xPM: Enhancing Exogenous Data Visibility

Adam Banham[a,*], Sander J.J. Leemans[b], Moe T. Wynn[a], Robert Andrews[a], Kevin B. Laupland[a,c], Lucy Shinners[d]

[a] *Queensland University of Technology, Brisbane, Queensland, Australia*
[b] *RWTH, Aachen, Germany*
[c] *Department of Intensive Care Services, Royal Brisbane and Women's Hospital, Brisbane, Queensland, Australia*
[d] *Southern Cross University, Bilinga, Queensland, Australia*

Abstract

Process mining is a well-established discipline with applications in many industry sectors, including healthcare. To date, few publications have considered the context in which processes execute. Little consideration has been given as to how contextual data (exogenous data) can be practically included for process mining analysis, beyond including case or event attributes in a typical event log. We show that the combination of process data (endogenous) and exogenous data can generate insights not possible with standard process mining techniques. Our contributions are a framework for process mining with exogenous data and new analyses, where exogenous data and process behaviour are linked to process outcomes. Our new analyses visualise exogenous data, highlighting the trends and variations, to show where overlaps or distinctions exist between outcomes. We applied our analyses in a healthcare setting and show that clinicians could extract insights about differences in patients' vital signs (exogenous data) relevant to clinical outcomes. We present two evaluations, using a publicly available data set, MIMIC–III, to demonstrate the applicability of our analysis. These evaluations show that process mining can integrate large amounts of physiologic data and interventions, with resulting discrimination and conversion to clinically interpretable information.

Keywords: process mining, multi-perspective, exogenous data, MIMIC-III

1. Introduction

Process mining is a discipline that uses historical event data extracted from an organisation about a business process to better understand process behaviour (process activities) and performance [1]. Process mining techniques may be grouped into three high-level categories [1]; process discovery, conformance, and enhancement. Process discovery techniques [2, 3] exploit historical data to recreate the structure of a process. Process conformance techniques [4, 5] verify how process executions historically matched with a pre-described understanding of a process. Process enhancement techniques [1, 6] enrich the description of a process with an additional aspect to provide greater clarity of decisions made. Typical process mining insights focus on presenting an objective view of a process, identifying bottlenecks that slow down process execution, or otherwise inefficient behaviour within process executions.

Process mining has been used in healthcare to study emergency departments of hospitals [7, 8, 9], and the study of clinical conditions and diseases such as sepsis [10, 11, 12, 13, 14, 15], heart failure/chest pain [16, 8] and tuberculosis [17]. Applying process mining to healthcare processes allows clinicians to explore years of historical data to find opportunities to improve processes [18, 12, 19, 20, 21]. Such improvements can contribute to positive clinical, and beneficial financial, outcomes. Thus, finding these opportunities is critical [22, 23, 24]. While process mining offers a range of techniques to

*Corresponding author
Email address: adam.banham@hdr.qut.edu.au (Adam Banham)

support healthcare experts and processes, the application of techniques is not without challenges [25, 19, 26].

Challenges faced by process mining practitioners in the healthcare domain are outlined in [26] and focus largely around the process of decision making in healthcare. For example, the variety of data sources that are needed to make a clinical decision, the need for an interdisciplinary or institutional perspectives, the dependence on healthcare professionals that have the knowledge to make complex decisions, and that process behaviour will evolve to reflect recent healthcare developments. In [19], the authors' present recommendations for future work to address these challenges, such as an interdisciplinary team approach to the project and a multi-perspective techniques that provide ways to include a holistic view and application of the data available within healthcare organisations. While calls for multi-perspective techniques are not new, little work has, so far, been presented that handles these challenges.

In this paper, we propose a multi-perspective framework that allows a distinction between process behaviour and rich data flows, thus providing the potential to address the previously mentioned challenges. We adopted an interdisciplinary approach with expert healthcare professionals, who provided context to the evaluation, thereby enriching the process mining technique. To distinguish between different types of data sources, we define the following terms; *endogenous* process data and *exogenous* data.

