e, k) to all higher layer nodes, and by associating the complete set of attributes to all new arcs.

We are now ready to present rigorously the proposed heuristic.

Algorithm 5 Our Process Model Optimization Algorithm

Step 1. Initialization: Let PsM^{best} be an empty process model;
Step 2. Random generation of a non-empty process model PsM by algorithm 4;
Step 3. Determine the marginal relevance matrix $\Delta(PsM, \gamma, e, k)$ by algorithm 2. Let Candidates be the list of all nodes (e, k) such that $\Delta(PsM, \gamma, e, k) > 0$ and sorted in descending order;
Step 4. While (Candidates not empty) do
4.1 $PsM' = Insert(PsM, (e^*, k^*))$ where (e^*, k^*) is the head of Candidates;
4.2 Repair PsM' by algorithm 3;
4.3 Evaluate PsM' by algorithm 1;
4.4 If PsM' better than PsM^{best}, $PsM^{best} \leftarrow PsM'$;
4.5 If PsM' better than PsM, $PsM \leftarrow PsM'$ and go to Step 3;
4.6 Otherwise, remove (e^*, k^*) from Candidates;
Step 5. Repeat Steps 2-4 either the maximal computation time $time_{max}$ or the maximal number $iter_{max}$ of iterations without improvement is reached.

E. Benchmark Algorithms

This subsection describes several benchmark algorithms with which the proposed algorithm will be compared. Our proposed algorithm is compared against the general process mining software Disco and the following more specialized heuristics:

- *Random* algorithm: it randomly generates a large number of initial solutions by Algorithm 4 and chooses the best. The number of solutions generated depends on the maximal allowable computational time.
- *RG* algorithm: It is an iterative Random Growth algorithm starting from an empty process model. At each iteration, it randomly selects a new node, adds the new node to the current process model, connects the new node to/from all existing nodes by arcs associated with all attributes, and then repairs the resulting process mode. The resulting model is set as the current model if it is better.
- *Reinsert* algorithm: It is a multi-start local optimization algorithm. It starts from an initial solution generated by Algorithm 4. At each iteration, it determines the node (e, k) of the lowest relevance, moves it to another layer (e, k'). Each local move from (e, k) to (e, k') is evaluated by the model relevance of the complete process model with an arc connecting any two nodes and the arc

associated with all attributes (feasible or not). The local move with the highest model relevance is selected and the corresponding complete model is repaired by Algorithm 3. The resulting feasible model is set as the current model if it is better than the current solution. Otherwise, the algorithm restarts from another new initial solution.

- *Relabeling* algorithm: it is similar to the *Reinsert* algorithm but with local move defined by the relabeling the least relevant node (e, k) as (e', k), i.e. replacing the current event label e by e'.

VI. Numerical Results

This section compares the proposed algorithm with benchmark algorithms of Section V on the basis of generated test instances. We first discuss the test instance generation in Section VI-A. Section VI-B addresses the algorithm parameter setting. Section VI-C compares the performances of various process model optimization algorithms.

For all test instances as well as the case studies of Section VII, the goodness of a process model is defined with the following weights: weight of arc relevance $\alpha = 1$ and weight of attribute relevance $\lambda = 0.5$. For each instance, the following gap $gap(PsM) = 1 - F^{model}(PsM)/F^{model}(PsM^*)$ or $gap(PsM) = 1 - F^{model}(PsM)/F^{model}(PsM^{best})$ is determined depending on whether the optimal process model is available where PsM^* is the optimal process model and PsM^{best} is the best process model among all algorithms. Note that the dependence on the event game is neglected here.

The programming language is C++, and the tests are run on an Intel Xeon E5-2660 v3 CPU, 2.60 GHz processor with 64 GB of memory. The proposed algorithm as well as the test instances of this section is accessible at https://github.com/zhihao007/Tase_github.

A. Test Instance Generation

This subsection proposes a generation of test instances for which the optimal process model is known. It starts from a base directed graph which is equivalent to a process model without event labels and attributes. It then derives various process models by assigning event labels and attribute sets to nodes and arcs. For each process model, it then generates event logs composed of traces that can be perfectly represented. Details of the instance generation is given below.

The base directed graph is an acyclic graph (N, A) with nodes grouped into K layers and arcs connecting lower layer nodes to higher layer nodes. Fig. 4 gives the base directed graph used in this paper with $K = 15$ layers, $|N| = 19$ nodes and $|A| = 36$ arcs. It is derived from our real case studies in order to generate test instance of realistic complexity. Note that the algorithms have also been tested on test instances derived from base directed graphs of various sizes and the results are similar and omitted.

For each given base directed graph (N, A), the generation of the process model PsM depends on the size $|E|$ of the event set and the size of the attribute set $|S|$. The event label $\varepsilon(PsM, n)$ of each node n is random selected from the event set E with equal probability $1/|E|$. For each arc $(PsM, n, n') \in A$, the set $\sigma(PsM, n, n')$ of its attributes is determined by random inclusion of each attribute $s \in S$ with probability 0.5. The random inclusion is repeated till $\sigma(PsM, n, n')$ becomes non-empty.

For each given process model PsM, event traces are generated as follows. We first randomly select the starting node n in PsM by sequentially testing the nodes from layer 1 to layer K. A node n of layer k is selected as the starting node with probability $p_0 q^k$ ($p_0 = 0.9$ and $q = 0.3$ in this paper) and otherwise, the next node is tested. If starting node n has no successor, then the generation stops. If starting node n has at least one successor n', the next node n' is selected randomly with equal probability among all successors, an attribute is also selected with equal probability among the set $\sigma(PsM, n, n')$. The new node n' is the terminating node with probability $p_0 q^{K-k}$. The generation stops if n' is terminating and otherwise, the generation continues from node n'. The event trace is defined by the sequence of events and attributes from the starting node to the terminating one.

For each given process model PsM and a generated event log, the PsM is modified by removing nodes-arcs-attributes that are not traversed by any trace and the process model optimization parameters LB^{node}, LB^{arc}, and $LB^{attribute}$ are defined as follows. Let $(U^{node}(n)$ $U^{arc}(n, n')$, $U^{attribute}(n, n', s))$ be the number of traces passing node n, $arc(n, n')$, $attribute(n, n', s)$ in the process model PsM. In the test instance,

$$LB^{node} = \min_{n \in N} U^{node}(n)$$
$$LB^{arc} = \min_{(n,n') \in A} U^{arc}(n, n')$$
$$LB^{attribute} = \min_{(n,n') \in A, s \in \sigma(PsM,n,n')} U^{attribute}(n, n', s)$$

By construction, PsM is an optimal process model and the event game defined by the event log generation process is an optimal event game. Further,

$$F^{model}(PsM^*) = \sum_{t \in L} (||t|| + \alpha(||t|| - 1))$$

In this paper, starting from the base directed graph of Fig. 4, 40 test instances are generated as follows. Four configurations with $(|E|, |S|) = (10, 2), (10, 3), (12, 2), (12, 3)$ are used. For each configuration, 10 event logs with 1000 traces for each are generated leading to 10 test instances.

B. Algorithm Parameter Setting

The algorithms of this paper have two most important parameters: the maximum computation time $time_{max}$ and the maximum number of iterations without improvement $iter_{imp}$. The selection of the parameters is more subtle and we resort to numerical experiments for this purpose.

