Process mining with event attributes and transition features for care pathway modelling

Omar Rifki, Zhihao Peng, Lionel Perrier, Xiaolan Xie

International Journal of Production Research

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tprs20

Process mining with event attributes and transition features for care pathway modelling

Omar Rifki, Zhihao Peng, Lionel Perrier & Xiaolan Xie

To cite this article: Omar Rifki, Zhihao Peng, Lionel Perrier & Xiaolan Xie (22 Nov 2024): Process mining with event attributes and transition features for care pathway modelling, International Journal of Production Research, DOI: 10.1080/00207543.2024.2427888

To link to this article: https://doi.org/10.1080/00207543.2024.2427888

Published online: 22 Nov 2024.

INTERNATIONAL JOURNAL OF PRODUCTION RESEARCH
https://doi.org/10.1080/00207543.2024.2427888

Taylor & Francis
Taylor & Francis Group

Process mining with event attributes and transition features for care pathway modelling

Omar Rifki [a], Zhihao Peng[b], Lionel Perrier[c] and Xiaolan Xie[b]

[a]LISIC, Univ. Littoral Côte d'Opale, Calais, France; [b]Mines Saint-Etienne, Univ Clermont Auvergne, INP Clermont Auvergne, CNRS, UMR 6158, LIMOS, Saint-Etienne, France; [c]Centre Léon Bérard, CNRS, Université Lumière Lyon 2, Université Jean Monnet Saint-Etienne, Emlyon Business School, GATE, Lyon, France

ABSTRACT

This paper proposes a formal optimisation framework and algorithms for data-aware process mining with event duplication that relaxes the usual one-event-label-one-process-model-node restriction. We put forward a hierarchical representation of the event attribute values and event labelling to achieve the best balance of the complexity and precision of the process model. We posit a new quality measure, relevance, which measures how well and how precisely a process model matches a given event log. The process model optimisation consists of determining (i) the process model with labels and attribute values for each node and transition functions for each arc and (ii) the event game stipulating how each trace of the event log is played in the process model. This article also proposes a dynamic programming algorithm for optimising event games, an exact method for optimal setting of node attributes and arc transition functions, and heuristic algorithms for process model optimisation. Numerical results show the efficiency of the algorithms with respect to relevant benchmarks and an 18% improvement in the model relevance. Applications on sarcoma care pathways reveal their dependency on attributes such as surgery quality and tumour size. Our approach clearly shows how both care event repetition and data impact sarcoma care pathways whereas other data-aware miners fail.

ARTICLE HISTORY
Received 23 April 2024
Accepted 28 October 2024

KEYWORDS
Optimal process mining; data-aware process mining; event attributes and transition features; upper bounds; cancer care pathways

SUSTAINABLE DEVELOPMENT GOALS
SDG 3: Good health and well-being

1. Introduction

This paper addresses the process mining of an event log with event attributes and inter-event features. Input data constitute a set of traces where each trace is a sequence of events, each event is defined by an event label (referred to as 'activity' in some literature) and a set of attributes with each defined by a finite domain, and a transition feature is also associated with two consecutive events. Acknowledging the importance of event repetition (multiple occurrences of the same event label in a trace), this paper seeks to determine the optimal process model (criterion to be discussed) while departing from existing process mining approaches in three significant ways: (i) here, a process model is defined as an acyclic graph with nodes associated with labels with event duplication (i.e. without the usual one-label-one-node restriction) and arcs associated with transition features; (ii) the process model proposed is data-aware with optimally set attribute values and transition features for each node and arc; (iii) and we present a formal optimisation framework of the process mining problem.

Event duplication relaxes the one-label-one-node restriction that is in place in previous published research and significantly extends our process mining approach. At the same time, however, it increases the complexity of process mining, which, in turn, depends on how to play a trace in a given process model, referred to here as an event game. We also address the problem of the optimal event game.

The literature on data-aware process mining is quite limited. The majority of process mining studies published to date simply ignore the large amount of data associated with the event log. Some studies rely on various preprocessing techniques to define event labels according to some level of granularity (van der Aalst and Carmona 2022). Existing data-aware process mining techniques use exhaustive or heuristic searches to determine data-dependent conditions of transitions between two activities (see Section 2 for details). They use the support and confidence of such local transitions instead of global process model quality measures. Most importantly, all such techniques collapse when the one-label-one-node restriction is relaxed.

CONTACT Lionel Perrier lionel.perrier@lyon.unicancer.fr Centre Léon Bérard, CNRS, Université Lumière Lyon 2, Université Jean Monnet Saint-Etienne, Emlyon Business School, GATE, Université Lumière Lyon 2, Université Jean Monnet Saint-Etienne, Lyon 69008, France; Xiaolan Xie xie@emse.fr Mines Saint-Etienne, Univ Clermont Auvergne, INP Clermont Auvergne, CNRS, UMR 6158, LIMOS, Saint-EtienneF-42023, France

With the exception of our previous studies, to our knowledge the formal setting of the process mining problem under consideration has been overlooked to date. It requires formal definitions of the event log, the process model, the event game, and the overall quality measure. The lack of literature on optimal process mining is probably related to the extreme difficulty that it results in a huge number of feasible process models, the lack of analytical expressions of the overall quality measures, and their highly nonlinear relations with the process models. The extensions of this paper make the problem even more complex, as extra decisions must be taken on a number of questions, such as the node attribute values. Nevertheless, the optimisation framework allows us to derive performance bounds, heuristic solutions, and metaheuristic solutions. Below, we will illustrate the benefits of a formal optimisation framework.

This paper is motivated by the data-driven modelling of care pathways of sarcoma patients. As outlined in our previous article (Peng et al. 2024), treatment repetitions such as multiple sarcoma-MDTB (multidisciplinary team board) often signal the degradation of the patient's health. As a result, grouping all MDTB events in a single process model node does not permit appropriate modelling of the disease progress, meaning that event duplication is required. Data-dependency is another important characteristic of sarcoma care pathways. As the diagnosis and treatment of sarcoma both present challenges to clinicians, it is highly recommended that a sarcoma-MDTB is held before surgery and that both are conducted in reference centres. The locations of MDTB and surgery events are crucial factors that have significant impacts on the evolution of care pathways. For instance, research has shown that major differences in care pathways were found between reference centre patients and non-expert centre patients (Blay et al. 2019; Gantzer et al. 2019). Surgery in one of the reference centres of the French NETSARC network was found to be associated with a reduction in the risk of local relapse, progression, and death. Another important factor in treatment is the patient's age; and elderly patients tend to be less likely to undergo surgery (Gingrich et al. 2019).

Although it is motivated by sarcoma care pathway modelling, our approach has further applications. It directly applies to the care pathways of other cancers, which also require data-awareness and event duplication, and to modelling many other healthcare processes. The care pathways of emergency department patients, for example, often involve a second consultation after initial medical examinations. Whether this second consultation is required depends to a large extent on the disease and its gravity. Event replication is also common in semiconductor manufacturing, where certain products return multiple times to the same machines (see Lu and Kumar 1991).

We have previously proposed various optimisation frameworks for process mining. Prodel et al. (2018) formally set out the optimal process mining problem under the one-label-one-node restriction and propose a tabu search approach. The one-label-one node restriction is relaxed by De Oliveira et al. (2020), where a tabu search approach is proposed with a simple event game of first possible node. Our research further relaxed the one-label-one-node restriction and proposed a local optimisation approach based on an optimal event game (Peng et al. 2024).

The purpose of this paper is to propose an optimisation framework and optimisation algorithms for data-aware process mining with event duplication. In brief, the process model allows for the same label in multiple nodes and data-awareness is handled by joint optimisation of the process model, attribute values of its nodes, and the event game of all traces. The proposed approach is illustrated in Figure 1. This paper makes specific contributions to the field by providing:

- an original representation of traces as sequences of event labels each associated with a set of attributes defined on a finite domain and inter-event transition features which are also defined on a finite domain;
- a hierarchical representation of attribute values to face the increased complexity of data-awareness and to allow for the flexible precision level of each attribute in the process model. More specifically, we introduce the following: macro-attribute values defined on an attribute tree for each attribute and macro-labels defined on a label tree;
- an original multi-layer process model built on the original concepts of macro-labels, macro-attribute values, and transition functions (see Peng et al. 2024 for justification of the multi-layer model). Each node in the process model is associated with a macro-label and a macro-attribute value for all relevant attributes with possible occurrence of the same label in multiple nodes. Each arc of the process model is associated with a transition function equal to a subset of transition features;
- a formal process model optimisation framework based on its size constraints and an extended quality measure, termed 'relevance', on the meaningfulness of the model as well as all model components (see Peng et al. 2024 for justification of the relevance). The extension takes into account the precision level of macro-labels, macro-attributes, and transition functions. It allows us to achieve the best balance between precision and the number of events/transitions represented;

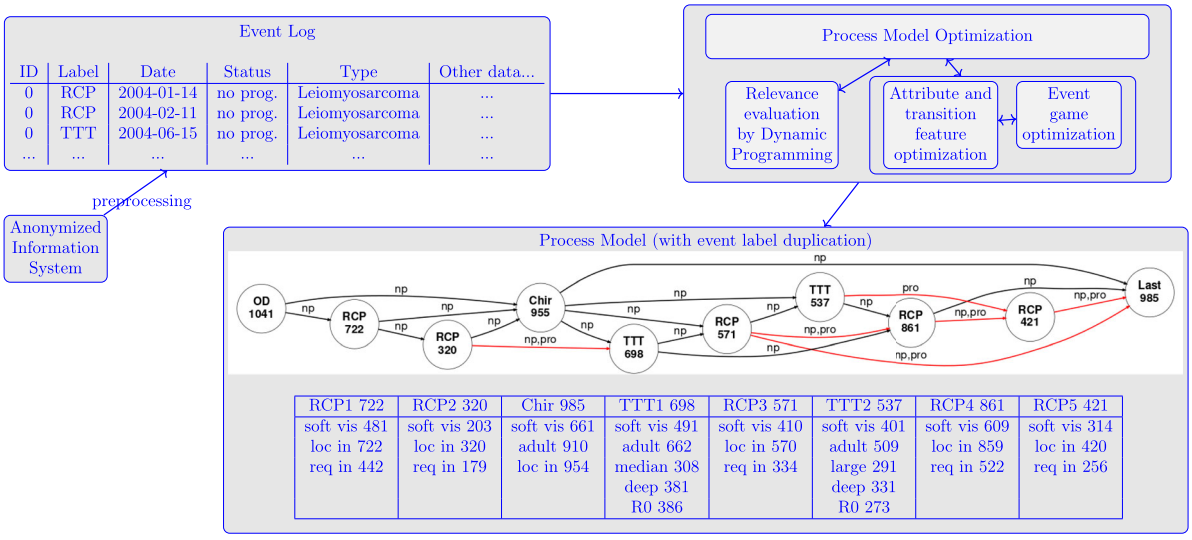


Figure 1. The proposed approach.

- efficient process model optimisation methods. These are based on a series of new theoretical results, including the optimal attributes/transition functions, the optimal event game that maps traces to process model for a given process model, and upper bounds of the optimal model reference. They all rely on an extended marginal relevance of adding a node. We then propose a multi-start local optimisation algorithm and an ant colony algorithm for the solution algorithms;
- an extended numerical experiment based on both generated instances and a real case study to assess the performance of the proposed methods, the benefits of introduction attributes, and how the attributes influence the process models. The introduction of attributes is shown to improve the model relevance by about 18%. The case study reveals a number of relations between the sarcoma care pathways and attributes, such as surgery location, surgery quality, and tumour size/depth. We also demonstrate that our approach can clearly show how both care event repetition and data impact the sarcoma care pathways, whereas other data-aware miners fail.

The following sections of this article are organised as follows: following a literature in Sections 2, 3 formally sets out the problem of process mining optimisation with event label attributes; Section 4 then addresses the optimisation of attributes/transition functions and the event game for a given process model, before Section 5 proposes process model optimisation algorithms; Sections 6–7 then provide numerical experiments, before Section 8 concludes the paper. Detailed algorithm presentations and numerical results are summarised in the main paper and full details are available in an electronic companion, found at: https://emse.fr/ ~ xie/ECpapers/ECoptisarc2024.pdf.

2. Literature review

This section provides a brief review of the relevant literature. It is first worth positioning our paper within the general process mining framework of van der Aalst and Carmona (2022). Our objective is to discover a process model from scratch. Other objectives include conformance checking, enhancement, and monitoring. We approach process modelling predominantly from a control-flow perspective, which focuses on the ordering of events. Other aspects, such as time, data, resources, and costs, can also be integrated into process models. Our research keeps data in its sights. To the best of our knowledge, the relaxation of the one-label-one-node restriction and the optimisation framework have not been considered in the literature, with the exception of our previous works, as mentioned in Section 1. We have therefore limited the scope of our review to the literature on data-aware process mining and a brief literature review on process mining algorithms and care pathway modelling. Van Der Aalst (2016) and van der Aalst and Carmona (2022) provide general introductions to process mining.

2.1. Data-aware process mining approaches

Table 1 positions our approach with respect to relevant data-aware process mining studies. Process discovery methods can be categorised into declarative, procedural, and hybrid approaches that combine elements of both paradigms (Augusto et al. 2018). Declarative approaches specify the rules governing a process, such as transitions

Table 1. Data-aware process mining approaches.

Study	Object-centric	Discovery	Conformance Checking	Enhancement	Monitoring	Declarative	Procedural	Integration
Rozinat and van der Aalst (2006)				✓			✓	
Maggi et al. (2013)		✓				✓		✓
De Leoni and van der Aalst (2013b)				✓			✓	
De Leoni and van der Aalst (2013a)			✓			✓		
Taghiabadi et al. (2014)			✓			✓		
Borrego and Barba (2014)			✓			✓		
Batoulis et al. (2015)				✓			✓	
Burattin, Maggi, and Sperduti (2016)			✓			✓		
Schönig et al. (2016)		✓				✓		✓
Mannhardt et al. (2016)			✓				✓	
Mannhardt et al. (2017)		✓					✓	✓
Li, de Carvalho, and van der Aalst (2017)	✓	✓				✓		✓
Leno et al. (2020)		✓				✓		✓
Bergami et al. (2021)			✓			✓		
Felli et al. (2021)			✓				✓	
Bano et al. (2021)				✓			✓	
Fahland (2022)	✓	✓				✓		✓
Alman et al. (2022)					✓	✓	✓	
Mannhardt et al. (2023)		✓		✓			✓	✓
Our		✓					✓	✓

Note: The column 'Integration' concerns the full integration of data perspective within the control-flow perspective for discovery approaches, meaning that data influence the control-flow structure.

between two events and their activation conditions. Procedural approaches, also called imperative approaches, specify the flows admitted by a process. As a result, the output of the former is a collection of rules whereas it is a complete process model (Petri net model or simple automata graphs) for the latter.

Procedural or imperative methods focus on the continuous evolution of the process objects (Fahland et al. 2009). Rozinat and van der Aalst (2006) were the first to incorporate a data perspective into process models by applying a classification algorithm, specifically decision trees, to determine whether decision points in Petri nets are influenced by data. De Leoni and van der Aalst (2013b) proposed an enhancement of this work, addressing 181 invisible transitions and multiple transitions in Petri nets, by defining a Petri net structure that integrates data. Batoulis et al. (2015) then introduced a semi-automatic method for identifying decision logic in BPMN (Business Process Model and Notation) models.