We define *endogenous* process data as data internal to a process, meaning data that is directly produced by process activities, or that describes the moment of activity. Examples of endogenous data that describe a moment of a (healthcare) process include the name of the procedure being performed, the nurse performing the procedure, and the patient's level of discomfort during the procedure. In contrast, we define *exogenous*

data as data external to a process, meaning data that is not tied directly to a process' activities, but which provides the context in which process instances execute. Examples of exogenous data sources include availabilities of surgeons, ICU bed availability, and patient vital sign measurements.

We present xPM, a framework for process mining with exogenous data. xPM prescribes the steps for connecting contextual data relevant to the endogenous moments of a process, and for conducting process discovery and enhancement. We also present a new process mining technique that uses xPM to understand how contextual factors influence process behaviour. Specifically, our analysis can visualise differences between changes in exogenous data and process outcomes. Furthermore, instead of manually processing our analysis, we propose an automatic ranking procedure to identify similar trends between process outcomes.

To evaluate our approach for process mining, we quantitatively compare results produced by data-aware techniques using both endogenous and exogenous data. We also evaluate if xPM can find trends within exogenous data that influences process behaviour. To do so, we use our technique to model activities in ICU wards and associated occurrences of (hospital acquired) sepsis. We use the publicly available MIMIC-III [27]dataset from the healthcare domain from which we extract both events relating to patient treatment in ICU (endogenous data), and patient vital signs measurements (exogenous data).

Using xPM, process mining analysts can better capture the reality in which the process executes by making multiple contextual data sources available to process mining techniques.Furthermore, xPM affords opportunities for process participants to describe decisions occurring within complex environments in terms of the context.

Our framework for process mining with exogenous data was initially presented in

[28]. We extended our previous work by presenting a new technique for process mining and an extension of the evaluation with our expert health professionals co-authors about sepsis/infection.

This article is structured as follows. In Section 2, we present related work. In Section 3, we present the formalisation needed to understand our proposed approach. In Section 4, we present the proposed approach for process mining analysis with exogenous data. In Section 5, we outline our evaluation designs and present results of applying our approach. In Section 6, we summarise our findings and explore future work.

2. Related Work

In this section, we highlight the use of process mining within the healthcare domain, categorisation of data for processes and multi-perspective process mining.

2.1. Process Mining and Healthcare

Mans et al. [25], provides an overview of 12 process mining applications within the healthcare domain, and a list of questions frequently asked by medical professionals about healthcare processes. All these questions related to learning how a process was executed (control-flow perspective). In a later literature review [18], the authors identified 74 papers related to applications of process mining in healthcare. The authors noted that most studies had focused on improving the control-flow perspective of processes, and concluded that there was a need for better visualisations, and a reliance on experts to apply techniques.

In a subsequent systematic review [29], the authors performed descriptive analysis on 55 articles and characterised the current dimensions of process mining in the healthcare

setting. The authors noted that there was a need for techniques to capture more details about the context of a hospital intervention within their characterisation.

In [19], the authors present ten recommendations from process mining researchers and healthcare experts to stimulate a more widespread use of process mining in healthcare. In addition, the authors encouraged efforts that enhance the usability and understandability of techniques. A key recommendation was the need for more multi-perspective studies.

In [20], the authors present a comprehensive literature review of 172 papers that overlap both process mining and healthcare domains between 2010 and 2020. The authors noted that the complexity of healthcare data and the heterogeneous nature of treatment pathways, has meant techniques still require fine-tuning for specific medical cases. Additionally, further work is needed to make process models more descriptive to be successfully employed and understood by healthcare professionals.

These reviews highlight a high level of involvement between the process mining community and the healthcare domain. However, these studies show that little work exists for multi-perspective techniques that incorporates diverse data sources alongside process behaviour.