Fig. 5 gives the numerical results for the proposed algorithm on four instances (one for each configuration) with $iter_{imp} \in$ 50, 100, 150, 200, 250 and 10 algorithm runs for each combination of instance and $iter_{imp}$. From these results, $iter_{imp} =$ 150, 200, 250 outperforms $iter_{imp} = 50, 100$ and the performances with $iter_{imp} = 150, 200, 250$ are quite close in terms of average gap and variance. For these reasons,

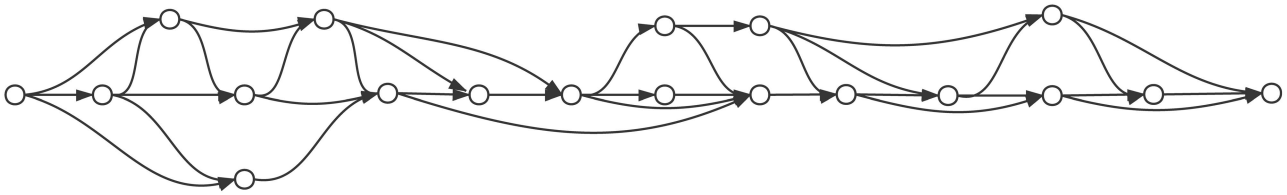

Fig. 4. Base directed graph.

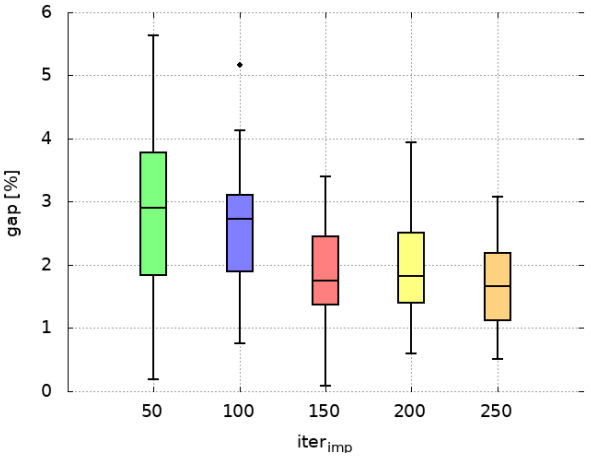

Fig. 5. Impact of the algorithm parameter $iter_{imp}$.

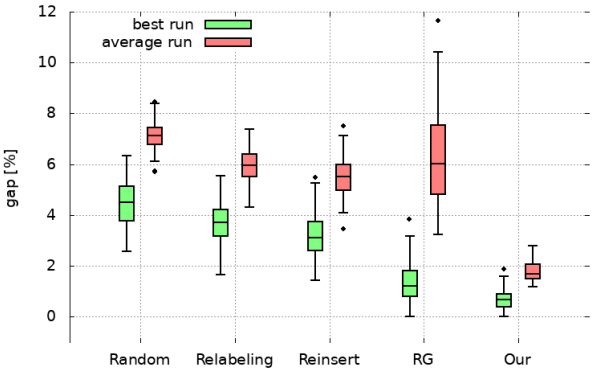

Fig. 6. Best gap and average gap on test instances.

$iter_{imp} = 200$ is used in the remaining of the paper apart for the comparison between different algorithms (section VI-C).

We set $time_{max} = 30$ minutes which is from our point of view a reasonable time limit. Furthermore, the average computation time (23.9 minutes) of all the runs does not exceed this value.

C. Algorithm Performance Assessment

This subsection compares the proposed algorithm denoted "Our" and the benchmark algorithms "Random", "Reinsert", "Relabeling", and "RG". For each of the 40 test instances, 10 runs are performed for each algorithm and the best of the 10 runs and the average are determined.

The best gap (gap of the best run) and the average gap of the 10 runs of the five algorithms are summarized in Fig. 6 and details are given in Table I and Table II.

The following observations can be made:

- "Random" also the worst performed algorithm still provides reasonable performance with mean best gap of 4.46%, mean average gap of 7.14% and worst gap of 11.6% across all instances and all runs. It provides other algorithms with good starting solution. Note however that the number of initial solutions tested in other algorithms is significantly smaller than the number of random solutions generated by "Random" in 30 minutes;
- The algorithms can be ranked as follows: (Our 0.70%, RG 1.38%, Reinsert 3.17%, Relabeling 3.63%, Random 4.46%) by mean best gap and (Our 1.80%, Reinsert 5.51%, Relabeling 5.93%, RG 6.33%, Random 7.14%) by mean average gap. Our proposed algorithm consistently outperforms competing algorithms and is quite close to the real optimum. Random performs the worst as expected;
- The comparison of "Our" with "Reinsert" and "Relabeling" suggests that the neighborhood structure matters.

TABLE I
RESULTS ON THE GAPS OF THE BEST RUN

Instance		Best gap (%)				
$(\|E\|,\|S\|)$	#	*RG*	*Rand*	*Reinsert*	*Relabel*	*Our*
(10,2)	0	0.88	3.73	3.32	4.28	0.83
	1	0.41	2.78	3.26	1.66	0.44
	2	1.08	4.59	4.17	2.09	0.77
	3	1.77	4.18	3.23	1.70	0.21
	4	1.40	5.17	2.93	4.05	1.59
	5	1.83	4.97	5.25	3.67	0.99
	6	0.86	5.00	4.02	2.53	0.70
	7	1.23	5.06	4.01	4.63	0.66
	8	1.52	2.97	3.19	3.59	0.47
	9	2.44	4.66	2.81	3.56	0.68
(10,3)	0	1.40	4.39	2.95	4.91	0.77
	1	2.08	4.43	4.41	3.37	0.52
	2	1.10	5.36	2.53	3.02	0.37
	3	0.36	4.75	3.17	3.36	0.36
	4	1.44	3.10	5.47	4.79	0.29
	5	1.22	5.11	2.77	3.93	0.63
	6	0.72	5.68	5.05	3.83	1.53
	7	1.08	2.58	1.82	2.62	0.68
	8	0.49	3.84	3.06	4.58	0.90
	9	1.04	5.40	2.77	3.92	0.91
(12,2)	0	1.18	3.89	2.75	3.76	1.00
	1	0.78	6.34	2.49	3.32	0.66
	2	0.81	4.36	3.76	3.34	1.37
	3	0.00	3.74	1.42	2.07	0.41
	4	2.47	2.76	2.02	4.75	0.00
	5	2.14	3.54	2.64	3.84	0.31
	6	1.41	3.83	3.02	4.20	1.12
	7	1.29	5.39	2.08	3.78	0.68
	8	1.75	4.84	3.50	2.74	1.26
	9	3.01	5.70	3.61	3.91	0.73
(12,3)	0	0.31	5.72	1.43	3.55	0.00
	1	2.02	4.11	2.59	5.55	0.58
	2	2.44	5.06	1.48	4.24	0.18
	3	0.79	3.08	4.09	3.68	0.96
	4	0.62	4.79	3.74	2.31	0.37
	5	3.84	6.12	3.30	4.49	0.23
	6	1.22	5.49	4.42	4.97	1.87
	7	1.26	3.64	2.25	3.02	0.82
	8	3.19	4.27	3.28	3.73	0.51
	9	0.48	3.94	2.63	3.66	0.80
Average		1.38	4.46	3.17	3.63	0.70