Mannhardt et al. (2017) were the first to propose an imperative method where data attributes directly impact the construction of the control-flow, rather than merely adding data attributes to an existing process model as previous approaches did. Their method can reveal infrequent paths depending on specific data attribute values using classifiers. Mannhardt et al. (2023) extended the data-aware perspective to stochastic Petri nets, which are an advanced form of Petri nets that explicitly encode the occurrence probabilities of transitions.

Declarative methods, however, focus on the logic that governs the overall interplay of actions and objects in a process (Fahland et al. 2009). The most commonly used declarative modelling language is Declare, introduced by Pesic, Schonenberg, and van der Aalst (2007). Maggi et al. (2013) were the first to incorporate data conditions into the semantics of the Declare language by means of First Order Linear Temporal Logic rules. Starting with an initial control-flow Declare constraint, this method uses classification techniques, such as decision trees, to discover conditions on data attributes that can differentiate between constraint fulfilment and violation.

Burattin, Maggi, and Sperduti (2016) introduce a formal multi-perspective version of Declare (MP-Declare). Schönig et al. (2016) present an implementation of MP-Declare, which relies on RXES, a standardised architecture for storing event log data. Conditional constraints can be easily discovered through standard SQL queries on the event data, although these SQL queries must be formulated.

Leno et al. (2020) propose two alternative approaches to discover MP-Declare constraints. The first approach uses clustering to identify groups in the target and activation payloads (representing both sides of the Declare constraint) combined with a rule mining technique. The second approach employs a redescription mining technique, which is an unsupervised knowledge discovery method. Experiments demonstrate the high effectiveness of the clustering-based approach in rediscovering

constraints artificially injected into a log, compared to the second approach.

A new process specification that moves beyond a 'case-centric' approach is currently emerging. These so-called 'object-centric' process specifications apply constraints to multiple objects and their interactions (see Di Ciccio and Montali 2022). Although still in its early stages, several studies adopt this richer representation, which accounts for event data (Fahland 2022; Li, de Carvalho, and van der Aalst 2017).

The approach of Leno et al. (2020) begins with the DECLARE constraints as outlined by Maggi, Bose, and van der Aalst (2012). These constraints are then refined with data-aware conditions through the following steps: (i) extracting fulfilment and violation feature vectors using the target and activation payloads of the constraints; (ii) applying the K-medoids clustering algorithm to identify groups with similar target payloads; and (iii) employing the RIPPER rule-based classification algorithm to describe the clusters. In an alternative approach, steps (ii) and (iii) are replaced by applying redescription mining algorithms to the features of both the activation and target payloads. Apart from the differences in paradigms, the key distinction between our approach and that of Leno et al. (2020) lies in the local/global consideration of data perspective. In Leno et al. (2020), the initial DECLARE constraints condition the subsequent attribute considerations, whereas in our approach attribute values are integrated at each step of the optimisation algorithm.

2.2. Process mining algorithms

Algorithms used in process discovery are often domain-specific, starting with the Alpha Miner (van der Aalst, Weijters, and Maruster 2004). This algorithm identifies specific patterns within the event log sequences and outputs in a Petri net. Although its runtime scales linearly with the size of the event log, the algorithm is highly sensitive to noise and incompleteness (van der Aalst 2011). Heuristics miners address this issue by considering both the order and frequency of events, making them the most widely used algorithms in practice (Gomes, de Lacerda, and da Silva Fialho 2022). The Inductive Miner (Bogarín Vega, Menéndez, and Romero 2018) further enhances the results of both Alpha and Heuristics miners, offering better handling of infrequent behaviours and large event logs. Other miners include the Fuzzy Miner (Günther 2009), which is more effective for processes that lack clear structure and behaviour, and Directly Follows Graphs, a method that relies on statistical analysis of how frequently one event follows another. Digging deeper into the relevant algorithms, it becomes clear that, with the exception of our previous work, none relies on a formal setting of the optimal process mining problem. Some algorithms, such as genetic miners (de Medeiros, Weijters, and van der Aalst 2007) do indeed use certain ingredients of metaheuristics but they do not rely on a formal optimisation problem setting. As a result, they do not fully explore the potential of optimisation algorithms and cannot evaluate the optimality gap with respect to the true optimum.

2.3. Process mining in healthcare

The majority of healthcare process mining applications are devoted to business process analysis for work flow analysis (see Chapter 14 in van der Aalst and Carmona 2022) and data flow (see Liu et al. 2021, 2023). None of these, however, is data-aware. From the literature review provided by Kusuma et al. (2021), we can see that there are even fewer applications of process mining in care pathway or disease trajectory modelling. Jensen et al. (2014) use a clustering approach to group the disease trajectories of 6.2 million patients into patterns centred on a small number of key diagnoses, such as chronic obstructive pulmonary disease (COPD) and gout. Mannhardt et al. (2017)'s approach was applied by Kusuma et al. (2020) to the disease trajectories of ICU patients and by Pang et al. (2021) to acute care in critical illness scenarios. As they are built on existing process mining approaches, however, these healthcare studies all suffer adhering to a one-label-one-node restriction and, in the majority of cases, they lack data awareness.

2.4. Summary

With respect to the goal of this paper, the existing process mining approaches all need the one-label-one-node restriction and none formally sets out the optimal process mining problem. While some data-aware process mining approaches do indeed exist, they all rely on local patterns regarding the next immediate events of any given one instead of the global performance of the process model. Our approach to process mining is, however, data-aware without adhering to the one-label-one-node restriction, based on a formal optimisation framework and algorithms for optimisation of global performances.

3. Problem setting

This section formally describes the problem with formal definitions of the input event log, the output process model, the event game stipulating how a trace traverses a process model, the goodness measure, and, finally, the process model optimisation formulation. A toy example is given in the electronic companion.

3.1. Event log

This subsection provides formal definitions of the key input data used for the process mining, which comprise an event log consisting of a set of traces. Each trace is an order sequence of labels and a transition feature between any two consecutive labels. Transition features are often referred to as transitions for the sake of brevity. Formally speaking, the event log is built upon two alphabet sets:

- B: a finite set of event labels;
- S: a finite set of transition features.

In order to take into account complex event logs, we associate with each label some attributes defined on a finite domain. More specifically,

- H: a finite set of label attributes or attributes for short;
- $H^b \in H$: the set of attributes of label $b \in B$;
- D^h: the finite domain of attribute $h \in H$.

Definition 3.1 (Event): An event denoted e is a label together with a feasible value for each of its attributes. Let E be the set of events, i.e. $E = \{(b, a) : b \in B, a \in \prod_{h \in H^b} D^h\}$.

Definition 3.2 (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 feature $s_i \in S$ 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 label of ith event of trace t, i.e. e_i. It will be called the label function;
- $\varphi(t, i, h)$: the attribute h of the ith event of trace t. It will be called attribute function and by convention $\varphi(t, i, h)$ is undefined if $h \notin H^{\varepsilon(t,i)}$;
- $\sigma(t, i, i+1)$: the feature associated with transition (e_i, e_{i+1}), i.e. s_i. It will be called the transition function. The transition function will also be extended to nonconsecutive events with $\sigma(t, i, j) = \{s_i, \ldots, s_{j-1}\}$.

Definition 3.3 (Event log): An event log L is a set of traces $L = \{t_1, \ldots, t_{card(L)}\}$. It consists of the first part of input data of our process mining problem.

Example: In our sarcoma application, the set of event labels is $B = OD$ = *original diagnosis, rcp* = *MTDB, chir* = *surgery,*..., the set of transition features is S = *pro* = *disease progression, np* = *no progression*, the set of attributes is H = loc = location, type, sev = severity, MD = surgeon, qlt = quality,..., attributes of label *chir* are H^{chir} = *loc, MD, qlt*, the attribute *loc* takes value over all French oncology services. Figure 2 is a trace of three events.

3.2. Process model representation

This subsection describes the solution (output) of our process mining problem. In approximate terms, it is a multi-layer network model in which each node is associated with a subset of labels and related attributes, arcs connect lower-layer nodes to higher-layer nodes, and each arc is associated with a subset of transition features. Data-awareness is represented by the attribute values of the nodes and transition features and two nodes can share the same event label. The introduction of label attributes makes it difficult to generate compact yet meaningful process models. In order to overcome this difficulty, we propose a hierarchical representation based on macro-labels, macro-events, event classes, label trees, and attribute trees.

Definition 3.4 (Macro-label): A macro-label n is a nonempty set of labels, i.e. $\widehat{b} \in 2^B \backslash \varnothing$. Its attribute set $H^{\widehat{b}}$ is naturally defined as the union of attribute sets of all its labels.

Definition 3.5 (Label tree): An label tree denoted $\widehat{B}$ is a set of macro-labels, i.e. $\widehat{B} \subseteq 2^B$ such that (1) $B \subseteq \widehat{B}$; (2) $\forall\, \widehat{b}, \widehat{b}' \in \widehat{B}$, either $\widehat{b} \cap \widehat{b}' = \varnothing$ or ($\widehat{b} \subset \widehat{b}'$ or $\widehat{b}' \subset \widehat{b}$).

Definition 3.6 (Attribute tree): An attribute tree of attribute h is a set $\widehat{D}^h \subseteq 2^{D^h} \backslash \varnothing$ such that (1) $D^h \subseteq \widehat{D}^h$, (2) $\forall\, x, x' \in \widehat{D}^h$, either $n \cap x' = \varnothing$ or ($x \subset x'$ or $x' \subset x$), 3) $D^h \in \widehat{D}^h$. We also call entities of the attribute tree macro-attribute. $x = D^h$ is called the root macro-attribute.

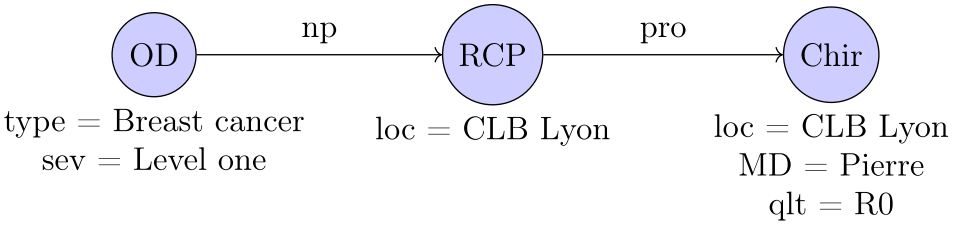


Figure 2. A trace.

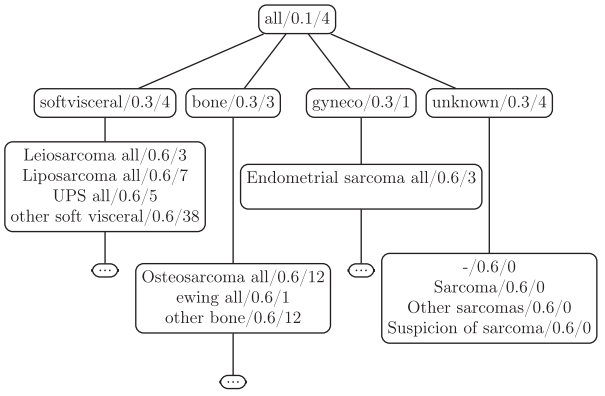


Figure 3. Attribute tree of attribute *type*.

Definition 3.7 (Macro-event): A macro-event is a macro-label together with a value on the attribute tree for each attribute. Let $\widehat{E}$ be the set of macro-events, i.e. $\widehat{E} = \{(\widehat{b}, a) : \widehat{b} \in \widehat{B}, a \in \prod_{h \in H^{\widehat{b}}} \widehat{D}^h\}$.

Figure 3, for example, presents the attribute tree of attribute *type* of sarcoma. It is a four-level tree that gives the macro-value, precision weight, and number of immediate descendants for each node. It divides sarcoma patients into four groups: soft tissue and visceral (*soft-visceral*); *bone*; *gynecology*; and *unknown*. Each group is further decomposed. The precision weight to be defined later serves to represent the process mining perspective the decision maker focuses on.

Remark 3.1: The tree structure allow a compact process model showing events in different precision level automatically. When designing the tree structure, it should be defined in a way which the clusters have a medical meaning.

Definition 3.8 (Process model): A process model denoted by *PsM* is an acyclic graph where each node is a macro-event and each arc a subset of transition features. More specifically, it is a five-uplet $(N, A, \varepsilon, \varphi, \sigma)$ where:

- $N = N_1, N_2, \ldots, N_K$ with K being the number of layers and $N_k \subset \widehat{E}$ being nodes of layer k. The notation N_k is extended to $N_{[\![k,k']\!]}$ to indicate nodes of layers k to $k' > k$;
- $\varepsilon(PsM, n) \in \widehat{B}$ returns macro-label of node n and ε is called the label function of nodes;
- $\varphi(PsM, n, h) \in \widehat{D}^h$ returns value of attribute h for node n and φ is called the attribute function of nodes;
- $\varsigma(PsM, n) \in \widehat{E}$ associate an event class to each node n and $\varepsilon(PsM, n) \neq \varepsilon(PsM, n') = \varnothing, \forall\, n, n' \in N_k$;
- $\sigma(PsM, n, n') \subset S$ associates to each arc $(n, n') \in A$ a set of transition features and σ is called the transition function.

Figure 1 gives an example of the process model with five nodes of label *RCP*. Note that the same notation is used for event (transition) 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.

3.3. 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 states of a trace can be represented by a given process model.

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 a macro-label containing the event label;
- events of a trace are represented by nodes in increasing order of layers.

Definition 3.9 (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 ith event of t, either its node mapping $\gamma(t, i) \in N$ or $\gamma(t, i)$ is undefined denoted as $\gamma(t, i) \uparrow$. Further, 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 3.10 (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 kth 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. $IM(\gamma, t) = \{\gamma(t, [1]), \gamma(t, [2]), \ldots\}$.

3.4. Goodness measures of a process model and an event game

This subsection proposes goodness measures of a process model controlled by an event game, referred to here as relevance. A trace generates a node relevance score at the nodes it visits and an arc relevance for arcs traversed. The model relevance is the sum of the total node and arc relevance.

More precisely, visiting a node n by a trace t generates, for node n, a label relevance, an attribute relevance and their linear combination called node relevance. Attribute

relevance depends on how well the node attribute values match those of the trace. Both label and attribute relevance scores depend on the precision weights $g^{precis}_{\widehat{B}}$ and $g^{precis}_{\widehat{D}^h}$ of the node label and attributes. Similarly, traversing an arc (n, n') by a trace t generates a cross relevance, a transition relevance and their linear combination called arc relevance with transition relevance depending on how well the arc transition feature matches that of the trace.

The followings are formal definitions of the relevance scores. 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 states associated to arcs.