2.2. Data for Processes

Categorisation of data sources used to describe processes has been discussed in several studies. The 'onion skin' model in [30] conceptualises the relationship between data and process as the viewpoint is moved further away from a process. This conceptualisation is then applied to process mining in [31], where data is categorised according to the likelihood of cause and effect between variables with the process. However, these frameworks are not seen as essential to process mining by recent reviewers of the field,

and the contextual component remains an optional consideration during event data extraction [32, 33, 34]. Our contribution is that we support separate entities for endogenous and exogenous data sources, such that they can studied in combination.

The benefits of including a variety of data categories are discussed in [35] which (i) motivates the use of data attributes for distinguishing between noise and conditional process behaviour, (ii) considers if data attributes influence decision points by creating an internal state as the process executes through boolean expressions and decision trees, and (iii) studied how alignments [5] can be extended such that they balance both the control flow perspective and the data perspective.

2.3. Multi-Perspective Process Mining

Our study extends the concepts presented in [35] and shows how exogenous data can be incorporated in a process mining analysis (instead of the analysis being limited to only the endogenous perspective of an event log).

Methodologies that encourage contextual data collection and log enrichment are few in number. However, some recent studies have focused on the enrichment of an event log with new types of data. In [36], the authors present a framework for intra- and inter-trace predictive monitoring and introduce the notion of *bi-dimensional coding* to deal with intra- and inter-trace dependencies. In [37], the authors present an approach to compare process variants using the endogenous data available within an event log with a *differential perspective graph*. In [38], the authors present an approach to characterise endogenous data, into *static, semi-dynamic, and dynamic* attributes to identify attributes of interest. In [39], the authors suggest that not all events within an event log are about the control flow, and are instead, about the data flow of a process. They use the concept

of *context events* to deal with the two types of events and show how distinguishing between the two can lead to discovering less complex models. However, this approach would incorporate exogenous context into the control flow perspective instead of clarifying whether the context influences process execution.

The benefits of additional data attributes that can be seen in the recent evolution of techniques that use such data, e.g., [40, 2]. In [40], the authors present a discovery algorithm that uses data attributes to create a hierarchical model to improve the simplicity of outcomes. Another approach, in [2], was to create a constraint operator for process trees notation, whereby data semantics can be expressed. While these techniques can create control flow sequences based on data attributes, no extensions have been proposed to use exogenous sources outside what can be found in the events within an event log.

The previously mentioned studies often focused on the imperative representation of models (strictly specifying how a process executes); however, a more flexible description for process execution exists in techniques that are focused on declarative modelling [41]. One such example of a declarative modelling approach is DECLARE, which has developed using a constraint-based system grounded in temporal logic [41].

Studies such as those in [42] and [43] use this approach, which incorporates multiple process perspectives. In [42], the authors present a framework for multi-perspective declarative process discovery, whereby different process perspectives (time, resource, or data) could be used to generate a set of constraints. In [43], the authors present declarative process discovery, focused on finding constraints based on correlated data attributes within constraint instances, explicitly focusing on the control-flow and data perspective of a process. While extensive work has been done with endogenous data within declarative process mining, it is difficult to extend these studies to exogenous data without a

representation available for exogenous data at the moment of process activities.

Similar work can be seen in the healthcare domain, where process mining has been applied to understand many perspectives across a health process. In [44], the authors proposed a comparison measurement for two models to compare clinical guidelines with historical health records. Subsequently, analysis was performed to relate clinical outcomes for stroke management with a computed distance between a mined process model and the query process model. However, this analysis did not correlate clinical outcomes with process activities or exogenous data, but instead with the computed distance.

In [16], the authors present a methodology for clustering patient treatment pathways and provide descriptive analysis on the clustering outcomes, such as finding clusters with low mortality rates. However, the clinical outcome was introduced in this study after the clustering was completed, meaning that clinical outcomes are not correlated with process activities. In [45], the authors studied readmission risk to ICU using time series data which consisted of physiological data and medication trends, and processing time-series data into frequent subgraphs. Then using a predictive model, subgraph membership was related to readmission. However, while this study did connect exogenous data to a clinical outcome, it does not consider the activities that occurred within the ICU stay and so cannot be considered a process analysis.