TABLE II
RESULTS ON THE GAPS OF THE AVERAGE RUN

Instance		Average gap (%)				
$(\|E\|,\|S\|)$	#	*RG*	*Rand*	*Reinsert*	*Relabel*	*Our*
(10,2)	0	8.60	7.49	6.08	6.17	1.84
	1	4.02	7.33	5.77	6.11	1.57
	2	3.24	6.42	6.31	5.81	1.40
	3	5.58	7.40	5.64	6.07	1.75
	4	5.24	7.13	5.37	6.02	2.38
	5	4.86	7.13	6.52	6.59	2.55
	6	4.89	6.32	5.54	5.14	1.38
	7	7.47	6.87	5.71	6.47	1.66
	8	4.93	6.12	4.84	6.41	1.17
	9	6.55	7.42	5.11	5.47	1.51
(10,3)	0	5.86	7.14	5.28	6.39	2.08
	1	6.48	7.22	5.43	5.94	1.69
	2	6.19	7.13	5.76	5.41	1.58
	3	4.04	5.73	4.10	4.56	1.63
	4	6.58	8.18	7.51	6.52	2.03
	5	3.61	6.85	6.06	5.98	1.65
	6	5.33	7.72	7.13	6.78	2.81
	7	6.20	5.74	3.46	4.32	1.52
	8	3.87	7.08	4.55	5.85	2.05
	9	4.96	7.33	4.61	5.66	1.69
(12,2)	0	9.41	6.86	4.89	5.57	1.58
	1	7.57	8.47	4.96	6.03	1.34
	2	5.85	7.39	5.44	5.51	2.07
	3	4.55	5.69	4.50	4.78	1.37
	4	9.21	6.84	5.86	6.93	1.40
	5	8.62	6.72	5.32	5.48	1.82
	6	4.65	6.43	5.48	6.52	1.50
	7	7.81	7.14	4.83	5.90	1.24
	8	7.48	6.85	5.18	5.14	2.39
	9	10.41	7.64	6.21	5.52	2.22
(12,3)	0	11.66	8.28	4.22	5.97	1.85
	1	3.45	8.27	5.90	7.39	1.90
	2	5.65	7.27	6.21	6.41	1.75
	3	4.75	8.39	5.92	5.69	1.93
	4	7.24	6.57	6.01	5.00	2.13
	5	6.21	8.03	5.52	6.86	2.39
	6	7.99	7.87	6.23	7.11	2.40
	7	6.22	7.09	5.48	5.79	1.51
	8	9.69	7.07	5.97	6.12	1.41
	9	4.37	6.68	5.02	5.84	1.66
Average		6.33	7.14	5.51	5.93	1.80

Both "Reinsert" and "Relabeling" do not change the number of nodes during the local optimization phase. Another reason of the superiority of "Our" over "Reinsert" and "Relabeling" is the ability of the marginal relevance measure to properly guide the search process;

- RG algorithm ranks second by mean best gap and even slightly outperforms "Our" for 2/40 instances. However it ranks poorly on mean average gap and hence exhibits poor robustness. The good best gap ranking and poor average gap ranking seem suggest a potential improvement of the RG algorithm by introducing restart;
- Detailed analysis of the computational time distribution shows that solution repair takes a significant part of the overall computation time in all algorithms. This is especially true for "Reinsert" and "Relabeling" which require the repair of a complete process model derived from a set of nodes. More efficient repair and appropriate repair strategy to avoid repair of all solutions are potential improvements of the algorithms.

VII. APPLICATION TO THE CARE PATHWAYS OF CANCER PATIENTS

This section applies the best performing algorithm "Our" to study the care pathways of sarcoma patients. We first present the background and the relevant data in Section VII-A. Section VII-B is a sensitivity analysis of the process model with respect to model precision level parameters. Section VII-C compares our data representation with alternative ones. Section VII-D compares the care pathways of various strategies for managing sarcoma patients. Section VII-E compares our approach with the generic process mining software Disco. Note that only the proposed algorithm ("Our" algorithm) is considered.

A. Backgroud and Data Description

According to World Health Organisation (WHO), nearly 10 million deaths worldwide were caused by cancer in 2020. The top three most common ones are breast, lung, colon and rectum. Classified as rare cancers, sarcomas describe a group of connective tissue cancers with heterogeneous histological subtypes. The overall estimated incidence is around 6.2/100,000/year according to the study of [33] conducted in a European region (Rhone-Alpes) of six million inhabitants. A complete overview of incidence and survival rates is reported in [34]. Due to its rarity, Sarcomas require a complex and specialized multidisciplinary management. In order to meet this need, the French National Cancer Institute (INCa) in collaboration with General Directorate of Healthcare Services (DGOS) have funded since 2009 several reference networks dedicated to sarcomas. NetSarc was a clinical network dedicated to patients of soft tissue and viscera sarcomas, while RRePS was created as its complementary network to guarantee the pathological review. ReSoS was funded for bone sarcomas patients by integrating both clinical and pathological functionalities. In 2019, the three networks were grouped into one network named NetSarc+, and share the same database [35]. The data used in this case study were from NetSarc+. Interested readers are invited to have more information at https://netsarc.sarcomabcb.org.

For each sarcoma patient, NetSarc+ records the following healthcare data: the sequence of care activities correspond to events with information on when-where-by whom for each. It also records the change of sarcoma state including metastatic progression and local progression. We model the later by transition attributes also called health states: no progress before any change of metastatic/local progress and progress after such event. The complete list of events and transition attributes used in the paper are summarized in Table III.

In the management of sarcoma patients, four different strategies exist:

- Strategy one: Patients who had a rcp before the initial surgery and complete initial management in the network (also including patients who had rcp after the initial surgery and complete initial management in the network);
- Strategy two: Patients who had a rcp before the initial surgery and initial management outside the network;
- Strategy three: Patients who had a rcp after initial surgery and initial management outside the network;
- Strategy four: Patients who had an initial management outside the network without rcp neither before nor after the initial surgery.

TABLE III
EVENTS AND TRANSITION ATTRIBUTES OF SARCOMA PATIENT CARE PATHWAYS

Events	
od	original diagnosis of Sarcoma
rcp	sarcoma multidisciplinary tumor boards
chir	surgery
ttt	treatment
last	last contact
bio	biopsy
neo	neo-adjuvant
Attributes	
np	no progression
pro	(metastatic or local) progression

There are in total 2203 sarcoma patients in our case study with 1069 patients of strategy 1, 143 patients of strategy 2, 720 patients of strategy 3 and 277 patients of strategy 4.

This case study allows to valid that the new process model is capable to represent the healthcare events along with the health state of patients (local/metastatic progression or not).

B. Process Model and Model Precision Level

This subsection addresses the impact of model precision level defined by the number of layer K and minimum relevance of nodes, arcs and attributes, i.e. LB^{node}, LB^{arc}, and $LB^{attribute}$. We limit ourselves to the 1068 patients of strategy 1.

We first consider the impact of these precision level parameters on the model relevance. The following combinations are considered: $K \in \{8, 10, 12, 14, 16\}$ and $(LB^{node}, LB^{arc}, LB^{attribute}) = \omega(1, 0.5, 0.25)|L|$ with $\omega \in \{10\%, 20\%, 30\%, 40\%, 50\%\}$ and $|L| = 1068$ traces. Results are given in Fig. 7. In this Figure, the impacts of K and of minimum relevance LB are evaluated by the percentage model relevance deviation with respect to that of $K = 8$ and $\omega \in 50\%$. The following observations are made:

- The model relevance increases as the model precision level increases, i.e. as K increases and $(LB^{node}, LB^{arc}, LB^{attribute})$ decreases. For a given $(LB^{node}, LB^{arc}, LB^{attribute})$, the model relevance increases till some point K_0 beyond which the change of K does not have any impact. Further, for a high $(LB^{node}, LB^{arc}, LB^{attribute})(\omega = 40\%\ or\ 50\%$ in this case), the process model has few nodes and hence the model relevance is insensitive to the change of K.

Note that the slight decrease of the model relevance at $K = 16$ for $\omega = 30\%$ and 40% is mainly due to the limited computation time of the algorithm.