3.4.1. Relevance with respect to a given trace

Definition 3.11: For a given process model $PsM = (N, A, \varepsilon, \varphi, \sigma)$, an event game γ and a trace $t = e_1(s_1)e_2(s_2)\dots e_m$, let $\{n_1, \dots n_J\}$ be the image of t and $e_{[1]}, \dots, e_{[J]}$ be the corresponding events of t. The local relevance with respect to trace t is defined as follows:

- $f^{label}(\gamma, t, n)$ local label relevance of a node n with $f^{label}(\gamma, t, n) = g^{precis}_{\widehat{B}}(\varepsilon(PsM, n))$ if n belongs to the image of t and 0 otherwise;
- $f^{attribute}(\gamma, t, n, h)$ local attribute relevance of an attribute h associate with a node n with $f^{attribute}(\gamma, t, n, h) = g^{precis}_{\widehat{D}^h}(\varphi(PsM, n, h))$ if n belongs to the image of t, h an attribute of the corresponding event e_i and its attribute value $\varphi(t, i, h) \in \varphi(PsM, n, h)$ and 0 otherwise;
- $f^{node}(\gamma, t, n)$ local node relevance of a node n with $f^{node}(\gamma, t, n) = (1 - \lambda_1) f^{label}(\gamma, t, n) + \lambda_1 \|H^n\|^{-1} \sum_{h \in H^n} f^{attribute}(\gamma, t, n, h)$ if n belongs to the image of t and 0 otherwise. By convention, $f^{node}(\gamma, t, n) = f^{label}(\gamma, t, n)$ if node n has no attribute;
- $f^{cross}(\gamma, t, n, n')$ local cross relevance of an arc (n, n') equal to 1 if $n = n_j$ and $n' = n_{j+1}$ for some j and 0 otherwise;
- $f^{trans}(\gamma, t, n, n')$ local transition relevance of (n, n') with $f^{trans}(\gamma, t, n, n') = ([j+1] - [j])^{-1} \|\sigma(t, [j], [j+1]) \cap \sigma(PsM, n, n')\|$ if $n = n_j$ and $n' = n_{j+1}$ for some j and 0 otherwise, i.e. the ratio of transitions between $e_{[i]}$ and $e_{[i+1]}$ having transition features in $\sigma(PsM, n, n')$;
- $f^{arc}(\gamma, t, n, n')$ local arc relevance of arc (n, n') with $f^{arc}(\gamma, t, n, n') = (1 - \lambda_2) f^{cross}(\gamma, t, n, n') + \lambda_2 * g^{precis}_{S}(\sigma(PsM, n, n')) f^{trans}(\gamma, t, n, n')$;
- $f^{model}(\gamma, t) = \sum_{n \in N} f^{node}(\gamma, t, n) + \alpha \sum_{(n,n') \in A} f^{arc}(\gamma, t, n, n')$ local model relevance.

where $\lambda_1, \lambda_2 \in [0,1)$ is the relative weight of attribute, transition with respect to node, arc, $\alpha > 0$ is the weight of arcs with respect to nodes, $g^{precis}_{\widehat{B}}(\widehat{b}) \le 1$ is the precision score of a macro-label $\widehat{b}$ equal to 1 if $\widehat{b} \in B$, $g^{precis}_{\widehat{D}^h}(x) \le 1$ is the precision score of a macro-attribute x with $x \subset x'$ implying $g^{precis}_{\widehat{D}^h}(x) \ge g^{precis}_{\widehat{D}^h}(x')$, $g^{precis}_{S}(\sigma) \le 1$ is the precision score of transition feature subset σ associated with an arc equal to 1 if σ is a singleton. In this paper, $g^{precis}_{S}(\sigma)$ equal to some nonincrceasing function of $\|\sigma\|$ will be used.

For example, in the attribute tree of Figure 3, the root node *all* has the least precision weight 0.1, the second level macro-attribute *bone* higher precision weight 0.3, the third level macro-attribute *Leiosarcoma* weight 0.6.

Note that we do not impose $g^{precis}_{\widehat{D}^h}(x)$ equal to 1 if $x \in D^h$ in order to account for imbalance of attribute trees. For example, in our case study, the attribute 'location' has two macro-value 'inside' and 'outside' and more precise location is given for 'inside' but not for 'outside'. 'Outside' is then considered as a basic attribute value but does not have the same precision as the other basic values.

3.4.2. Relevance with respect to an event log

Definition 3.12: 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:

- $F^{label}(\gamma, n) = \sum_{t \in L} f^{label}(\gamma, t, n)$ label relevance of node n;
- $F^{attribute}(\gamma, n, h) = \sum_{t \in L} f^{attribute}(\gamma, t, n, h)$ attribute relevance of attribute h at node n;
- $F^{node}(\gamma, n) = \sum_{t \in L} f^{node}(\gamma, t, n)$ node relevance of node n;
- $F^{cross}(\gamma, n, n') = \sum_{t \in L} f^{cross}(\gamma, t, n, n')$ cross relevance of arc (n, n');
- $F^{trans}(\gamma, n, n') = \sum_{t \in L} f^{trans}(\gamma, t, n, n')$ transition relevance of arc (n, n');
- $F^{arc}(\gamma, n, n') = \sum_{t \in L} f^{arc}(\gamma, t, n, n')$ arc relevance of arc (n, n');
- $F^{model}(\gamma) = \|L\|^{-1} \sum_{t \in L} f^{model}(\gamma, t)$ model relevance.

Note that the model relevance is normalised to the size of the event log. By definition, we also have:

$$F^{model}(\gamma) = \|L\|^{-1} \sum_{n \in N} F^{node}(\gamma, n) + \alpha \|L\|^{-1} \times \sum_{(n,n) \in A} F^{arc}(\gamma, n, n') \quad (1)$$

$$F^{node}(\gamma, n) = (1 - \lambda_1) F^{label}(\gamma, n) + \lambda_1 \|H^n\|^{-1} \times \sum_{h \in H^n} F^{attribute}(\gamma, n, h) \quad (2)$$

$$F^{arc}(\gamma, n, n') = (1-\lambda_2)F^{cross}(\gamma, n, n') + \lambda_2 * g_S^{precis}(\sigma(PsM, n, n'))F^{trans}(\gamma, n, n') \quad (3)$$

3.4.3. Upper bounds of the model relevance

Assuming the perfect representation of all events and transitions leads to the node relevance $\|t\|$ and arc relevance $\|t\|-1$ for each trace t and the following upper bound:

$$F^{model}(\gamma) \le Bound1 = \|L\|^{-1}\sum_{t\in L}(\|t\| + \alpha(\|t\|-1)) \quad (4)$$

We further derive a better bound for the case without macro-label, i.e. $\widehat{B} = B$. Assuming perfect representation of each event if played and the perfect representation of transition between played events of the same trace, each trace t has $\|\gamma(t)\|$ events played and we have the following bound:

$$F^{model}(\gamma) \le \|L\|^{-1}\sum_{t\in L:\|\gamma(t)\|>0}(\|\gamma(t)\| + \alpha(\|\gamma(t)\|-1)) \quad (5)$$

and hence

$$F^{model}(\gamma) \le \|L\|^{-1}\left((1+\alpha)\sum_{t\in L}\|\gamma(t)\| - \alpha\sum_{t\in L}1(\|\gamma(t)\|>0)\right) \quad (6)$$

where $\sum_{t\in L}\|\gamma(t)\|$ is the total number of events played and $\sum_{t\in L}1(\|\gamma(t)\|>0)$ the total number of traces played. The number of events played is bounded by:

$$\sum_{t\in L}\|\gamma(t)\| \le Z = \sum_{i=1}^{UB^{node}}\|L_{(b_{[i]},j_{[i]})}\| \quad (7)$$

where $L_{(b,j)}$ with $j \ge 1$ is the set of traces containing at least j events of label b and sorted as follows: $\|L_{(b_{[1]},j_{[1]})}\| \ge \|L_{(b_{[2]},j_{[2]})}\| \ge \ldots$. The notation $L_{(b,j)}$ is used to take into account nodes of same label at different layers. Further,

$$\sum_{t\in L}1(\|\gamma(t)\|>0) \ge K(L,Z) \quad (8)$$

where $K(L,Z) = \inf\{K : \|t_{[1]}\| + \cdots + \|t_{[K]}\| \ge Z\}$ with the traces of event log L being sorted as $\|t_{[1]}\| \ge \|t_{[2]}\| \ge \ldots$. To summarise:

$$F^{model}(\gamma) \le Bound2 = (1+\alpha)\|L\|^{-1}Z - \alpha\|L\|^{-1}K(L,Z) \quad (9)$$

3.5. Process model optimisation formulation

This subsection gives the formal definition of the process model optimisation problem. It consists of determining a process model and an event game in order to maximise the model relevance subject to model size constraints. Note that constraints are needed to avoid the generation of a spaghetti-like, messy, and over-complicated model.

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

$$\max_{PsM,\gamma} F^{model}(\gamma) = \|L\|^{-1}\sum_{n\in N}F^{node}(\gamma, n) + \alpha\|L\|^{-1}\sum_{(n,n')\in A}F^{arc}(\gamma, n, n') \quad (10)$$

subject to:

$$PsM = (N, A, \varepsilon, \sigma) \quad (11)$$

$$N = N_1 \cup \cdots \cup N_K, \quad \text{with } N_k \subseteq \widehat{E} \quad (12)$$

$$\varsigma(PsM, n) \in \widehat{E}, \quad \forall\, n \in N \quad (13)$$

$$\varepsilon(PsM, n) \ne \varepsilon(PsM, n'), \quad \forall\, n, n' \in N_k \quad (14)$$

$$\sigma(PsM, n, n') \ne \varnothing, \quad \forall\, (n, n') \in A \quad (15)$$

$$\|N\| \le UB^{node} \quad (16)$$

$$\|A\| \le UB^{arc} \quad (17)$$

where constraint (12) defines the maximum number of layers, (13) associates event node with a macro-event, (14) imposes different macro-labels for the same layer, (15) restricts to arcs of nonempty transition feature sets, (16)–(17) are size constraints of nodes and arcs.

Note that constraint (15) can be removed as $g_S^{precis}(s) = 1$ for all singleton s implies the existence of nonempty optimal transition function for all arcs. Further an alternative process mining problem can be defined by replacing the size constraints (16)–(17) minimal relevances $LB^{node}, LB^{arc}, LB^{trans}$ of nodes, arcs and transition features.

4. Optimization of attribute/transition functions and event game

This section addresses the optimisation of the event game and attribute/transition functions for a given process model. We consider first the optimisation of attribute/transition functions with all others being given, then the optimisation of the event game with a similar dynamic programming algorithm of our previous paper (Peng et al. 2024). Finally, we explore the joint optimisation of the event game and attribute/transition functions.

4.1. Optimal attribute/transition functions

This subsection considers the optimisation of attributes and transition functions for any given process model and the event game. In this case, the images and footprints of all traces are given. Each node then has a given set of events represented by it and each arc has a given set of traces traversing it. Under such conditions, the attribute functions and the transition functions can be determined independently. Furthermore, the attribute functions can be determined separately for different nodes and different attributes. Similarly, the transition functions can be determined separately for different arcs.

Consider first the attribute function optimisation. The value of attribute h of node n only affects the attribute relevance $F^{attribute}(\gamma, n, h)$. Let $x = \varphi(PsM, n, h)$ be the macro-value of attribute h at node n, by definition,

$$F^{attribute}(\gamma, n, h) = g^{precis}_{\widehat{D}^h}(x) P^h(n, x) \tag{18}$$

where $P^h(n, x)$ is the frequency of triplet (n, h, x), i.e. the total number of traces visiting node n with value of attribute h belonging to x. For each macro-value x,

$$P^h(n, x) = \sum_{x' \in x \cap D^h} P^h(n, x') \tag{19}$$

The above provides a simple way to determine the optimal value of h for node n. We first determine the basic attribute value frequencies $P^h(n, x)$, use them for simple computation of the attribute relevance for all possible values and then determine the optimal attribute value.

Consider now the transition function optimisation. Let $s = \sigma(PsM, n, n')$ be the feature set of arc (n, n'). It only affects the transition relevance $g^{precis}_S(s) F^{trans}(\gamma, t, n, n')$ denoted by $V(s, n, n')$. If s is a singleton, then $V(s, n, n') = F^{trans}(\gamma, t, n, n')$. For other transition function value,

$$V(s, n, n') = g^{precis}_S(s) \sum_{s' \in s} V(s', n, n') \tag{20}$$

Since this paper focuses on precision function of the form $g^{precis}_S(s) = g(\|s\|)$, the optimal transition function value can be determined by the maximum among the following:

$$\max\{g(1) V(s_1, n, n'), g(2)(V(s_1, n, n') + V(s_2, n, n')), \ldots\} \tag{21}$$

where $V(s, n, n')$ for all $s \in S$ are sorted in nondecreasing order $V(s_i, n, n') \geq V(s_{i+1}, n, n')$. This provides the algorithm for optimisation of the transition function. By scanning the images of all traces, we can first compute all basic singleton transition function $V(s)$. These values are then sorted in nondecreasing order and then used for determined the optimal transition function as in the above.

We now address the efficient computation of $P^h(n, x)$ for basic attribute value $x \in D^h$ and $V(s, n, n')$ for singleton s. To start, we set to 0 all these values. For all traces $t = e_1(s_1) e_2(s_2) \ldots e_m$, let $\{n_1, \ldots n_J\}$ be the image of t and $e_{[1]}, \ldots, e_{[J]}$ be the corresponding events of t. For each node n_i and all attribute $h \in H^{\varepsilon(t,[i])}$, $P^h(n, x) \leftarrow P^h(n, x) + 1$ with $x = \varphi(t, [i], h)$. For all (n_i, n_{i+1}) being an arc of PsM, $V(s_j, n_i, n_{i+1}) \leftarrow V(s_j, n_i, n_{i+1}) + ([i+1] - [i])^{-1}$ for all transition feature s_j between event $e_{[i]}$ and $e_{[i+1]}$.

Algorithm 1: Attribute and transition function optimisation

Step 1. $P^h(n, x) \leftarrow 0$, $V(s, n, n') \leftarrow 0$ for all base attribute value s and transition feature s
Step 2. Play all traces t, $P^h(n, x) \leftarrow P^h(n, x) + 1$ for each node visit n with attribute (h, x), $V(s, n, n') \leftarrow V(s, n, n') + 1/m$ for each transition s if t crosses (n, n') in m transitions
Step 3. For each node n and attribute h, determine attribute relevance by (18)- -(19) for all macro-value x and select the optimal macro-value;
Step 4. For each arc (n, n'), determine the optimal transition function s by (21).