The following works present in [46, 47, 48], highlight the importance of ICU activities or procedures that occurred within the hospital stay. Each piece of work in this set, has performed process mining efforts on a timeline of episodes to model personalised treatment pathways to improve the efficiency of health services. In [49], the authors argue that valuable analysis can come from performing these techniques on a population of patients with similar health conditions.

In summary, it is clear that while context has been recognised as influencing process behaviour, only limited progress has been made in including contextual data (mostly as case or event attributes) for process mining. Our review found no techniques capable of exploiting exogenous data, as we define it, to discern its influence on process behaviour. Furthermore, pre-processing exogenous data remains a mostly manual task that requires domain knowledge or expertise. These limitations constitute a research gap which we address in the remainder of this paper.

3. Theory

This section introduces events logs, exogenous data sets, Petri nets and the sub-formalisation for process descriptions we use, xDPN. To formalise these concepts, we adopt the notation from [50] and include them here for completeness.

Universes. $\mathcal{U}_E$ is the universe of events, $\mathcal{U}_D = \{case, act, time, \dots\}$ is the universe of data variables, $\mathcal{U}_V$ is the universe of values, and $\mathcal{U}_{\text{map}}$ is the universe of variable-value mappings. $\mathcal{U}_\Gamma$ is the universe of timestamps, $\mathcal{U}_A$ is the universe of activities. Hence, $\mathcal{U}_A \cup \mathcal{U}_\Gamma \subset \mathcal{U}_V$ [50] .

Event logs are a collection of events. An event e is an execution step of a process, recorded through data variables. Formally, an event log is (E, π), where $E \subseteq \mathcal{U}_E$ is a set of events and $\pi \in (\mathcal{U}_E \to \mathcal{U}_{\text{map}})$ is a function that maps an event e to a variable-value mapping. We use the shorthand $e_{\pi,\text{A}}$ to denote the value of a data variable $\text{A} \in \text{dom}(\pi(e))$ for event e, i.e. $\pi(e)(\text{A}) = e_{\pi,\text{A}}$ [50].

We assume an event e to have at least the following data variables: $\{case, act, time\}$, such that *case* is the trace related to an event, *act* is the process activity related to an event, and *time* denotes when an event occurred. We use the following shorthand

to denote these data variables as needed, $\lambda : \mathcal{U}_E \to \mathcal{U}_A$ and $\lambda(e) = \pi(e)(act)$, also $\Gamma : \mathcal{U}_E \to \mathcal{U}_\Gamma$ and $\Gamma(e) = \pi(e)(time)$, finally $\mathcal{C} : \mathcal{U}_E \to \mathcal{U}_V$ and $\mathcal{C}(e) = \pi(e)(case)$.

Traces. An end-to-end execution of a process, or a trace ($\mathrm{T} \in \mathbb{T}$) is a sequence of events derived from an event log (E, π), such that $\mathbb{T} = \{\langle e_1, \ldots, e_n \rangle \mid e_1, \ldots, e_n \in E \wedge \forall_{1 \leq i < n}\ \mathcal{C}(e_i) = \mathcal{C}(e_{i+1}) \wedge \Gamma(e_i) \leq \Gamma(e_{i+1})\}$. We extend π to also apply to traces, where it denotes the set of all values present in the trace, i.e. $\Pi : \mathbb{T} \to (\mathcal{U}_D \nrightarrow 2^{\mathcal{U}_V})$, with $\Pi(\mathrm{T})(\mathrm{A}) = \{\pi(e)(\mathrm{A}) \mid e \in \mathrm{T}\}$ or denoted using the shorthand of $\mathrm{T}_{\Pi,\mathrm{A}}$.