We now consider the process models obtained at different precision level. Fig. 8 gives the process models obtained with $K = 12$ and $\omega = 30\%$ and 50%, i.e. $(LB^{node}, LB^{arc}, LB^{attribute}) = (30\%, 15\%, 7.5)|L|$ and $(LB^{node}, LB^{arc}, LB^{attribute}) = (50\%, 25\%, 12.5\%)|L|$. The following observation is made:

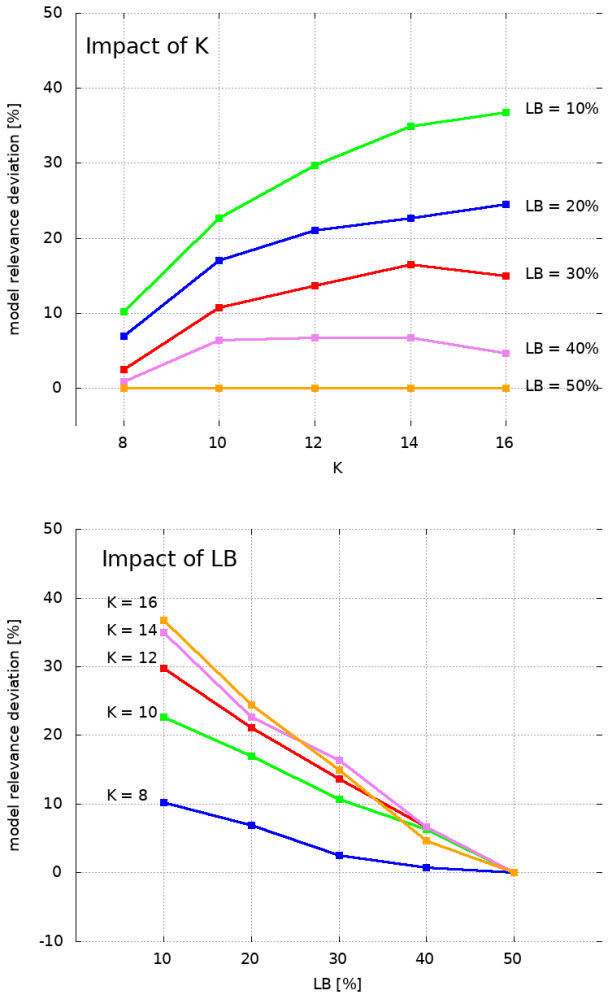

Fig. 7. Model relevance vs model precision parameters K and $(LB^{node}, LB^{arc}, LB^{attribute})$.

- As the precision level decreases (LB increases), the process model is aggregated into the most common pathways. As the precision level increases, the structure of aggregated pathway is kept with some part split into more detailed pathway model.
- From a practical point of view, the process models clear show that metastatic /local progress intervenes at the end of the care pathways and is followed by extra rcp for better monitoring.

C. Impact of Care Pathway Representations

This subsection compares the care pathway representation proposed in this paper against alternative ones on the basis of model relevance and the ability to highlight the impact of health state on care pathways.

We introduce the representations by example. Consider the care pathway $od - rcp - chir - ttt - rcp - * - rcp - last$ with a local progress observed after the second rcp identified by $*$ above. Four representations are considered:

- **Our**: $od \xrightarrow{np} rcp \xrightarrow{np} chir \xrightarrow{np} ttt \xrightarrow{np} rcp \xrightarrow{np} rcp \xrightarrow{np} last$
- **Fictitious event**: $od \longrightarrow rcp \longrightarrow chir \longrightarrow ttt \longrightarrow rcp \longrightarrow$ pro $\longrightarrow rcp \longrightarrow last$

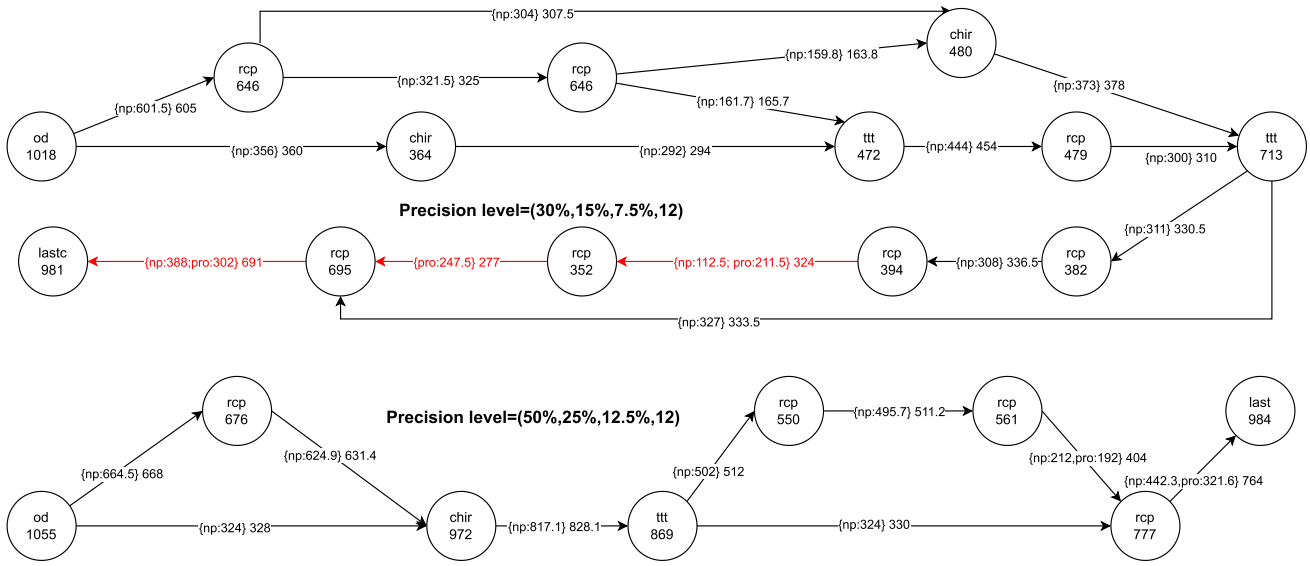

Fig. 8. Process models for strategy one.

- **Expanded event label**: $(od, np) \longrightarrow (rcp, np) \longrightarrow (chir, np) \longrightarrow (ttt, np) \longrightarrow (rcp, np) \longrightarrow (rcp, pro) \longrightarrow (last)$ where the "last" event is not expanded.
- **No state**: $od \longrightarrow rcp \longrightarrow chir \longrightarrow ttt \longrightarrow rcp \longrightarrow rcp \longrightarrow last$

Fig.9 gives the process models of the four representation of strategy 1 patients with precision level $K = 12$ and $(LB^{node}, LB^{arc}, LB^{attribute}) = (30\%, 15\%, 7.5\%)|L|$. The following observations can be made:

- The model relevance is ranked as follows: "Fictitious event" 13422, "No state" 13063, "Our" 12985.3, "Expanded event label" 11598. The apparent better model relevance of "Fictitious event" vs "No state" is basically due to the extra event "pro" in the traces;
- The similar model relevance of "No state" and "Our" shows the ability of our representation to model the impact of the health state in care pathways with minor degradation of the model relevance. The model "Our" clearly shows that the health state intervenes at the later stage of the care pathways;
- The "Expanded event label" representation not only degrades significantly the model relevance but also captures less information on the impact of the health state than "Our" representation;
- The "Fictitious event" representation has difficulty to capture the impact of the health state with enough relevance. The fictitious event "pro" does not appear in its process model. It is expected that lower LB^{node} (i.e. higher precision) would generate process model with nodes of event label "pro" but at the cost of large process model.

To summarize, the representation proposed in this paper allows model the impact of health state with minor degradation of the model relevance.

D. Care Pathways of the Four Strategies

This subsection compares the care pathways of the four strategies. Recall that patients of strategy 1 have their surgery in the reference network with rcp either before or after the surgery, patients of strategy 2, 3 have their surgery outside the network with rcp before (after) the surgery, patients of strategy 4 have their surgery outside without rcp.

Fig. 10 gives their process models at the precision level of $K = 12$ and $(LB^{node}, LB^{arc}, LB^{attribute}) = (30\%, 15\%, 7.5)|L|$.

- The process model of strategy 4 has only one node of event "od" and no other information as the patients are treated outside the network;
- Compared with the recommended strategy 1, disease progression (highlighted in orange colour) intervenes much earlier in strategies 2 and 3. We conjecture that this can be partly attributed to the better treatment patterns of strategy-1 patients by the sarcoma experts in the network, which is in line with the literature [36], [37] and to the variation of diagnostic intervals, especially possible longer diagnostic intervals when patients undergo several magnetic resonance imaging before treatments in strategies 2-3 [38];
- As a result, the care pathways of patient with disease progression tend to be longer and more complex, which is also consistent with the medical literature [39].