4.2. Optimal event game

This subsection addresses the optimal event game of a given trace for a given process model with given attribute and transition functions. Let $PsM = (N, A, \varepsilon, \varphi, \sigma)$ be the process model and $t = e_1(s_1) e_2(s_2) \ldots e_m$ be the trace. Deriving an optimal event game consists in optimising the local model relevance of t, that is

$$f^{model}(\gamma^*, t) = \max_{\gamma} f^{model}(\gamma, t) \tag{22}$$

We propose a dynamic programming algorithm. Let $G_i(n)$ be the optimal local model relevance of the partial trace $t_i = e_1(s_1) e_2(s_2) \ldots e_i$ with event e_i being represented by node n. By this definition, $G_i(n) = 0$ for all $n \notin N(e_i)$, where $N(e_i) = \{n \in N : \varepsilon(t, i) \in \varepsilon(PsM, n)\}$ denotes the set of nodes having the same label of event e_i. Hence, we focus on nodes $n \in N(e_i)$. The recursive Bellman equation can be written as,

$$G_i(n) = \begin{cases} l_{node}(n, i), & \text{if } PRE(i, n) = \varnothing \\ \max\limits_{(i', n') \in PRE(i,n)} l_{node}(n, i) & \\ \quad + \alpha l_{arc}(n', n, i', i) & \\ \quad + G_{i'}(n'), & \text{otherwise,} \end{cases} \tag{23}$$

for all $n \in N(e_i)$, where $PRE(i,n)$ is the set of couples (i',n') indicating a preceding event $i' < i$ can be played by a node n' on lower layers $1,\ldots,layer(n)-1$, $l_{node}(n,i)$ is the local node relevance of playing event e_i at node n, and $l_{arc}(n',n,i',i)$ is the local relevance of arc (n',n) by playing event $e_{i'}$ at node n' and e_i at node n.

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

$$f^{model}(\gamma^*,t) = \begin{cases} 0 & \text{if } N(t) = \varnothing \\ \max\limits_{i\in[\![1,m]\!],n\in N(e_i)} G_i(n) & \text{otherwise} \end{cases} \tag{24}$$

Algorithm 2: Optimal event game of a trace $t = e_1(s_1)e_2(s_2)\ldots e_m$

Step 1. For $i = 1$ to m, apply (23) to compute the partial trace relevances $G_i(n)$ for all nodes n
Step 2. Apply (24) to determine the optimal local relevance, the last played event i^* and the node n^*
Step 3. Backtraking $G_{i^*}(n^*)$ by (23) to determine the other played events and their nodes.

4.3. Joint optimisation of event game and attribute/transition functions

This subsection addresses the joint optimisation of the event game and attribute/transition functions. The exact optimisation is too complex and we propose an iterative local improvement procedure for the joint optimisation. Consider first the case of a given event game. Section 4.1 applies to optimisation of the attribute/transition functions. With the new attribute/transition functions, the event game can further be optimised by the dynamic programming algorithm of Section 4.2. The above procedure then repeats until convergence is achieved. Similar local improvement can be performed by starting from given attribute/transition functions.

5. Process model optimisation

This section presents algorithms for process model optimisation. We first introduce two preliminaries: an extended concept of marginal model relevance introduced in our previous paper (Peng et al. 2024) and the solution repair. We then propose two solution algorithms: a multi-start local optimisation algorithm; and an ant colony algorithm.

5.1. Marginal relevance of a new node

This subsection evaluates the benefit of adding a new node n of macro-label $\widehat{b}$ to layer k. More specifically, we determine

$$\Delta(\widehat{b},k)$$

the optimal node relevance of the new node n without alternating the event game γ for all other existing nodes with the convention $\Delta(\widehat{b},k) = 0$ if node $(\widehat{b},k)$ exists in the current process model. Let $L(\widehat{b},k)$ be the set of traces t for which a new event can be represented by the new node n, i.e. there exists an event e_i of t that is not represented and $\varepsilon(t,i) \in \widehat{b}$, all preceding events are either not represented or played at lower layers, all following events are either not represented or played in higher layers. Let $\varepsilon(t,\widehat{b},k)$ be the label of the first such event e_i. As a result, each trace t of the set $L(\widehat{b},k)$ can be played by the new node n with event label $\varepsilon(t,\widehat{b},k)$.

If the attribute functions of node n are given, then $\Delta(\widehat{b},k)$ can be determined as for $F^{node}(\gamma,n)$. Otherwise, its attribute functions can be easily optimised as in Section 4.1. To summarise, the marginal relevance $\Delta(\widehat{b},k)$ not only provides the benefit of adding a new node but also its attribute functions $\varepsilon(\widehat{b},k,h)$.

5.2. Solution repair

This subsection addresses the repair of an infeasible process model PsM, i.e. with the violation of the maximal numbers of nodes or arcs. By construction, all process models of this paper meet the node size constraint and hence we limit ourselves to the repair of the arc constraint. The basic idea is to derive a feasible process model PsM' by removing $\max(\|A\| - UB^{arc}, 0)$ arcs of the least arc relevance. More specifically, we first determine the arc relevances of PsM, then sort the arcs in ascending order of their arc relevance, and then remove the first $\max(\|A\| - UB^{arc}, 0)$ arcs. The resulting process model PsM' is feasible and is our repaired process model. We then update the event game of the new process model PsM' by the algorithm of Section 4.1. Note that the attribute/transition functions are not updated for the sake of computational burden.

5.3. A multi-start local optimisation heuristic

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

LocalOpt starts with the random generation of an initial process model. It first generates randomly the nodes by scanning all combinations of macro-label and layer

Algorithm 3: LocalOpt-A multi-start local optimisation heuristic

Step 1. Random generation of an initial process model *PsM*
Step 2. Determine the maximal marginal relevance $\Delta(\widehat{b}^*, k^*)$. Note that the computation of $\Delta(\widehat{b}^*, k^*)$ comes with its optimal attribute value
Step 3. Add a node of macro-label $\widehat{b}^*$ at layer k^* and connect it to all existing nodes;
Step 4. Apply Algorithm 1 to update the event game and Algorithm 2 to update the attribute/transition functions;
Step 5. Repair the solution;
Step 6. Repeat 2- -5 till UB^{node} is reached;
Step 7. Repeat 1- -6 till a given computation time budget is reached.

$(\widehat{b}, k)$ with each being selected with some given probability p_{rand} as soon as the node size UB^{node} is not reached. This step repeats as soon the set of selected nodes is empty. It then adds an arc between any two nodes of different layers. For any node, each attribute h is assigned its root macro-attribute D^h. Each arc is associated with the transition function S. The random generation terminates with the event game optimisation of the resulting process model by the algorithm of Section 4.1 and the solution repair if needed.

LocalOpt then iteratively improves the process model by adding new nodes. At each iteration, it first evaluates the marginal relevance $\Delta(\widehat{b}, k)$. Let $\Delta(\widehat{b}^*, k^*)$ be the largest marginal relevance and $\varepsilon(\widehat{b}^*, k^*, h)$ its attribute value. If $\Delta(\widehat{b}^*, k^*) = 0$, the process stops. Otherwise, a node node $n^* = (\widehat{b}^*, k^*)$ with attribute values $\varepsilon(\widehat{b}^*, k^*, h)$ is added. An arc with transition function S is then added between n^* and any other nodes of different layers. The event game is updated by the algorithm of Section 4.1. The algorithm of Section 4.2 is used to update the transition functions. The event game is then updated again. LocalOpt continues adding new nodes as above as soon as the node size is not reached.

LocalOpt is a multi-start local optimisation heuristic. It restarts from a new initial solution as soon as the termination condition is not reached. The joint optimisation of Section 4.3 is applied to further improve the final process model.

5.4. Ant colony optimisation

This section presents an ACO (Ant colony optimisation) algorithm for process model optimisation (see Dorigo, Maniezzo, and Colorni 1996; Mohan and Baskaran 2012 for more details). Our ACO is a population-based metaheuristic with m ants.

Algorithm 4: ACO algorithm

Step 1. Initialization of the pheromone $pher(n, n')$ for all arcs (n, n') of the macro-label-layer network G_{aco} of nodes $(\widehat{b}, k)$
Step 2. Each ant randomly traverses the network G_{aco} and construct a process model
Step 3. Update the pheromones;
Step 4. Repeat 1- -3 till a given computation time budget is reached.

At each iteration, ants move on a network G_{aco} composed of $K\|\widehat{B}\|$ nodes $(\widehat{b}, k)$ plus a starting node at layer 0 with arcs from the starting node to all other nodes and from any node of layer $k > 0$ to any other node of a different layer $k' > 0$. Each ant moves randomly on G_{aco} along a random route of UB^{node} nodes, constructs a feasible process model composed of the same nodes and leaves pheromone trail on arcs of G_{aco} traversed by the ant.

The pheromones $pher(n, n')$ of the arcs of G_{aco} equal to a given constant τ initially is updated at the end of each iteration as follows. Each ant i generating a process model PsM^i leaves a pheromone equal to the model relevance of the PsM^i on each arc (n, n') of G_{aco} traversed. Let $deltapher(n, n')$ be the total pheromone leaved by all m ants in the current iteration. The pheromones $pher(n, n')$ is updated according to the usual rule:

$$pher(n, n') \leftarrow (1 - \rho_{aco}) pher(n, n') + \rho_{aco} \times q_{aco} \times deltapher(n, n') \quad (25)$$

where $\rho_{aco} \in [0, 1]$ and ρ_{aco} are parameters of the ACO. q_{aco} is determined according to one the four rules IB-update, BC-update, MMAS and Hyper-cube (HC) given in the electronic companion EC1 and suggested by the ACO literature.

We now address the construction of the process model. For each ant, the process model is built progressively by adding the new node reached by the ant and the construction terminates when a process model with UB^{node} is built or no improvement is achieved in $iter_{ant}$ consecutive iterations. Let $PsM^{current}$ be the current process model with the nodes and $n^{current} = (\widehat{b}^{current}, k^{current})$ the current node. We first evaluate $\Delta(\widehat{b}, k)$ the marginal relevance for all nodes not traversed by the ant. We then select randomly a new node $n = (\widehat{b}, k)$ with the following probability

$$C \times (pher(n^{current}, n))^{\alpha_{aco}} \times \Delta(\widehat{b}, k)^{\beta_{aco}} \quad (26)$$

where C is a normalising constant, $\alpha_{aco} > 0$ and $\beta_{aco} > 0$ are ACO parameters. A new process model $PsM^{candidate}$ is obtained by adding the new node n with attribute values related to its marginal relevance and connecting it to all nodes of different layers. The new model is evaluated by first updating its event game, updating its transition functions and then repairing if needed. If the new model is strictly better, i.e. $F^{model}(PsM^{candidate}) > F^{model}(PsM^{current})$, then (i) the ant moves to the new node n with process model $PsM^{candidate}$ and (ii) if $F^{model}(PsM^{candidate}) > p_{adj}F^{model}(PsM^{best})$ where PsM^{best} is the current best of the ACO, then the joint optimisation of the event game and attribute/transition function is performed and PsM^{best} is updated as needed. If the new model is not strictly better, the ant stays at the current node.

6. Numerical results

This section presents numerical results on randomly generating instances for two purposes: (i) to assess the performances of our proposed LocalOpt and ACO algorithms against benchmark values; and (ii) to numerically evaluate the benefits of introducing attributes.

6.1. Experimental settings

This subsection provides a brief presentation of the setting of the numerical experiment, the details of which are given in the electronic companion. All test instances are available upon request.

Random instance generation: In order to generate instances that are realistic enough, test instances are generated from process models of LocalOpt for the real case study. 12 process models are derived for 3 model sizes $UB^{node} \in \{10, 15, 20\}$ and $UB^{arc} = 2 \times UB^{node}$ for 4 groups of patients following strategy *str* 0, 1, 2, and 3 (see Section 7.1 for definitions of strategies).

All test instances share the same labels, attributes, transition features, label tree, attribute trees, precision weights as for the real case. Each instance *str*-UB^{node} corresponds to an event log of $N = 2000$ traces generated from the corresponding process model as follows. For each trace t, the starting event is generated by repeating the uniform sampling of a node n and random acceptance with probability $p_{start} = 0.9 \times 0.2^{k-1}$ where k is the layer of n. n is the terminating event with probability $p_{stop} = 0.9 \times 0.2^{K-k}$ or if n has no successors. Otherwise, it randomly moves to one of its successors with transition feature randomly selected from that of the arc. For each event e, its label and attributes are derived from that of its corresponding node n. The label of e is randomly sampled from the whole label set with probability $p_{mut} = 0.05$ and otherwise from the macro-label of n. For each attribute h of e, if h is an attribute of n, with probability $1 - p_{mut}$, its value is randomly sampled from the macro-value of the node. Otherwise, the value of h is randomly sampled from its entire domain.

Alternative algorithms: Our proposed LocalOpt and ACO algorithms are compared against the benchmarking ones including *Random*, *RG*, *Reinsert*, and *Relabeling* (see Appendix 1). For all algorithms starting with random initial solution, $p_{rand} = 0.05$ is used for selection of a given macro-label.

Algorithm parameter tuning: We use the following default parameters $K = 10$ layers, attribute weight $\lambda_1 = 0.4$ transition function weight $\lambda_2 = 0.4$, and arc relevance weight $\alpha = 0.8$. Each algorithm runs independently 10 times with a default computation time of 15 minutes, which is large enough time for the algorithms to converge, as shown in electronic companion EC2. ACO algorithm parameters are tuned in electronic companion EC1.

The justifications of the model parameters are as follows and sensitivity analysis will be performed for other settings. $K = 10$ as the length of longest event trace of the original dataset is 10. UB^{node} and UB^{arc} are selected in the ranges of the starting models used for instance generation. λ_1, λ_2 and α are set based on the following desires of the practitioners: matching nodes 4 times more important than matching arcs with $\alpha = 0.8$, matching label (arc) slightly more important than attributes (transition features) with $\lambda_1 = 0.4$ ($\lambda_2 = 0.4$), precision weight varying linearly from 0.1 for root node to 1 for leaves.

Performance indicators: For each algorithm and each instance, we compute the best, the worst and the average model relevance F^{model}. We also determine a normalised quality measure denoted by *ir* and called *information ratio* and defined as follows:

$$ir = F^{model}/Bound1 \tag{27}$$

where $Bound1$ is the model reference if all traces were perfectly played and hence can be considered as the total amount of information of the event log and *ir* represents then the percentage of the event log information captured by the process model. For a precise spaghetti-like process model, we achieve faithful representation of all traces with $ir = 1$.