E. Comparison to Other Process Mining Tools

This subsection compares our approach with two process mining tools: Disco and Directly Follows visual Miner (DFvM) [40]. The first one has been widely used in process mining domain and the second one is an extension [41] of Inductive visual Miner (IvM) [42]. Both tools can generate a directly follows process model with a given event log.

All the three process mining tools are tested on the event log of strategy-1 patients. The "Expanded event label" representation is used for Disco and DFvM. Note that nodes are of different labels in these two models. In order to handle multi-occurrence events, we associate to each event the order of occurrence among all events of the same label in the

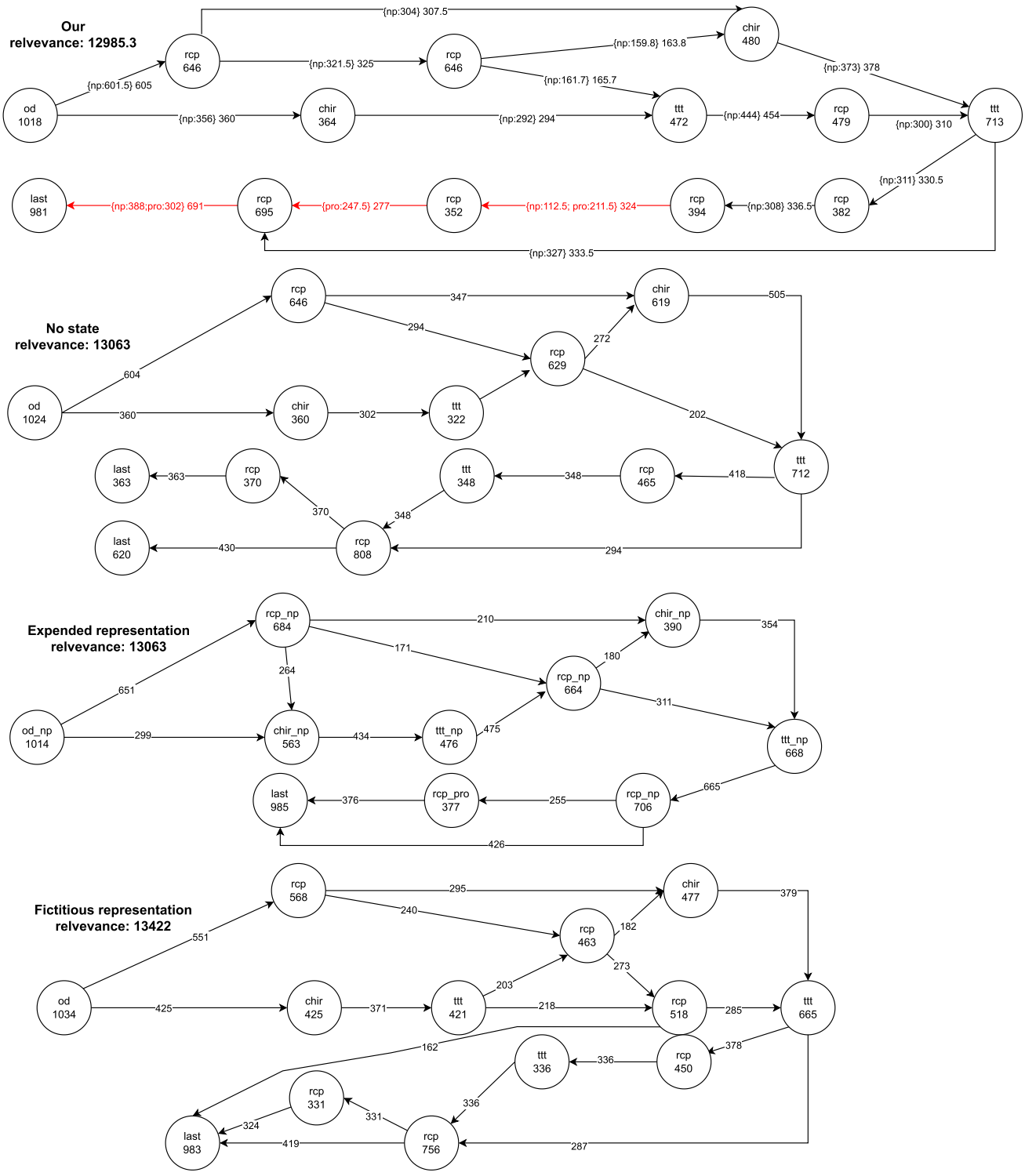

Fig. 9. Process models for strategy one under different representations.

trace. For example, the trace od-rcp-chir-ttt-rcp-$*$-rcp-last with progress at the $*$ position is represented as (od1,np)-(rcp1,np)-(chir1,np)-(ttt1,np)-(rcp2,np)-(rcp3, pro)-(last) in our Disco and DFvM model.

Our approach uses the prevision level of $K = 12$ and $(LB^{node}, LB^{arc}, LB^{attribute}) = (30\%, 15\%, 7.5)|L|$. In order to make a fair comparison, our model is first generated. Then the models of Disco and DFvM are adjusted until the three models have the same number of nodes with parameters (activities, paths) = (10.6%, 0%) for Disco and (activities, paths) = (1, 0.404) for DFvM.

All three models are compared both quantitatively and qualitatively. Qualitative comparison is based on analysis of key informations captured by the three process models. Quantitative comparison is based on common quality measures: our non standard relevance and the traditional fitness.

Consider now the quantitative comparison. We first define the event game for replaying traces in Disco and DFvM