6.2. Comparison of algorithms

This subsection compares our proposed LocalOpt and ACO algorithms against benchmarking algorithms with 10 runs for each. Table 2 gives the best run information ratio. Detailed results are given in electronic companion EC3, including the model relevances of worst, best, and

Table 2. Best run information ratio of all algorithms.

ins	LocalOpt	ACO	Rand	RG	Reinsert	Relabel
0–10	0.90	**0.91**	0.82	0.74	0.85	0.89
	0.92	**0.93**	0.87	0.83	0.90	0.92
	0.93	**0.93**	0.90	0.87	0.88	0.93
0–15	0.89	**0.89**	0.76	0.76	0.86	0.87
	0.92	**0.92**	0.88	0.81	0.88	0.91
	0.93	0.92	0.88	0.87	0.89	0.92
0–20	**0.79**	0.79	0.74	0.69	0.75	0.77
	0.90	**0.90**	0.80	0.76	0.85	0.87
	0.92	**0.93**	0.82	0.78	0.83	0.91
1–10	0.90	**0.91**	0.81	0.75	0.87	0.89
	0.91	**0.91**	0.87	0.82	0.91	0.91
	0.92	**0.92**	0.89	0.85	0.90	0.91
1–15	0.86	**0.87**	0.82	0.77	0.83	0.86
	0.91	**0.91**	0.86	0.82	0.87	0.90
	0.92	**0.92**	0.89	0.89	0.88	0.91
1–20	**0.81**	**0.81**	0.75	0.72	0.80	0.81
	0.87	**0.87**	0.82	0.77	0.84	0.85
	0.90	0.90	0.85	0.81	0.85	0.88
2–10	0.92	**0.93**	0.82	0.77	0.86	0.90
	0.93	**0.94**	0.89	0.87	0.91	0.93
	0.94	**0.94**	0.90	0.91	0.90	0.93
2–15	0.81	**0.83**	0.74	0.69	0.79	0.81
	0.91	**0.92**	0.79	0.77	0.86	0.88
	0.92	**0.93**	0.84	0.80	0.81	0.90
2–20	0.80	**0.81**	0.74	0.72	0.79	0.79
	0.89	0.88	0.80	0.77	0.86	0.87
	0.93	0.93	0.86	0.85	0.86	0.90
3–10	**0.92**	**0.92**	0.84	0.81	0.87	0.91
	0.93	**0.93**	0.89	0.84	0.89	0.93
	0.93	**0.93**	0.90	0.90	0.91	0.93
3–15	0.90	**0.91**	0.81	0.78	0.86	0.89
	0.92	**0.93**	0.88	0.81	0.88	0.92
	0.93	**0.93**	0.90	0.86	0.90	0.92
3–20	0.85	0.85	0.78	0.77	0.85	**0.85**
	0.89	**0.90**	0.85	0.81	0.88	0.89
	0.91	0.91	0.88	0.86	0.88	0.90
avg ir	0.898	0.902	0.837	0.803	0.861	0.888
% best	41.67	77.78	0	0	0	2.78

average run model plus the best information ratio for all algorithms, and a figure on model relevance ranges for LocalOpt and ACO for three instances *str*-20 generated from a model of size 20.

- The overall ranking from the best to the worst is: ACO-LocalOpt, Relabeling, Reinsert, Random, RG. Except for *ins*-3-20 where Relabeling generates a slight better best run for size 10, all other best results are achieved by either ACO or LocalOpt. The superior performance of Relabeling/Reinsert over Random/RG is achieved thanks to the use of certain theoretical results arising from this paper.
- The average best-run information ratio is as follows: ACO (90,2%), LocalOpt (89.8%), Relabeling (88.8%), Reinsert (86.1%), Random (83.7%), RG (80.3%). ACO captures 10% more information of the event log than RG.
- ACO is slightly better and more robust than LocalOpt. ACO produces more best best-runs, 77.78% versus 41.67%. The better robustness is shown by the 88.99% best worst runs for ACO and 16.67% for LocalOpt. The slight superiority and better robustness of ACO is further supported by the model relevance distribution.

6.3. Benefits of attributes

The goal of this subsection is to measure quantitatively the benefits of introducing attributes in process mining. All results of this subsection are obtained by ACO for three runs of 8h to insure against any bias that might arise as a result of algorithm convergence. Results for LocalOpt are similar and omitted.

To measure the benefits of attributes, we consider the following four models:

- PsM^{old} is the process model given by ACO without taking into account attributes and without macro-labels. The model relevance is evaluated with the same event game and with the root value for each attribute. Note that PsM^{old} corresponds to the model of our previous paper Peng et al. (2024);
- PsM^{att} is the process model PsM^{old} but with the final event game and attributes/transition functions jointly optimised;
- PsM^{new} is the process model determined by ACO of this paper;
- PsM^{Exp} is the process model determined by ACO of a modified event log without attributes and macro-labels. The initial event log is modified by associating each properly valued label an expanded label, i.e. two expanded labels differ either by label or by at least one attribute value.

The conversion to expanded labels in PsM^{Exp} allows direct application of our previous methods as described in Peng et al. (2024). The price to pay is the huge number of event labels, and hence the impossibility of a process model of reasonable size being able to capture enough events. As a result, the resulting model relevance is likely to be small. For this reason, our LocalOpt and ACO algorithms do not apply due to the huge number of event labels and instead PsM^{Exp} is evaluated by the upper bound $Bound2$ defined in Section 3.4.2.

We also determine the following gaps:

- $Gap^{att} = PsM^{att}/PsM^{old} - 1$ to measure the benefits of introducing attributes without modifying the process model structure;
- $Gap^{new} = PsM^{new}/PsM^{old} - 1$ to measure the benefits of process model optimisation taking into account attributes;
- $Gap^{Exp} = Bound2^{Exp}/PsM^{old} - 1$ to measure the effects of expanded labels;

Table 3. Model relevance of different attribute models with *ir* in parentheses.

	PsM^{old}	PsM^{att}	PsM^{att_bis}	PsM^{new}	PsM^{Exp}
avg	4.78 (0.64)	6.12 (0.82)	6.11 (0.82)	6.19 (0.83)	2.40 (0.32)
% best	0	22.22	16.67	97.22	0

- $Gap^{att,new} = PsM^{new}/PsM^{att} - 1$ to measure the additional benefits of process model structure optimisation taking into account attributes.

Table 3 gives the average model relevance, information ratio and percentage of the best runs for all instances with $\lambda_1 = 0.8$ and Figure 4 gives the corresponding gaps. Detailed results for both best runs and average runs for both $\lambda_1 = 0.8$ and $\lambda_1 = 0.4$ are given in electronic companion EC4. Results for the best runs and for other values of λ are similar and hence omitted. The following observations are made:

- Gap^{att} shows that introduction of attributes significantly improves the process model and the information ratio *ir* shows 18% more event log information captured. For the test instances, optimising attributes alone without alternating the event games PsM^{att_bis} brings nearly the same improvement and alternating the event games PsM^{att} brings further but moderate improvement. The latter is likely to be due to the real case-derived test instances for which many features have unique dominating value (feature 'age' is equal to 'adult' for over 95% of traces). For general event logs without dominating values for features, we conjecture that the introduction of attributes significantly alters how traces traverse the process model and joint optimisation of attributes. The event game is needed to full explore the benefits of the attributes;
- Gap^{new} and $Gap^{att,new}$ show further improvement of jointly optimising the process model structure. PsM^{new} achieves 97% of the best best runs in contrast to the 22% for PsM^{att}. The moderate 1.1% improvement of PsM^{new} over PsM^{att} is likely related to the dominating values of many features. Note that PsM^{new} achieves 100% of the best average runs;
- Gap^{Exp} is inadequate, confirming our conjecture that expanding the event labels to account for attributes is not a good solution due to the huge number of expanded event labels. An average information ratio of only 32% is achieved thanks to two event labels without attributes appearing in most traces; and the amount of event log information captured by truly expanded event labels is very small.

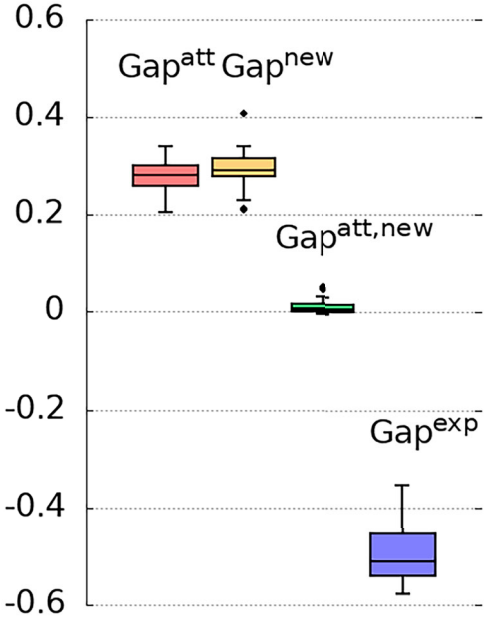


Figure 4. Percentage improvement thanks to event attributes.

7. Application to cancer care pathways

In this section, the ACO solution algorithm is applied to discovering care pathway models from a dataset of sarcoma patients. We first present the dataset, the label tree and attribute trees, and finally the numerical findings with a special focus on the benefits of attributes on care pathway discovery. It is worth noting that a recent literature review by Kusuma et al. (2021) identified only four direct applications of process mining to disease trajectory modelling and highlighted a lack of awareness of these methods.

Sarcomas are a large family of rare tumours that affect men and women at all ages. We extracted all data from the French database NETSARC regarding patients diagnosed with sarcoma in 2013 who underwent surgery for their primary tumour. The total of 2203 patients were treated according to four care management strategies: (1) complete initial management in the network with a sarcoma MDTB before/after the initial surgery ($n = 1068$); (2) outside initial management with a sarcoma MDTB before the initial surgery ($n = 108$); (3) similar to 2 but with a sarcoma MDTB after the initial surgery ($n = 750$); and (4) outside initial management and no sarcoma MDTB ($n = 277$). Strategy 0 denotes all patients. While both the quality of the data and missing data present significant challenges to any real case studies, these are only minor concerns in our methodology-oriented paper. Data quality is not an issue as the dataset has been used and corrected in multiple clinical studies (e.g. Blay et al. 2019). Missing data are assigned a specific value (−) in this paper. Fortunately, they do not appear in all of our process models, implying that their impact on the care pathways is only minor.

Table 4. Attributes and labels of the case study.

	type	age	size	depth	site	quality	location	requesting center
OD								
RCP	■						■	■
Chir	■	■					■	
RChir	■	■				■		
TTT	■	■	■	■		■		
Bio			■	■	■			■
Last								

7.1. Label and attribute representation

This subsection presents the formal representation of the sarcoma care pathways and related weight for process mining. They are based on discussions with medical experts.

In this study, sarcoma care pathways are characterised by

- Transition features: S = progression (*pro*), no progression (*np*) corresponding to the cancer progression state;
- Event labels: original diagnosis of sarcoma (*OD*), sarcoma multidisciplinary tumour boards (*RCP*), biopsy before surgery (*Bio*), surgery (*Chir*), second surgical excision/Re-excision (*RChir*), neoadjuvant/adjuvant treatment (*TTT*), last contact (*Last*);
- Label attributes: histological subtypes (*type*), *age*, tumour size (*size*), tumour depth (*depth*), tumour location (*site*), surgery quality (*quality*), location for RCP or Chir (*location*), *requestingcentre* for RCP (see Appendix 2 for explanations). Attributes associated with labels are in Table 4. For event label, only attributes that can explain the occurrence of the event are considered;
- Label tree: the label tree with a single macro-label '*all*' is given in Figure 5. The precision weight is 0.1 for *all* and 1 for others;
- Attribute trees: attribute trees are given in Figure 3 for *type* and in Table 5 for other attributes with for each node macro-value, precision weight, and number of immediate descendants and with detailed leaf nodes in the electronic companion. The tree can be unbalanced with different numbers of levels due to missing detailed data or no need of further decomposition of nodes;

From the statistics given in electronic companion EC5, liposarcomas accounted for 17% of histological subtypes, followed by leiomyosarcomas (14%). The mean tumour size was 91.85 mm (SD 76.11). Three quarters of tumours were deep-seated, with soft tissue the most common site (80%). 42% of resections had R0 negative histological margins (57% in strategy 1 versus 18% in strategy 3). For ease of reading, attribute *location* and *requesting centre* have three macro-values (all, inside, outside) for inside/outside the NETSARC network; and surgery *quality* has three base values (R0, R1, and R2 from best to worst). The other attribute values are straightforward (Figure 5).

Table 5. Attribute trees.

Attribute	Root	Level-1
loc RCP	all/0.1/2	inside/0.3/71 outside/0.3/0
loc Chir	all/0.1/3	inside/0.3/32 outside/0.3/0 –/0.3/0
req centre	all/0.1/3	inside/0.3/44 outside/0.3/138 –/0.3/0
age	all/0.1/3	child/1/0 adult/1/0 –/1/0
depth	all/0.1/4	superficial/1/0 deep/1/0 deep+superficial/1/0 –/1/0
quality	all/0.1/4	R0/1/0 R1/1/0 R2/1/0 –/1/0
size	all/0.1/4	small/1/0 median/1/0 large/1/0 –/1/0
site	all/0.1/4	superior/0.3/16 inferior/0.3/14 trunk/0.3/15 others/0.3/45

7.2. Numerical findings

This subsection describes relevant care pathway features discovered by our new process mining approach with event label attributes. More specifically, the following issues will be addressed: (i) benefits of attributes; and (ii)

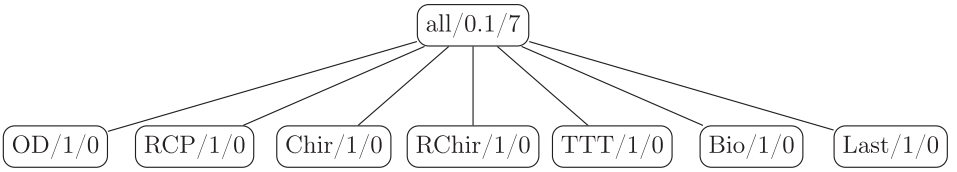


Figure 5. The label tree of the case study.

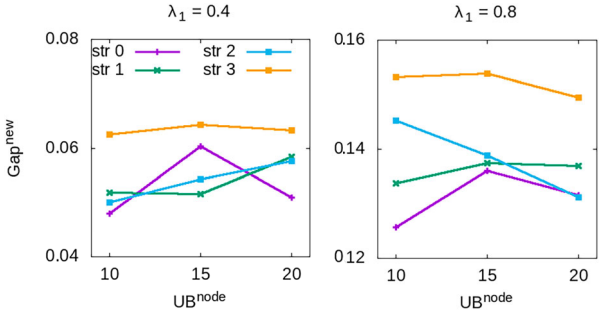


Figure 6. Percentage relevance improvement versus model size.

impact of model size, attribute weight λ_1, and attribute precision weight. Only partial numerical results are given in the main paper to show the key numerical findings and the complete numerical results are given in a companion online document. Note that this subsection aims to provide evidence of the usefulness of label attributes and we do not provide a complete sensitivity analysis. Further, the clinical study using the proposed approach is beyond the scope of this paper.

Model relevance improvement: Figure 6 compares model relevance of process model PsM^{old} without attributes and process model PsM^{new} with attributes. As in Section 6, *the introduction of attributes improves the model relevance* by 4–6% for small attribute weight $\lambda_1 = 0.4$ and by 12–16% for large attribute weight $\lambda_1 = 0.8$.

Process models with attributes vs without attributes: Figure 7 gives the process models with attributes for $UB^{node} = 10$ nodes and the corresponding models without attributes are given in electronic companion EC7. The following observations could be observed:

- Impact on the process model: the introduction of attributes has a more significant impact on the arcs than on the nodes of the process model. For example, with $UB^{node} = 10$ nodes, all strategies except *str*-2 have the same nodes. For *str*-2, a node of macro-label *all*, which is forbidden without attributes, is used in the process model with attributes to capture the highly diverse care pathways. For all process models with or without attributes, cancer progression highlighted in red appears on the right and increases as the care pathways advance from left to the right;
- *Value of attributes*: (i) confirmation of the strategy with surgery *Chir* done *inside* in *str*-1 and *outside* in *str*-2 and 3; (ii) new information of potential clinical values: *TTT* treatment by far for *deep* tumour, second event *TTT* in *str*-0 (*TTT2* named rightward) significantly due to lower surgery quality *R1*, important second surgery *RChir* in *str*-3 (250/741) by far for lower quality *R1* surgery, longer post-*Chir* pathways in *str*-1 for larger tumour size (*TTT2*) than smaller tumour size (*TTT1*).