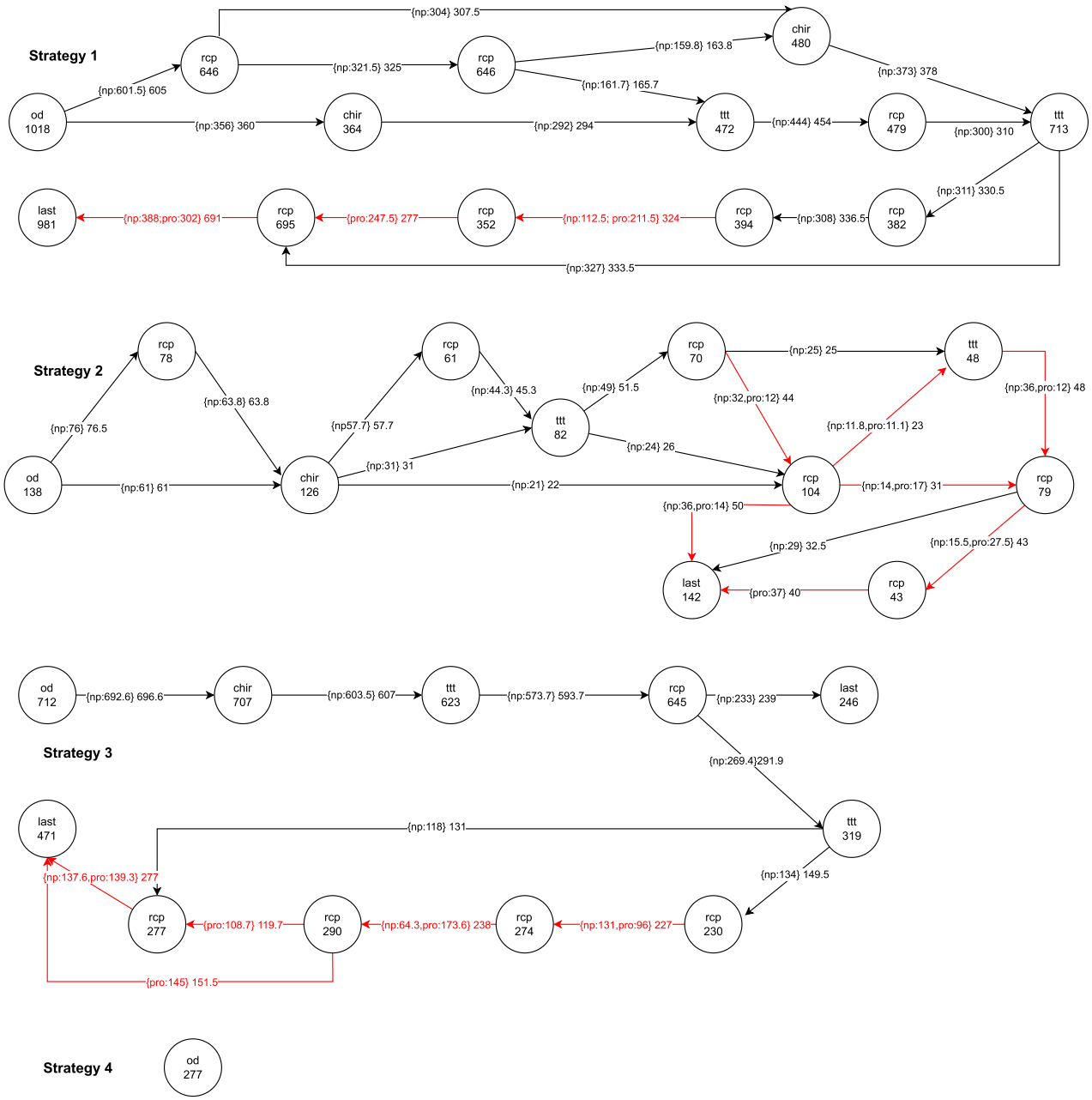

Fig. 10. Comparisons of process models for four strategies.

models. Each trace t of m events is converted into a sequence of m expanded event label. An event is replayed if its expanded label matches exactly the event label of a node if such a node exists and is not replayed if not (the node is unique as nodes have different labels in Disco and DFvM).

For all three models, we define the fitness as the percentage of event log events replayed by each model. For the relevance of Disco and DFvM models, the node relevance is straightforward. For each trace t and for each arc, its arc relevance equals 1 if two events of t are replayed by adjacent nodes of the arc in the order of the arc and no intermediate events are replayed.

Table IV gives the relevance and fitness of all three models. The two quality measures are consistent and both rank the models from the best to the worst as follows: $Our > DFvM > Disco$. Nevertheless, relevance allows a unique and consistent measure of the meaningfulness of all model components with respect to an event log. This makes the rigorous formulation of the optimal process mining subject to minimal meaningfulness for different model components.

TABLE IV
COMPARISON OF PROCESS MINING MODELS BASED ON DIFFERENT METRICS

	Disco	DFvM	Our
Relevance	11468	11959	12985
Fitness	0.815	0.841	0.860

Consider then the qualitative comparison of the three models given by Fig. 11. The following observations are made:

- Better pathway information captured in our model. It clearly shows each pathway with the number of patients and their related health states. The antecedent of disease

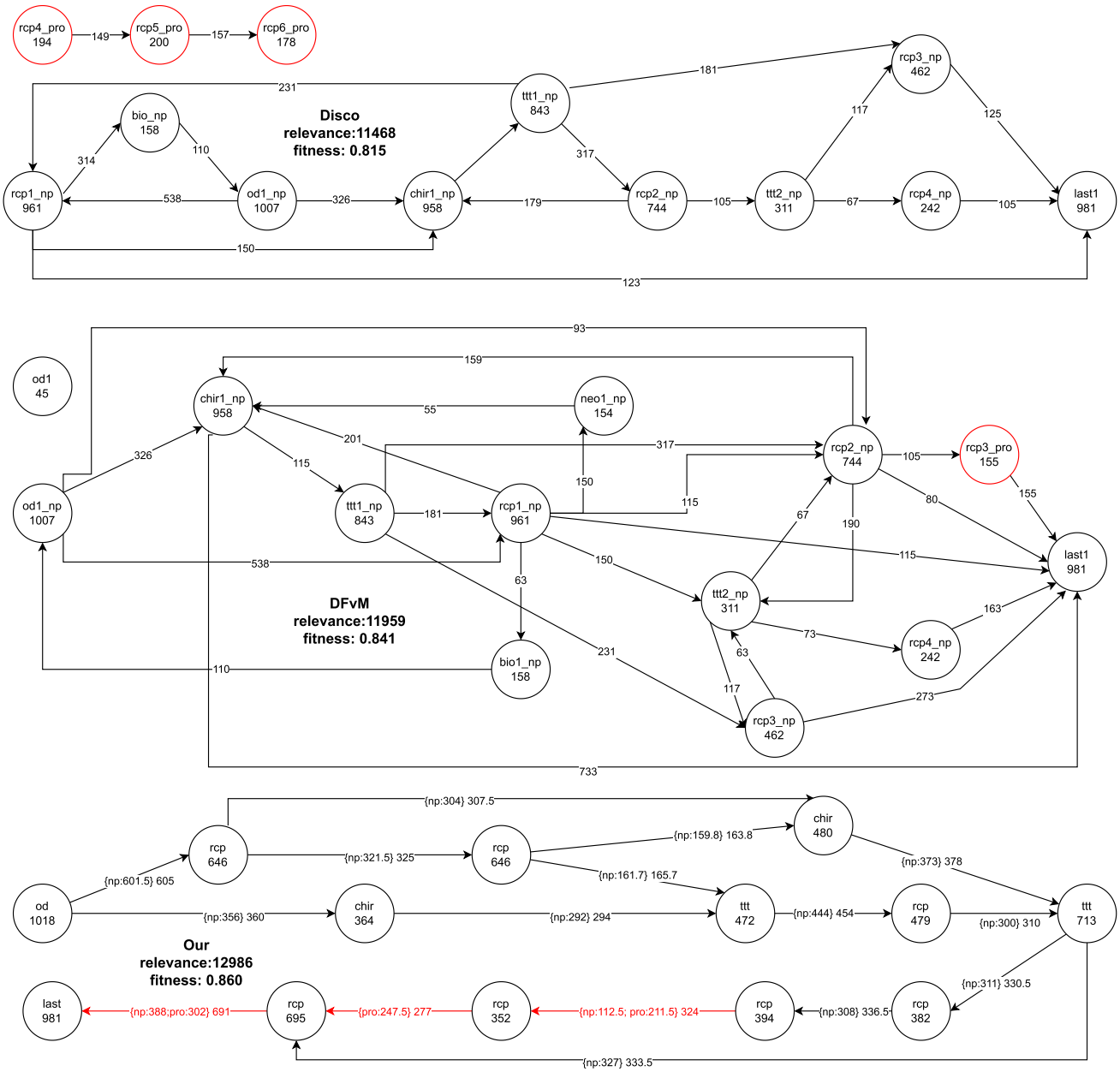

Fig. 11. Comparison of Our model with Disco and DFvM.

progression is modeled to show what has happened before. Disco represents the pathway for the patients in progression separately from the other ones and does not give any information on the antecedent of disease progression. DFvM just highlights rcp3 as the progression event, which is not a precised information.

- Characteristics of strategy 1 better highlighted in "Our" than in "Disco" and "DFvM". The characteristics of rcp either before or after is captured clearly in our model but not evident in the other two models.
- Our model contains only meaninful enough model components and no such guarantee is possible with Disco and DFvM. Disco and DFmV mined less frequent events (od1, bio_np, neo1_np, bio1_np). Thanks to the precision level, our model can filter these events and thus give a more meaningful representation of the event log.

VIII. Conclusion, Extensions and Perspectives

This paper has studied the problem of identifying the optimal process model underlying an event log. The scientific contribution is multiple: 1) The joint consideration of events and attributes; 2) The "relevance" notion is proposed and used to model the studied problem; 3) A dynamic programming approach is developed to calculate the event game in an exact way; 4) An instance generator is proposed; 5) A constructive heuristic algorithm is proposed and compared to other benchmark methods; 6) First application to care pathways of Sarcoma patients and exhaustive comparisons have been made.