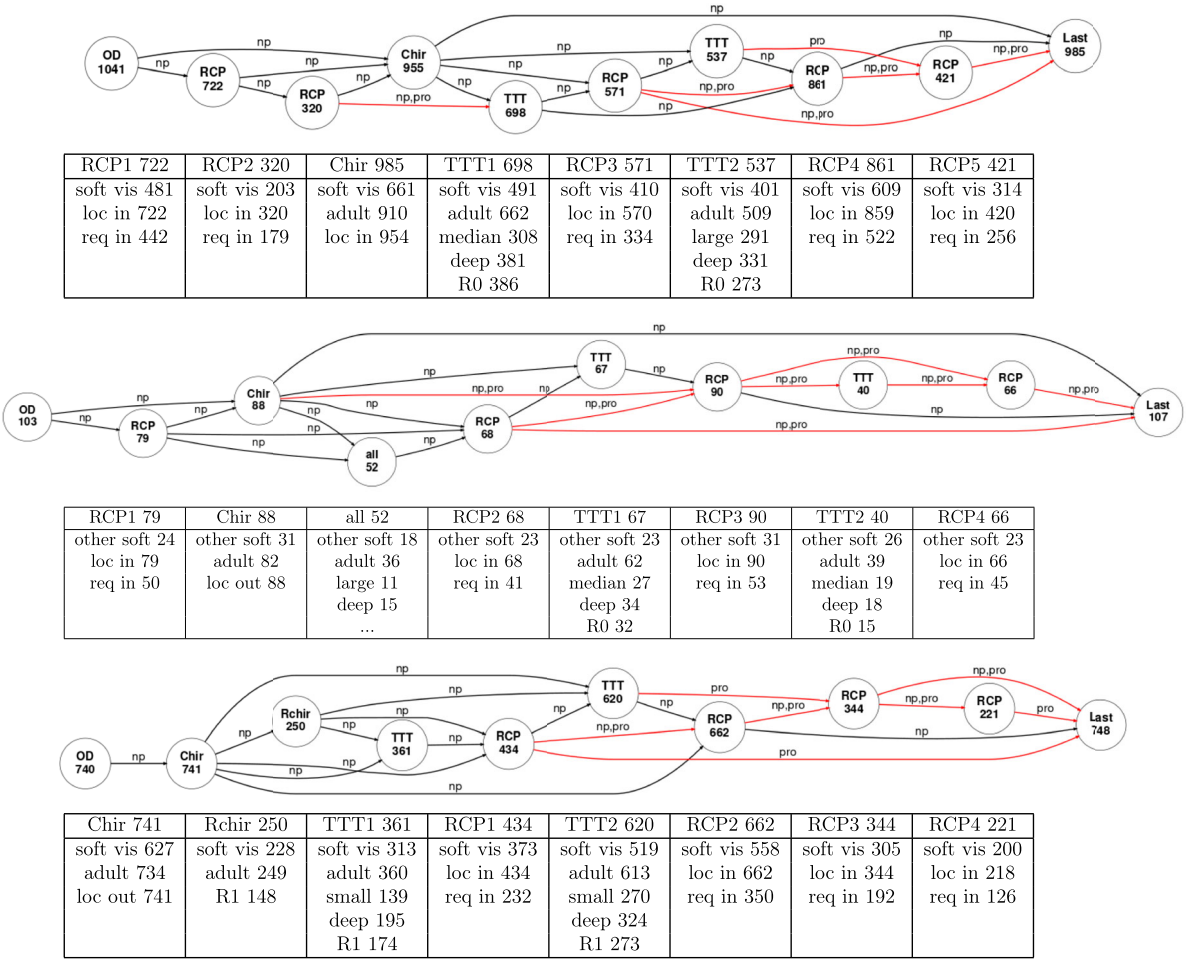
Increasing model size (Figure 7 *versus* Figure 8): (i) more care events including less frequent *RChir* and *Bio* events and duplicated *Chir* events to better represent different sub-populations/strategies; (ii) less under macro-label (*all* in *str*-2 now split); (iii) *richer value of attributes*: *Bio* in *str*-1 and 2 for *large* and *deep* tumours, second surgery *RChir* in both *str*-2 and 3 due to lower surgery quality *R1* and with similar ratio of *RChir*/*Chir* (ratio significantly lower in *str*-1).

Impact of increasing attribute weight λ_1: using the sensitivity results and process models provided in electronic companion EC9, increasing λ_1 makes the attribute value matching more important, leading to decreasing model relevance and increasing usage of macro-labels (*all* in this case) to improve the number of attribute value matching for the corresponding nodes.

Impact of attribute precision weight with sensitivity results and process models in electronic companion EC10: (i) decreasing the weight of layer-2 macro-value *inside* and *outside* from 0.3 to 0.1 for all location attributes, location value gives less precise *all* macro-label and the model relevance decreases slightly; (ii) increasing the weight of layer-3 macro-values for *Leiosarcoma*, etc., from 0.6 to 1 for attribute *type*, *type* value goes from level-2 value *softvisceral* to more precise level-3 value *other softvisceral* and the model relevance slightly increases.

7.3. Comparison with data-aware process miners

In this section, we compare our approach with two prominent process miners that incorporate the data perspective: the declarative method developed by Leno et al. (2020), known for its strong ability to discover attribute-dependent constraints; and the procedural approach of Mannhardt et al. (2017), which is



RCP1 722	RCP2 320	Chir 985	TTT1 698	RCP3 571	TTT2 537	RCP4 861	RCP5 421
soft vis 481 loc in 722 req in 442	soft vis 203 loc in 320 req in 179	soft vis 661 adult 910 loc in 954	soft vis 491 adult 662 median 308 deep 381 R0 386	soft vis 410 loc in 570 req in 334	soft vis 401 adult 509 large 291 deep 331 R0 273	soft vis 609 loc in 859 req in 522	soft vis 314 loc in 420 req in 256

RCP1 79	Chir 88	all 52	RCP2 68	TTT1 67	RCP3 90	TTT2 40	RCP4 66
other soft 24 loc in 79 req in 50	other soft 31 adult 82 loc out 88	other soft 18 adult 36 large 11 deep 15 ...	other soft 23 loc in 68 req in 41	other soft 23 adult 62 median 27 deep 34 R0 32	other soft 31 loc in 90 req in 53	other soft 26 adult 39 median 19 deep 18 R0 15	other soft 23 loc in 66 req in 45

Chir 741	Rchir 250	TTT1 361	RCP1 434	TTT2 620	RCP2 662	RCP3 344	RCP4 221
soft vis 627 adult 734 loc out 741	soft vis 228 adult 249 R1 148	soft vis 313 adult 360 small 139 deep 195 R1 174	soft vis 373 loc in 434 req in 232	soft vis 519 adult 613 small 270 deep 324 R1 273	soft vis 558 loc in 662 req in 350	soft vis 305 loc in 344 req in 192	soft vis 200 loc in 218 req in 126

Figure 7. Process models of 10 nodes of the case study for strategies 1–3.

implemented in the interactive Data-aware Heuristics Miner (iDHM) (Mannhardt, De Leoni, and Reijers 2017).

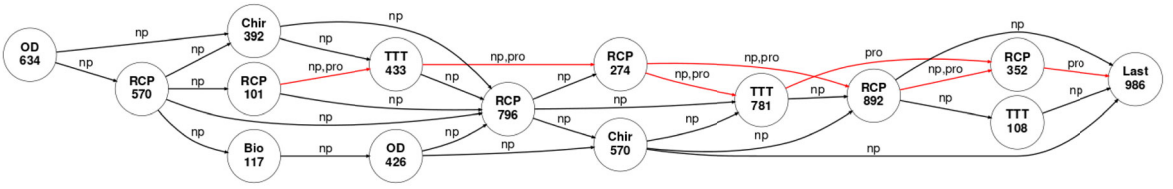
The three methods differ in their nature, semantics, and representation of process models. However, of the methods presented in the literature, the selected benchmark methods are the closest to ours in terms of integrating the event log data attributes in constructing the control-flow perspective of process models. All process model miners are tested on the event log of strategy 1 patients in our sarcoma case study.

We begin by running the iDHM method, which requires four thresholds $(\theta_{obs}, \theta_{dep}, \theta_{bin}, \theta_{cond})$ (refer to Mannhardt et al. 2017 for more details). We set $\theta_{obs} = 1$ and $\theta_{cond} = 0$ to favour the discovery of conditional dependency relations over standard ones. In the plugin options, we select all event log attributes, check the 'all tasks connected' option, and choose the algorithms Flexible Heuristic Miner, C4.5 (Cohen's Kappa), and C4.5 (F1-score) for dependency, conditional, and decision heuristics, respectively. By setting $\theta_{dep} = 0.75$ and $\theta_{bin} = 0.1$, we obtain the causal net shown in Figure 9(a). Three conditional relations are discovered, highlighted in red, and listed in the subsequent table. This result represents the maximum number of conditional relations achieved through fine-tuning the thresholds.

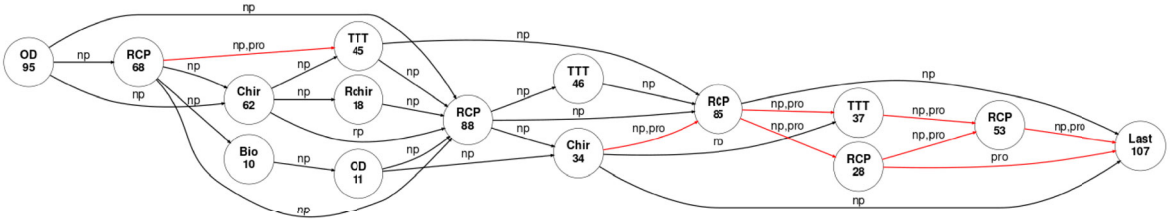
We then apply the declarative method with *response* input rules (A, B, Cond) implying that if A occurs and Cond holds, B must occur afterward. We use the K-Medoids + RIPPER (clustering activation + target payloads) version with $K = 2$ as we find in Leno et al. (2020)'s algorithm to derive the conditions.

To ensure a fair comparison, the declarative method focuses on rules corresponding to transitions in Figure 9(a), i.e. conditions of (*RCP*, *Bio*), (*RCP*, *Last*), etc. We run our method by setting $UB^{arc} = 9$, $UB^{node} = 7$, and $K = 6$. Results are given in Figure 9(b,c), respectively. The following observations can be made:

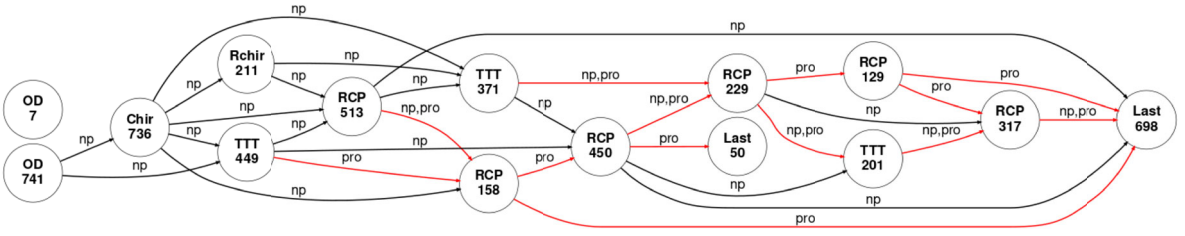
- Procedural vs declarative: comparing iDHM and our approach with MP-Declare, procedure methods (especially ours) offer a better understanding of the care pathways, i.e. the evolution over time of the patient's care. MP-Declare, however, sheds only scant light on the same question;



RCP1 570	Chir1 392	RCP2 101	Bio 117	TTT1 433	RCP3 796	RCP4 274	Chir2 570	TTT2 781	RCP5 892	RCP6 352	TTT3 108
soft vis 378 loc in 570 req in 325	soft vis 291 adult 383 loc in 391	soft vis 65 loc in 101 req in 54	large 61 deep 72 site thigh 44 req in 54	soft vis 317 adult 410 median 173 deep 248 R0 227	soft vis 546 loc in 795 req in 491	soft vis 204 loc in 273 req in 168	soft vis 381 adult 534 loc in 570	soft vis 554 adult 739 large 303 deep 446 R0 424	soft vis 629 loc in 890 req in 536	soft vis 264 loc in 351 req in 229	soft vis 83 adult 102 median 42 deep 65 R0 56



RCP1 68	Chir1 62	Bio 10	TTT1 45	Rchir 18	RCP2 88	TTT2 46	Chir2 34	RCP3 85	TTT3 37	RCP4 28	RCP5 53
other soft 20 loc in 68 req in 40	other soft 25 adult 59 loc out 62	large 6 deep 7 site Retrope- ritoneum 4 CHU Nancy 2	other soft 17 adult 41 median 19 deep 26 R0 20	other soft 8 adult 18 R1 7	other soft 27 loc in 88 req in 55	soft vis 30 adult 44 median 19 deep 24 R0 19	Ewing 6 adult 28 loc out 34	other soft 30 loc in 85 req in 52	other soft 13 adult 35 median 16 deep 15 R0 14	other soft 11 loc in 28 req in 19	other soft 18 loc in 53 req in 35



Chir1 736	Rchir 211	TTT1 449	RCP1 513	TTT2 371	RCP2 158	RCP3 450	RCP4 229	RCP5 129	TTT3 201	RCP6 317
soft vis 623 adult 729 loc out 736	soft vis 190 adult 210 R1 142	soft vis 368 adult 443 small 170 deep 277 R1 149	soft vis 430 loc in 513 req in 259	soft vis 321 adult 370 small 172 deep 178 R1 195	soft vis 136 loc in 158 req in 84	soft vis 391 loc in 450 req in 256	soft vis 208 loc in 229 req in 127	soft vis 115 loc in 128 req in 77	soft vis 175 adult 200 small 88 deep 103 R1 113	soft vis 281 loc in 316 req in 170

Figure 8. Process models of 15 nodes of the case study for strategies 1–3.

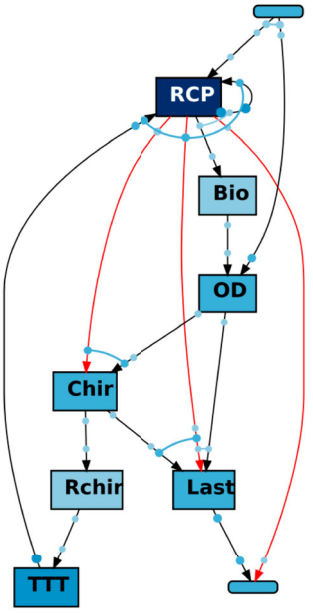
- Benefit of event duplication: grouping all RCP events in the same node makes the iDHM model poorly informative of the sarcoma care pathways. Conversely, our method clearly explains the hidden care pathway: all start by *OD*, *Chir* event follows *OD* or *OD-RCP*, *TTT* follows ether *Chir* or *RCP*, cancer progress (*pro*) correlated with the length of care pathway, dominant care pathway *OD-RCP-Chir-TTT-RCP-Last*;
- Benefit of data-awareness: while the three conditional relations of iDHM do not provide much information, the attribute values of our model provide significant details, such as: (i) *Chir* and *RCP* nearly all with *location* = *inside*; and (ii) *TTT* events with dominant attributes *depth* = *deep*, *quality* = *R*0, *size* = *large*. While (i) demonstrates the compliance with recommendation of strategy 1, (ii) reveals interesting clinical information. The benefit of the data-awareness central to our approach in comparison with iDHM can be attributed to event duplication and our optimisation framework. When compared with our approach, MP-Declare allows more general conditions, such as *quality*! = *R*2, and provides both support and confidence. By contrast, our approach provides only one macro-attribute value and support. Unfortunately, MP-Declare's potential is limited by the one-label-one-node restriction.

In summary, our approach allows us to show how care event repetition and data impact the sarcoma care pathways that other data-aware miners fail.

8. Conclusion

This paper proposes a novel formal optimisation framework for process mining with two specific features: event duplication and data-awareness, both of which are vitally

(a) iDHM



Source	Target	Condition
RCP	Chir	patient_status = np
RCP	Last	patient_status = pro
RCP	End	age = -

(b) MP-Declare

Number	Activation/Target	Activation/Target condition	Support	Confidence
C_1	RCP Bio	site != Retroperitoneum	0.83	0.83
C_2	**RCP** **Last**	**patient_status = pro**	**0.61**	**0.61**
C_3	**RCP** **Chir**	**patient_status = np and age = adult**	**0.9**	**0.9**
C_4	RChir TTT	quality != R2 margin	0.73	1
C_5	TTT RCP	patient_status = np	0.85	0.85
C_6	Bio OD	patient_status = np	0.99	0.99
C_7	OD Chir	patient_status = np and age = adult	0.92	0.92
C_8	Chir Last	patient_status = np	0.59	0.59
C_9	Chir Rchir	atient_status = np and age = adult	0.94	0.94

(c) Our

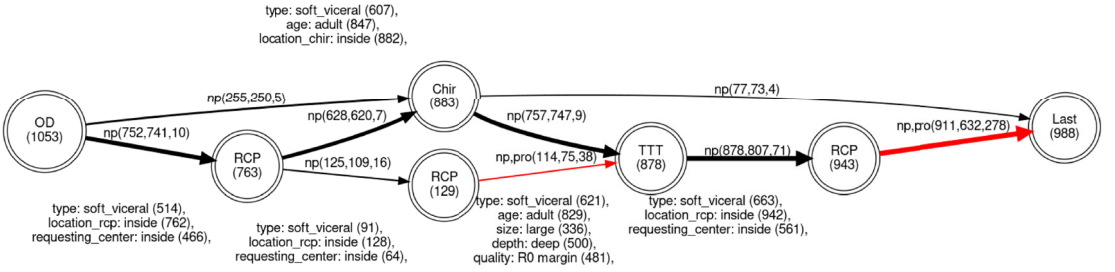


Figure 9. Comparison of data-aware process miners on strategy 1 event log. (a) iDHM. (b) MP-Declare and (c) Our.

important in cancer care pathway modelling and yet barely addressed by the existing literature. Hierarchical representation of event attribute values allows us to limit the additional complexity of data-awareness and to set the process model node data at the appropriate precision level. Event duplication is addressed by a dynamic programming algorithm for optimal replay of any event trace. The optimisation framework allows the optimal

attribute value setting of process model nodes and upper bounds of overall quality measures. We propose a multi-start local optimisation algorithm and an ant colony optimisation algorithm (ACO) as means to optimise our process model.

From numerical results on randomly generated instances, we can see that data-awareness leads to an increase of 18% of represented information. When applied to a sarcoma care pathway case study, this model demonstrates that data-awareness adds rich information concerning clinical values and confirms the strategies used to generate the event logs. We also show that our approach clearly illustrates the impact of care event repetition and data on the pathways.

The practical implication of this methodology-oriented paper will be explored in follow-up clinical studies of sarcoma care pathways. The process mining approach outlined in this paper will be applied in companion clinical papers to show the impacts on the care pathways of various attributes (soft tissue vs bone, adult vs elderly, R0 vs other surgery quality, tumour size/depth, etc.). The clinical studies should be designed and results consolidated by medical experts. Furthermore, a decision-aid tool is needed to help practitioners with setting the parameters of relevant perspectives of the process models.

Future research can be pursue in multiple directions:

(1) *General data representation in process models*: one limitation of our approach is the selection of one value per attribute for each node whereas MP-Declare allows for a more general condition, such as $quality! = R2$. Multiple extensions are possible: (i) multiple values per attribute; (ii) a logical constraint in terms of attributes per node and per arc. Both significantly increase the solution space, meaning that the approach outlined in this paper would not apply. Combining combinatorial optimisation and machine learning techniques may, however, limit the solution space;
(2) *Alternative quality measures*: another limitation is that our *relevance* score relies purely on positive experience of a trace, i.e. whether it passes a node and whether its attribute values match those of the node. It is reasonable to include negative quality measures to account for the relevant nodes not visited and attribute value mismatches. Classical machine learning measures of false positive and false negative could be explored;
(3) *Near-optimal event game*: whereas the exact event game optimisation by dynamic programming is reasonable for relatively small event logs, it does not scale up for large event logs. Combination with machine-learning methods could also provide means to swiftly estimate the event game and related relevance measures;
(4) *Process model optimisation for large event logs*: process model optimisation with general data representation and general quality measures for large event logs faces the challenges of a combinatorially larger solution space and longer evaluation time for each. Novel optimisation algorithms are needed;
(5) *Process model optimisation with missing data*: despite the quality dataset of this paper, missing data are not uncommon in practical datasets, especially healthcare datasets. The application of our approach requires innovative techniques for event game optimisation and relevance score evaluation.

Acknowledgments

The authors would like to thank the clinical network for sarcomas (NETSARC+) supported by the French National Cancer Institute (INCa), in particular Prof. Jean-Yves BLAY (Centre Léon Bérard, Lyon). The authors would also like to thank anonymous reviewers for their helpful comments and feedback.

Data availability statement

All test instances are available upon request.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the French Ministry of Health (Ministère des Affaires Sociales et de la Santé) [Programme de Recherche Médico Economique-PRME-18-0162] and the National Natural Science Foundation of China [grant number 72192822].

Notes on contributors

Omar Rifki is an assistant professor of computer science at the University of Littoral Opal Coast. He earned his Ph.D. in Economics from Kyushu University, where he focussed on the notion of robustness in optimisation. His current research interests include optimisation in the transportation and healthcare sectors, particularly in logistics and process mining. He also explores models of propagation and decision-making.

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 Institut Henri FAYOL, Department of Mathematical and Industrial Engineering, Mines Saint-Etienne, Saint-Étienne, France. His research interests include in applying

optimisation tools to industrial problems, such as logistics, manufacturing and healthcare.

Lionel Perrier received the Ph.D. degree in economics and the Habilitation à 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 also with the GATE UMR 5824, CNRS, Université Lumière Lyon 2, Université Jean Monnet Saint-Etienne, emlyon business school. L. Perrier, PhD, HDR, is a member of the selection committees for grants Programme de Recherche Médico-Economique (PRME) & Programme Hospitalier de Recherche Clinique National (PHRC-N), French Ministry of Health, and a member of the Board Organisation of European Cancer Institutes (OECI) Cancer Economics Working Group. He was a member of the Economic and Public Health Evaluation Committee (CEESP), French National Authority for Health (HAS), from 2015 to 2024. As a scientific coordinator, a work package leader or scientist, he is in charge of the health economics part of numerous national and international projects, and solicited for teaching in this field amongst other in Emlyon Business School and Centrale Lyon.

Xiaolan Xie is a professor of industrial engineering at the École des Mines de Saint Etienne. His research interests include healthcare system engineering, optimisation and data analytics. He is author/coauthor of 350+ publications including over 130+ journal articles and six books. He has been PI/co-PI for various collaborative research at national level (ANR-TECSAN HOST, ANR HEASY-PLAT, PREPS-eSI), international level (NSF China, FP6-IST6 IWARD, FP6-NoE I∗PROMS) and industrial level (7 CIFREs with Heva, Lomaco, Aesio, Cetaf, CH-Chalon). Prof. Xie is a fellow of IEEE. He was the founding chair of the Technical Committee on Automation in Health Care Management of the IEEE Robotics & Automation Society. He has been editor/associate editor for various international journals (IEEE TASE, IEEE TAC, IEEE TRA, IJPR) and special issue guest editor. He was general chair of the 2021 IEEE Conference on Automation Science and Engineering.

ORCID

Omar Rifki http://orcid.org/0000-0002-6074-9131

References

Alman, Anti, Fabrizio Maria Maggi, Marco Montali, Fabio Patrizi, and Andrey Rivkin. 2022, June 6–10. "Multi-Model Monitoring Framework for Hybrid Process Specifications." In *Advanced Information Systems Engineering - 34th International Conference, CAiSE 2022, Proceedings*, 319–335. Leuven: Springer.

Augusto, Adriano, Raffaele Conforti, Marlon Dumas, Marcello La Rosa, Fabrizio Maria Maggi, Andrea Marrella, Massimo Mecella, and Allar Soo. 2018. "Automated Discovery of Process Models from Event Logs: Review and Benchmark." *IEEE Transactions on Knowledge and Data Engineering* 31 (4): 686–705. https://doi.org/10.1109/TKDE.2018.2841877.

Bacon, Andrew, Kwok Wong, Malee S. Fernando, Brian Rous, Roger J. W. Hill, Shane D. Collins, John Broggio, and Sandra J. Strauss. 2023. "Incidence and Survival of Soft Tissue Sarcoma in England between 2013 and 2017, an Analysis from the National Cancer Registration and Analysis Service." *International Journal of Cancer* 152 (9): 1789–1803. https://doi.org/10.1002/ijc.v152.9.

Bano, Dorina, Francesca Zerbato, Barbara Weber, and Mathias Weske. 2021. "Enhancing Discovered Process Models with Data Object Lifecycles." In *2021 IEEE 25th International Enterprise Distributed Object Computing Conference (EDOC)*, 124–133. IEEE.

Batoulis, Kimon, Andreas Meyer, Ekaterina Bazhenova, Gero Decker, and Mathias Weske. 2015, June 8–12. "Extracting Decision Logic from Process Models." In *Advanced Information Systems Engineering: 27th International Conference, CAiSE 2015, Proceedings 27*, 349–366. Springer International Publishing.

Bergami, Giacomo, Fabrizio Maria Maggi, Andrea Marrella, and Marco Montali. 2021. "Aligning Data-Aware Declarative Process Models and Event Logs." In *Business Process Management: 19th International Conference, BPM 2021, Proceedings 19*, 235–251. Springer.

Blay, J.-Y., Charles Honore, Eberhard Stoeckle, Pierre Meeus, M. Jafari, François Gouin, P. Anract, et al. 2019. "Surgery in Reference Centers Improves Survival of Sarcoma Patients: A Nationwide Study." *Annals of Oncology* 30 (7): 1143–1153. https://doi.org/10.1093/annonc/mdz124.

Bogarín Vega, Alejandro, Rebeca Cerezo Menéndez, and Cristóbal Romero. 2018. "Discovering Learning Processes Using Inductive Miner: A Case Study with Learning Management Systems (LMSs)." *Psicothema* 30 (3): 322–329.

Borrego, Diana, and Irene Barba. 2014. "Conformance Checking and Diagnosis for Declarative Business Process Models in Data-Aware Scenarios." *Expert Systems with Applications* 41 (11): 5340–5352. https://doi.org/10.1016/j.eswa.2014.03.010.

Brennan, Murray F., Cristina R. Antonescu, Nicole Moraco, and Samuel Singer. 2014. "Lessons Learned from the Study of 10,000 Patients with Soft Tissue Sarcoma." *Annals of Surgery* 260 (3): 416–422. https://doi.org/10.1097/SLA.0000000000000869.

Burattin, Andrea, Fabrizio M. Maggi, and Alessandro Sperduti. 2016. "Conformance Checking Based on Multi-Perspective Declarative Process Models." *Expert Systems with Applications* 65:194–211. https://doi.org/10.1016/j.eswa.2016.08.040.

Crago, Aimee M., and Murray F. Brennan. 2015. "Principles in Management of Soft Tissue Sarcoma." *Advances in Surgery* 49 (1): 107–122. https://doi.org/10.1016/j.yasu.2015.04.002.

De Leoni, Massimiliano, and Wil M. P. van der Aalst. 2013a. "Aligning Event Logs and Process Models for Multi-Perspective Conformance Checking: An Approach Based on Integer Linear Programming." In *Business Process Management: 11th International Conference, BPM 2013. Proceedings*, 113–129. Springer.

De Leoni, Massimiliano, and Wil M. P. van der Aalst. 2013b. "Data-Aware Process Mining: Discovering Decisions in Processes using Alignments." In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, 1454–1461. Association for Computing Machinery.

de Medeiros, Ana Karla A., Anton J. M. M. Weijters, and Wil M. P. van der Aalst. 2007. "Genetic Process Mining: An Experimental Evaluation." *Data Mining and Knowledge Discovery* 14:245–304. https://doi.org/10.1007/s10618-006-0061-7.

De Oliveira, Hugo, Vincent Augusto, Baptiste Jouaneton, Ludovic Lamarsalle, Martin Prodel, and Xiaolan Xie. 2020. "Optimal Process Mining of Timed Event Logs." *Information Sciences* 528:58–78. https://doi.org/10.1016/j.ins.2020.04.020.

De Pinieux, Gonzague, Marie Karanian, Francois Le Loarer, Sophie Le Guellec, Sylvie Chabaud, Philippe Terrier, Corinne Bouvier, et al. 2021. "Nationwide Incidence of Sarcomas and Connective Tissue Tumors of Intermediate Malignancy Over Four Years Using an Expert Pathology Review Network." *PLoS One* 16 (2): e0246958. https://doi.org/10.1371/journal.pone.0246958.

Di Ciccio, Claudio, and Marco Montali. 2022. "Declarative Process Specifications: Reasoning, Discovery, Monitoring." In *Process Mining Handbook*, 108–152. Cham: Springer.

Dorigo, Marco, Vittorio Maniezzo, and Alberto Colorni. 1996. "Ant System: Optimization by a Colony of Cooperating Agents." *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 26 (1): 29–41. https://doi.org/10.1109/TSMCB.3477.

Fahland, Dirk. 2022. "Process Mining Over Multiple Behavioral Dimensions with Event Knowledge Graphs." In *Process Mining Handbook*, 274–319. Cham: Springer.

Fahland, Dirk, Daniel Lübke, Jan Mendling, Hajo Reijers, Barbara Weber, Matthias Weidlich, and Stefan Zugal. 2009. "Declarative versus Imperative Process Modeling Languages: The Issue of Understandability." In *Enterprise, Business-Process and Information Systems Modeling: 10th International Workshop, BPMDS 2009, and 14th International Conference, EMMSAD 2009, held at CAiSE 2009, Proceedings*, 353–366. Springer.

Felli, Paolo, Alessandro Gianola, Marco Montali, Andrey Rivkin, and Sarah Winkler. 2021. "CoCoMoT: Conformance Checking of Multi-Perspective Processes via SMT." In *Business Process Management: 19th International Conference, BPM 2021, Rome, Italy, September 6–10, 2021, Proceedings 19*, 217–234. Springer.