We first present some immediate extensions. First, the constraint of at least one attribute associated with each arc of the process model can be easily relaxed and all results and algorithms trivially extend. Second the precision level is defined by some kind of chance constraints (5), (6), (7) represented by

minimal relevance LB of nodes, arcs and attributes. Although we think it more meaningful for practitioners, all results of this paper hold to classical size constraints (maximal number of nodes, arcs) and mixed size and chance constraints. Third the direct optimization of some extended fitness stated in Section III-D allows better comparison with existing algorithms.

Although the proposed process model optimization algorithm is good enough for our case studies, it is quite time consuming and can be improved in various ways. Upper and lower bounds and optimal conditions can help design efficient algorithms. Other neighborhood structure and other metaheuristics are other possible improvements. Further, although the dynamic programming gives the optimal event game, it will be very time consuming when the size (length/number of traces and number of nodes in a process model) of problem is large. In this case, a classification method can be proposed to regroup similar traces. Therefore, the calculation time can be reduced by relaying only the most representative ones.

The healthcare application of this paper is limited to a limit number of care events and the health state of cancer progress or not. It can be extended to other application contexts. The health states can be replaced by other features of patients such as age, sex, etc. This extension leads to process model integrating heterogeneous pathways so that medical practitioners can have a global view over different groups of patients. Unfortunately, this extension leads to large number of events and attributes. The process mining framework of this paper does not apply directly. Future research is needed to take into account automatic event label merging and attribute label merging. Further, for some applications, it is also meaning to consider non acyclic process models.

From the application point of view, the next step of the process mining is the simulation of care pathways of sarcoma patients to evaluate the important performance measures and for what if scenario analysis. How to turn the qualitative process model of this paper into quantitative simulation model requires machine learning of relevant quantitative parameters such as transition probability and inter-event time.

ACKNOWLEDGMENT

The authors would like to thank the NetSarc Network and Community, in particular Prof. Jean-Yves Blay for his valuable comments, Françoise Ducimetiere, Ph.D., and Claire Chemin, M.Sc., for access to the clinical and biological national database dedicated to sarcoma (https://sarcomabcb.org/). The author Zhihao Peng is grateful to Dr. Omar Rifki for his help in experimentation and fruitful discussions.

REFERENCES

[1] J.-Y. Blay, "Getting up-to-date in the management of soft tissue sarcoma," *Future Oncol.*, vol. 14, no. 10s, pp. 3–13, May 2018.

[2] W. van der Aalst, "Process mining," *Commun. ACM*, vol. 55, no. 8, pp. 76–83, 2012.

[3] H. De Oliveira, V. Augusto, B. Jouaneton, L. Lamarsalle, M. Prodel, and X. Xie, "Optimal process mining of timed event logs," *Inf. Sci.*, vol. 528, pp. 58–78, Aug. 2020.

[4] W. van der Aalst, *Process Mining: Data Science in Action*. Berlin, Germany: Springer, 2016.

[5] W. van der Aalst, "The application of Petri nets to workflow management," *J. Circuits, Syst. Comput.*, vol. 8, no. 1, pp. 21–66, Feb. 1998.

[6] A. R. C. Maita et al., "A systematic mapping study of process mining," *Enterprise Inf. Syst.*, vol. 12, no. 5, pp. 505–549, May 2018.

[7] E. De Roock and N. Martin, "Process mining in healthcare—An updated perspective on the state of the art," *J. Biomed. Informat.*, vol. 127, Mar. 2022, Art. no. 103995.

[8] J. Munoz-Gama et al., "Process mining for healthcare: Characteristics and challenges," *J. Biomed. Inform.*, vol. 127, Mar. 2022, Art. no. 103994.

[9] G. P. Kusuma, A. P. Kurniati, E. Rojas, C. D. McInerney, C. P. Gale, and O. A. Johnson, "Process mining of disease trajectories: A literature review," *Student Health Technol. Inform.*, vol. 281, pp. 457–461, May 2021.

[10] E. Rojas, J. Munoz-Gama, M. Sepúlveda, and D. Capurro, "Process mining in healthcare: A literature review," *J. Biomed. Inform.*, vol. 61, pp. 224–236, Jun. 2016.

[11] K. Baker et al., "Process mining routinely collected electronic health records to define real-life clinical pathways during chemotherapy," *Int. J. Med. Inform.*, vol. 103, pp. 32–41, Jul. 2017.

[12] L. Placidi et al., "Process mining to optimize palliative patient flow in a high-volume radiotherapy department," *Tech. Innov. Patient Support Radiat. Oncol.*, vol. 17, pp. 32–39, Mar. 2021.

[13] L. Zhang et al., "Intelligent diagnosis of cervical cancer based on data mining algorithm," *Comput. Math. Methods Med.*, vol. 2021, Nov. 2021, Art. no. 7690902.

[14] A. P. Kurniati, O. Johnson, D. Hogg, and G. Hall, "Process mining in oncology: A literature review," in *Proc. 6th Int. Conf. Inf. Commun. Manage. (ICICM)*, Oct. 2016, pp. 291–297.

[15] A. P. D. Tos, S. Bonvalot, and R. Haas, "Evolution in the management of soft tissue sarcoma: Classification, surgery and use of radiotherapy," *Expert Rev. Anticancer Therapy*, vol. 20, no. 1, pp. 3–13, Apr. 2020.

[16] J.-Y. Blay, "Sarcoma management: Expertise and balance," *Future Oncol.*, vol. 17, no. 21s, p. 1, Jul. 2021.

[17] W. van der Aalst, T. Weijters, and L. Maruster, "Workflow mining: Discovering process models from event logs," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 9, pp. 1128–1142, Sep. 2004.

[18] W. van der Aalst, A. Medeiros, and A. J. Weijters, "Genetic process mining," in *Proc. Int. Conf. Appl. Theory Petri Nets*. Berlin, Germany: Springer, 2005, pp. 48–69.

[19] A. Weijters, W. M. van der Aalst, and A. A. De Medeiros, "Process mining with the HeuristicsMiner algorithm," Technische Universiteit Eindhoven, Eindhoven, The Netherlands, Tech. Rep. 166, 2006, pp. 1–34.

[20] C. W. Günther and W. M. P. van der Aalst, "Fuzzy mining—Adaptive process simplification based on multi-perspective metrics," in *Proc. Int. Conf. Bus. Process Manage.* Berlin, Germany: Springer, 2007, pp. 328–343.

[21] A. Augusto, R. Conforti, M. Dumas, and M. La Rosa, "Split miner: Discovering accurate and simple business process models from event logs," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2017, pp. 1–10.

[22] S. K. L. M. V. Broucke and J. De Weerdt, "Fodina: A robust and flexible heuristic process discovery technique," *Decis. Support Syst.*, vol. 100, pp. 109–118, Aug. 2017.

[23] S. J. Leemans, D. Fahland, and W. M. van der Aalst, "Discovering block-structured process models from event logs—A constructive approach," in *Proc. 34th Int. Conf. Appl. Theory Petri Nets Concurrency*. Milan, Italy: Springer, Jun. 2013, pp. 311–329.

[24] J. M. E. van der Werf, B. F. van Dongen, C. A. Hurkens, and A. Serebrenik, "Process discovery using integer linear programming," in *Proc. Int. Conf. Appl. Theory Petri Nets*. Berlin, Germany: Springer, 2008, pp. 368–387.

[25] M. Prodel, V. Augusto, X. Xie, B. Jouaneton, and L. Lamarsalle, "Discovery of patient pathways from a national hospital database using process mining and integer linear programming," in *Proc. IEEE Int. Conf. Autom. Sci. Eng. (CASE)*, Aug. 2015, pp. 1409–1414.

[26] C. Liu, H. Li, S. Zhang, L. Cheng, and Q. Zeng, "Cross-department collaborative healthcare process model discovery from event logs," *IEEE Trans. Autom. Sci. Eng.*, vol. 20, no. 3, pp. 2115–2125, Jul. 2023.

[27] C. Liu, L. Cheng, Q. Zeng, and L. Wen, "Formal modeling and discovery of hierarchical business processes: A Petri net-based approach," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 53, no. 2, pp. 1003–1014, Feb. 2023.

[28] M. Prodel, V. Augusto, B. Jouaneton, L. Lamarsalle, and X. Xie, "Optimal process mining for large and complex event logs," *IEEE Trans. Autom. Sci. Eng.*, vol. 15, no. 3, pp. 1309–1325, Jul. 2018.

[29] J. Theis, W. L. Galanter, A. D. Boyd, and H. Darabi, "Improving the in-hospital mortality prediction of diabetes ICU patients using a process mining/deep learning architecture," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 1, pp. 388–399, Jan. 2022.

[30] M. Pishgar, J. Theis, M. D. Rios, A. Ardati, H. Anahideh, and H. Darabi, "Prediction of unplanned 30-day readmission for ICU patients with heart failure," *BMC Med. Inform. Decis. Making*, vol. 22, no. 1, p. 117, Dec. 2022.

[31] M. Pishgar et al., "A process mining-deep learning approach to predict survival in a cohort of hospitalized COVID-19 patients," *BMC Med. Inform. Decis. Making*, vol. 22, no. 1, p. 194, Dec. 2022.

[32] W. M. P. van der Aalst, "A practitioner's guide to process mining: Limitations of the directly-follows graph," *Proc. Comput. Sci.*, vol. 164, pp. 321–328, 2019.

[33] F. Ducimetière et al., "Incidence of sarcoma histotypes and molecular subtypes in a prospective epidemiological study with central pathology review and molecular testing," *PLoS ONE*, vol. 6, no. 8, Aug. 2011, Art. no. e20294.

[34] A. Bacon et al., "Incidence and survival of soft tissue sarcoma in England between 2013 and 2017, an analysis from the National Cancer Registration and Analysis Service," *Int. J. Cancer*, vol. 152, no. 9, pp. 1789–1803, May 2023.

[35] E. Bompas et al., "Management of sarcomas in children, adolescents and adults: Interactions in two different age groups under the umbrellas of GSF-GETO and SFCE, with the support of the NETSARC+ network," *Bull. du Cancer*, vol. 108, no. 2, pp. 163–176, Feb. 2021.

[36] S. Bonvalot et al., "Survival benefit of the surgical management of retroperitoneal sarcoma in a reference center: A nationwide study of the French sarcoma group from the NetSarc database," *Ann. Surgical Oncol.*, vol. 26, no. 7, pp. 2286–2293, Jul. 2019.

[37] J.-Y. Blay et al., "Surgery in reference centers improves survival of sarcoma patients: A nationwide study," *Ann. Oncol.*, vol. 30, no. 7, pp. 1143–1153, Jul. 2019.

[38] C. E. Sharon, R. J. Straker, and G. C. Karakousis, "The role of imaging in soft tissue sarcoma diagnosis and management," *Surgical Clinics*, vol. 102, no. 4, pp. 539–550, 2022.

[39] S. Pokras, W.-Y. Tseng, J. L. Espirito, A. Beeks, K. Culver, and E. Nadler, "Treatment patterns and outcomes in metastatic synovial sarcoma: A real-world study in the U.S. oncology network," *Future Oncol.*, vol. 18, no. 32, pp. 3637–3650, Oct. 2022.

[40] S. J. Leemans, E. Poppe, and M. T. Wynn, "Directly follows-based process mining: Exploration & a case study," in *Proc. Int. Conf. Process Mining (ICPM)*, Jun. 2019, pp. 25–32.

[41] S. Leemans, E. Poppe, and M. Wynn, "Directly follows-based process mining: A tool," in *Proc. ICPM Demo Track, CEUR Workshop*. Aachen, Germany: SunSITE Central Europe, vol. 2374, 2019, pp. 9–12.

[42] S. J. Leemans, D. Fahland, and W. M. van der Aalst, "Process and deviation exploration with inductive visual miner," in *Proc. BPM Demo Sessions*, 2014, vol. 1295, no. 8, pp. 46–50.

Zhihao Peng received the Ph.D. degree from Université de Technologie de Belfort Montbéliard, Belfort, France, in 2019.

He is currently a Researcher with the Center for Biomedical and Healthcare Engineering, Department of Healthcare Engineering, Mines Saint-Etienne, Saint-Étienne, France. His research interests include in applying optimization tools to industrial problems, such as logistics and healthcare.

Vincent Augusto received the Ph.D. degree in industrial engineering from École Nationale Superieure des Mines de Saint-Étienne, France, in 2008, and the "Habilitation 'a Diriger des Recherches" degree in health territories engineering in 2016. From 2009 to 2015, he was a Visiting Professor with the Interuniversity Research Center on Business Networks, Logistics and Transport (CIRRELT), Laval University, Quebec, Canada. He is currently a Professor and a Permanent Faculty Member with the Center for Biomedical and Healthcare Engineering, Mines Saint-Étienne. He is also co-responsible for the living laboratory MedTechLab, where he approach design, experiments, and innovation are combined with health care engineering. Since January 2020, he has been the Director of the Center for Biomedical and Healthcare Engineering, Mines Saint-Étienne, for 70 people. His research interests include the application of industrial engineering techniques, such as simulation and optimization, to health care systems, health data, and medical decision aids.

Lionel Perrier received the Ph.D. degree in economics and the Habilitation Diriger des Recherches degree from the University of Lyon, Lyon, France, in 2002 and 2010, respectively. He is currently responsible for the innovations and strategies unit in the Clinical Research Direction of the Léon Bérard Cancer Centre, Lyon. He is a member of the Economic and Public Health Evaluation Committee (CEESP), French National Authority for Health (HAS). He is also with the Léon Bérard Cancer Centre, GATE UMR 5824, University of Lyon, and the Human and Social Sciences Department and the Ethics Focus Group, Léon Bérard Cancer Centre. As a scientific coordinator, a work package leader or scientist, he is in charge of the health economist part of numerous projects and solicited for teaching in this field amongst other in Emlyon Lyon Business School and École Centrale de Lyon.

Xiaolan Xie (Fellow, IEEE) received the Ph.D. degree from the University of Nancy I, Nancy, France, in 1989, and the Habilitation Diriger des Recherches degree from the University of Metz, France, in 1995.

He is currently a Professor of industrial engineering and the Head of the Department of Healthcare Engineering, Center for Biomedical and Healthcare Engineering, École Nationale Supérieure des Mines (ENSMSE), Saint Étienne, France. He is also a Chair Professor with Shanghai Jiao Tong University, China. Before Joining ENSMSE, he was the Research Director with the Institut National de Recherche en Informatique et en Automatique (INRIA) from 2002 to 2005, a Full Professor with École Nationale d'Ingénieurs de Metz from 1999 to 2002, and a Senior Research Scientist with INRIA from 1990 to 1999. He is the author/coauthor of more 350 publications, including over 130 journal articles and six books. He has rich industrial application experiences with European industries. He is a PI of various national and international projects. His research interests include design, planning, and scheduling, supply chain optimization, and performance evaluation, of healthcare and manufacturing systems. He was the Founding Chair of the Technical Committee on Automation in Health Care Management of the IEEE Robotics and Automation Society. He is the General Chair of IEEE CASE2021. He has been an Editor/an Associate Editor of IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING, IEEE TRANSACTIONS ON AUTOMATIC CONTROL, IEEE TRANSACTIONS ON ROBOTICS AND AUTOMATION, and *International Journal of Production Research*. He has a guest editor of various special issues on healthcare engineering and manufacturing systems.