Gamboa, Adriana C., Alessandro Gronchi, and Kenneth Cardona. 2020. "Soft-Tissue Sarcoma in Adults: An Update on the Current State of Histiotype-Specific Management in an Era of Personalized Medicine." *CA: A Cancer Journal for Clinicians* 70 (3): 200–229.

Gantzer, Justine, Antonio Di Marco, Thibaut Fabacher, Noelle Weingertner, Jean-Baptiste Delhorme, David Brinkert, Guillaume Bierry, Jean-Pierre Ghnassia, Jérémie Jégu, and Jean-Emmanuel Kurtz. 2019. "Conformity to Clinical Practice Guidelines at Initial Management in Adult Soft Tissue and Visceral Tumors since the Implementation of the NetSarc Network in Eastern France." *The Oncologist* 24 (8): e775–e783. https://doi.org/10.1634/theoncologist.2018-0751.

Gingrich, Alicia A., Sarah B. Bateni, Arta M. Monjazeb, Steven W. Thorpe, Amanda R. Kirane, Richard J. Bold, and Robert J. Canter. 2019. "Extremity Soft Tissue Sarcoma in the Elderly: Are We Overtreating Or Undertreating this Potentially Vulnerable Patient Population?" *Journal of Surgical Oncology* 119 (8): 1087–1098. https://doi.org/10.1002/jso.v119.8.

Gomes, André Filipe Domingos, Ana Cristina Wanzeller Guedes de Lacerda, and Joana Rita da Silva Fialho. 2022. "Comparative Analysis of Process Mining Algorithms in Process Discover." In *New Trends in Disruptive Technologies, Tech Ethics and Artificial Intelligence: The DITTET Collection 1*, 258–270. Springer.

Gronchi, A., A. B. Miah, A. P. Dei Tos, N. Abecassis, J. Bajpai, S. Bauer, R. Biagini, et al. 2021. "Soft Tissue and Visceral Sarcomas: ESMO–EURACAN–GENTURIS Clinical Practice Guidelines for Diagnosis, Treatment and Follow-Up☆." *Annals of Oncology* 32 (11): 1348–1365. https://doi.org/10.1016/j.annonc.2021.07.006.

Günther, Christian W. 2009. "Process Mining in Flexible Environments." PhD diss., Eindhoven University of Technology.

Jensen, Anders Boeck, Pope L. Moseley, Tudor I. Oprea, Sabrina Gade Ellesøe, Robert Eriksson, Henriette Schmock, Peter Bjødstrup Jensen, Lars Juhl Jensen, and Søren Brunak. 2014. "Temporal Disease Trajectories Condensed from Population-Wide Registry Data Covering 6.2 Million Patients." *Nature Communications* 5 (1): 4022. https://doi.org/10.1038/ncomms5022.

Kusuma, Guntur P., Angelina Prima Kurniati, Ciarán D. McInerney, Marlous Hall, Chris P. Gale, and Owen Johnson. 2020. "Process Mining of Disease Trajectories in MIMIC-III: A Case Study." In *International Conference on Process Mining*, 305–316. Springer.

Kusuma, Guntur P., Angelina Prima Kurniati, Eric Rojas, Ciarán D. McInerney, Chris P. Gale, and Owen A. Johnson. 2021. "Process Mining of Disease Trajectories: A Literature Review." *Studies in Health Technology and Informatics* 281:457–461.

Lemma, Jasmiini, Sari Jäämaa, Jussi P. Repo, Kirsi Santti, Juho Salo, Carl P. Blomqvist, and Mika M. Sampo. 2023. "Local Relapse of Soft Tissue Sarcoma of the Extremities Or Trunk Wall Operated on with Wide Margins without Radiation Therapy." *BJS Open* 7 (2): zrac172. https://doi.org/10.1093/bjsopen/zrac172.

Leno, Volodymyr, Marlon Dumas, Fabrizio Maria Maggi, Marcello La Rosa, and Artem Polyvyanyy. 2020. "Automated Discovery of Declarative Process Models with Correlated Data Conditions." *Information Systems* 89:101482. https://doi.org/10.1016/j.is.2019.101482.

Li, Guangming, Renata Medeiros de Carvalho, and Wil M. P. van der Aalst. 2017. "Automatic Discovery of Object-Centric Behavioral Constraint Models." In *Business Information Systems: 20th International Conference, BIS 2017, Proceedings 20*, 43–58. Springer.

Liu, Cong, Huiling Li, Shuaipeng Zhang, Long Cheng, and Qingtian Zeng. 2023. "Cross-Department Collaborative Healthcare Process Model Discovery from Event Logs." *IEEE Transactions on Automation Science and Engineering* 20 (3): 2115–2125. https://doi.org/10.1109/TASE.2022.3194312.

Liu, Cong, Qingtian Zeng, Long Cheng, Hua Duan, and Jiujun Cheng. 2021. "Measuring Similarity for Data-Aware Business Processes." *IEEE Transactions on Automation Science and Engineering* 19 (2): 1070–1082. https://doi.org/10.1109/TASE.2021.3049772.

Lu, Steve H., and P. R. Kumar. 1991. "Distributed Scheduling Based on Due Dates and Buffer Priorities." *IEEE Transactions on Automatic Control* 36 (12): 1406–1416. https://doi.org/10.1109/9.106156.

Maggi, Fabrizio M., R. P. Jagadeesh Chandra Bose, and Wil M. P. van der Aalst. 2012. "Efficient Discovery of Understandable Declarative Process Models from Event Logs." In *Advanced Information Systems Engineering: 24th International Conference, CAiSE 2012, Proceedings 24*, 270–285. Springer.

Maggi, Fabrizio Maria, Marlon Dumas, Luciano García-Bañuelos, and Marco Montali. 2013. "Discovering Data-Aware Declarative Process Models from Event Logs." In *Business Process Management: 11th International Conference, BPM 2013, Proceedings*, 81–96. Springer.

Mannhardt, Felix, Massimiliano De Leoni, and Hajo A. Reijers. 2017. "Heuristic Mining Revamped: An Interactive, Data-Aware, and Conformance-Aware Miner." In *Business Process Management: 15th International Conference, BPM 2017, Proceedings*, 1–5. CEUR-WS.org.

Mannhardt, Felix, Massimiliano De Leoni, Hajo A. Reijers, and Wil M. P. van der Aalst. 2016. "Balanced Multi-Perspective Checking of Process Conformance." *Computing* 98 (4): 407–437. https://doi.org/10.1007/s00607-015-0441-1.

Mannhardt, Felix, Massimiliano De Leoni, Hajo A. Reijers, and Wil M. P. Van Der Aalst. 2017. "Data-Driven Process Discovery-Revealing Conditional Infrequent Behavior from Event Logs." In *Advanced Information Systems Engineering: 29th International Conference, CAiSE 2017, Proceedings 29*, 545–560. Springer.

Mannhardt, Felix, Sander J. J. Leemans, Christopher T. Schwanen, and Massimiliano de Leoni. 2023. "Modelling Data-Aware Stochastic Processes-Discovery and Conformance Checking." In *International Conference on Applications and Theory of Petri Nets and Concurrency, Proceedings*, 77–98. Springer.

Mohan, B. Chandra, and R. Baskaran. 2012. "A Survey: Ant Colony Optimization Based Recent Research and Implementation on Several Engineering Domain." *Expert Systems with Applications* 39 (4): 4618–4627. https://doi.org/10.1016/j.eswa.2011.09.076.

Oberoi, Sapna, Edwin Choy, Yen-Lin Chen, Thomas Scharschmidt, and Aaron R. Weiss. 2023. "Trimodality Treatment of Extremity Soft Tissue Sarcoma: Where Do We Go Now?" *Current Treatment Options in Oncology* 24 (4): 300–326. https://doi.org/10.1007/s11864-023-01059-2.

Pang, Jianfei, Haifeng Xu, Jun Ren, Jun Yang, Mei Li, Dan Lu, and Dongsheng Zhao. 2021. "Process Mining Framework with Time Perspective for Understanding Acute Care: A Case Study of AIS in Hospitals." *BMC Medical Informatics and Decision Making* 21 (1): 1–10. https://doi.org/10.1186/s12911-020-01362-0.

Peng, Zhihao, Vincent Augusto, Lionel Perrier, and Xiaolan Xie. 2024. "Optimal Process Mining of Traces with Events and Transition Attributes with Application to Care Pathways of Cancer Patients." *IEEE Transactions on Automation Science and Engineering* 21 (3): 4364–4381. https://doi.org/10.1109/TASE.2023.3295947.

Pesic, Maja, Helen Schonenberg, and Wil M. P. van der Aalst. 2007. "Declare: Full Support for Loosely-Structured Processes." In *11th IEEE International Enterprise Distributed Object Computing Conference (EDOC 2007)*, 287–287. IEEE.

Prodel, Martin, Vincent Augusto, Baptiste Jouaneton, Ludovic Lamarsalle, and Xiaolan Xie. 2018. "Optimal Process Mining for Large and Complex Event Logs." *IEEE Transactions on Automation Science and Engineering* 15 (3): 1309–1325. https://doi.org/10.1109/TASE.2017.2784436.

Rozinat, Anne, and Wil M. P. van der Aalst. 2006. "Decision Mining in ProM." In *Business Process Management: 4th International Conference, BPM 2006, Proceedings 4*, 420–425. Springer.

Schönig, Stefan, Claudio Di Ciccio, Fabrizio M. Maggi, and Jan Mendling. 2016. "Discovery of Multi-Perspective Declarative Process Models." In *Service-Oriented Computing: 14th International Conference, ICSOC 2016, Proceedings 14*, 87–103. Springer.

Strönisch, Annika, Sven Märdian, and Anne Flörcken. 2023. "Centralized and Interdisciplinary Therapy Management in the Treatment of Sarcomas." *Life* 13 (4): 979. https://doi.org/10.3390/life13040979.

Taghiabadi, Elham Ramezani, Vladimir Gromov, Dirk Fahland, and Wil M. P. van der Aalst. 2014. "Compliance Checking of Data-Aware and Resource-Aware Compliance Requirements." In *On the Move to Meaningful Internet Systems: OTM 2014 Conferences: Confederated International Conferences: CoopIS, and ODBASE 2014, Proceedings*, 237–257. Springer.

van der Aalst, Wil M. P. 2011. "Process Discovery: An Introduction." In *Process Mining: Discovery, Conformance and Enhancement of Business Processes*, 125–156. Berlin, Heidelberg: Springer.

Van Der Aalst, Wil M. P. 2016. *Data Science in Action*, 3–23. Berlin, Heidelberg: Springer.

van der Aalst, Wil M. P., and Josep Carmona. 2022. *Process Mining Handbook*. Cham: Springer Nature.

van der Aalst, Wil M. P., Ton Weijters, and Laura Maruster. 2004. "Workflow Mining: Discovering Process Models from Event Logs." *IEEE Transactions on Knowledge and Data Engineering* 16 (9): 1128–1142. https://doi.org/10.1109/TKDE.2004.47.

Wytiaz, Victoria, Eric Schwartz, John D. Rice, Lili Zhao, Rama Jasty, Scott Schuetze, and Rashmi Chugh. 2024. "Disparate Outcomes, Biologic and Therapeutic Differences in Pediatric versus Adult Patients with Ewing Sarcoma." *Oncology* 2 (1): 1–8. https://doi.org/10.1159/000533412.

Appendices

Appendix 1. Benchmarking heuristics

For the sake of drawing comparisons, we include a number of random-based and local search-based heuristics. The following offers descriptions of these heuristics.

- *Random* algorithm: this algorithm randomly generates a large number of initial solutions similar to the initial solution generation of LocalOpt and chooses the best. Attribute/transition function optimisation is applied to each solution. The number of solutions generated depends on the maximum allowable computational time.
- *RG* algorithm: this 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 with root macro-attributes to the current process model, connects the new node to/from all existing nodes by arcs with transition function equal to *S*, updates the event game, repairs the resulting process mode and updates the attribute/transition functions. The resulting model is set as the current model if it is better.
- *Reinsert* algorithm: this is a multi-start local optimisation algorithm. It starts from an initial solution of LocalOpt. At

each iteration, it determines the node $(\widehat{b}, k)$ of the lowest relevance and moves it to another layer $(\widehat{b}, k')$. Each local move from $(\widehat{b}, k)$ to $(\widehat{b}, k')$ is evaluated by the model relevance of the complete process model with an arc connecting any two nodes and the transition function equal to S. The local move with the highest model relevance is selected and the corresponding complete model is repaired and then improved by optimisation of attribute/transition functions. 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: this is similar to the *Reinsert* algorithm but with local move defined by relabelling the least relevant node $(\widehat{b}, k)$ as $(\widehat{b}', k)$, i.e. replacing the current macrolabel $\widehat{b}$ by $\widehat{b}'$.

As for LocalOpt and ACO, all these algorithms end with the joint optimisation of the event game and attribute/transition function.

Appendix 2. Introduction to sarcoma care pathways

Sarcomas are a large family of rare tumours that affect men and women at all ages. The estimated incidence is 70.7 per million inhabitants (about 4700 patients per year) in France (De Pinieux et al. 2021) with a similar incidence in England (Bacon et al. 2023). Given their rarity and the heterogeneity, sarcoma management is complex, and this sometimes results in sub-optimal management (Blay et al. 2019; Crago and Brennan 2015). We consider the following characteristics of the patient, disease, and healthcare organisation to have a significant impact on care pathways:

- *Histological subtypes*: these are key factors to consider when choosing a chemotherapy agent available for use in sarcoma patients. For example, angiosarcoma is highly sensitive to taxanes, which can be a treatment option in this histological type, whereas trabectedin has proven effective in treating advanced liposarcoma (Gronchi et al. 2021).
- *Tumor size*: radiation therapy can be considered for lesions > 5 cm in size (Gamboa, Gronchi, and Cardona 2020), and patients with extremity soft tissue sarcoma lesions (the majority of patients with soft tissue sarcoma) greater than 10 cm in size have a disease-specific survival rate that is significantly lower than those with lesions ≤ 5 cm (less than 40% vs. around 80% at 15 years, $p < 0.001$) (Brennan et al. 2014).
- *Tumor depth*: 75% of the patients with deep high-grade tumours treated with wide surgery alone developed distant metastases and expected to benefit from radiation therapy (Lemma et al. 2023). Neoadjuvant chemoradiotherapy, which is given prior to the primary treatment, may facilitate resection of large deep-seated tumours (Oberoi et al. 2023).
- *Age*: in Ewin's sarcoma, for example, children may have more cycles of first-line chemotherapy than adults, regardless of the stage of their cancer (Wytiaz et al. 2024).
- *Treatment location*: multidisciplinary management by reference centres is recommended (Blay et al. 2019; Strönisch, Märdian, and Flörcken 2023). The French network NETSARC includes 26 reference centres. Gantzer et al. (2019) observed a higher rate of quality resection surgeries (R0 and R1) in reference centres, 48.6% versus 32%. Initial sarcoma management in reference centres was shown to improve the overall survival and local relapse-free survival rate (Blay et al. 2019).