Exogenous Data. An *exogenous measurement* (exo–measurement) m^t is a single measurement $m \in \mathcal{U}_V$ at timestamp $t \in \mathcal{U}_\Gamma$. For instance, $37^{18-01-202211:55:23}$ may represent the result of a nurse measuring the temperature of a particular patient at the given timestamp. An *exogenous series* (exo–series) $\langle m_1^{t_1}, \ldots m_n^{t_n} \rangle^A$ is a sequence of exo–measurement, annotated with a variable-value mapping A. Furthermore, we denote the universe of exo–series as $\mathcal{U}_{es}$. For instance, $\langle 37^{18-01-202211:55:23}, 37.5^{18-01-202212:13:58}, 38^{18-01-202213:14:32} \rangle^{\text{patient\_id:1034847}}$ may represent the repeated measuring of body temperature by a nurse for a particular patient. We use the following shorthand to denote the ith element of an exo–series es, such that $es(i)$ is the ith exo–measurement recorded within an exo–series. Furthermore, to denote the value of a variable A for an exo–series es, we use $es_{\#_\mathrm{A}}$.

An *exogenous panel* (exo–panel or $\mathbb{P}$) is a collection of exo–series ($\mathbb{P} \subseteq \mathcal{U}_{es}$). For instance, $\{\langle 37^{18-01-202211:55:23}, 38^{18-01-202213:14:32} \rangle^{\text{patient\_id:1034847}}, \langle 37.8^{18-01-202215:47:39}, 37.8^{18-01-202220:15:41} \rangle^{\text{patient\_id:1034882}}\}$. An *exogenous description* (exo–description or $\mathcal{U}_{\text{exo}}$) is a collection of exo–panels. Typically, an exo–panel would contain all related exo–measurements; for instance: all blood pressure measurements for all patients. An exo–description then contains several exo–panels; for instance: one for blood pressure, one for heart rate, and one for body

temperature.

Labelled Petri Nets. A *Petri net* is a triple $N = (P, T, F)$, where P is a finite set of places, T is a finite set of transitions such that $P \cap T = \emptyset$ and $F \subset (P \times T) \cup (T \times P)$ is a set of directed arcs, called a flow relation [1]. A *labelled Petri net* is a quintuple, $(P, T, F, \Sigma, \lambda)$, where (P, T, F) is a Petri net, $\Sigma \subseteq \mathcal{U}_A$ is a set of activity labels and λ is overloaded to apply to transitions or an event labelling function $\lambda : T \to \mathcal{U}_A$ [1]. Places may hold tokens, which are produced and consumed when transitions fire according to the flow relation. A transition is *enabled* if each input place contains a token. The state of a Petri net is a *marking*, which records what places have tokens and how many. An enabled transition l can *fire*, which updates the marking according to the flow relation F and, if l is labelled by λ, denotes the execution of activity $\lambda(l)$. An *initial marking* denotes the initial state of a Petri net before the first transition is fired. Elements of $P \cup T$ are called nodes. A node a is an input node of another node b if and only if there is a directed arc from a to b. A node a is an output node of another b if and only if there is a directed arc from b to a. For any $a \in P \cup T$, $\bullet a = \{b \mid (b, a) \in F\}$ and $a\bullet = \{b \mid (a, b) \in F\}$.

Petri Nets with Exogenous Data (xDPN). A *precondition* (ϕ) is a boolean expression describing a subset of values for data variables (e.g. temperature is higher than 20°C). A *Petri Net with Exogenous Data* (xDPN) is a sextuple $(P, T, F, \Sigma, \lambda, \Phi)$, where $(P, T, F, \Sigma, \lambda)$ is a labelled Petri net and $\Phi : T \to \phi$ associates transitions with preconditions. The state of an xDPN is a marking and an assignment of endogenous and exogenous variables. A transition is *data enabled* if the precondition attached to a transition is satisfied by the current assignment of variables, or if there is no attached precondition. In an xDPN, a transition can fire if it is enabled and data enabled. An xDPN is a sub-formalism of DPN (a complete formalisation of DPN can be found in [51, 35, 52]): in contrast to xDPN, DPN distinguish between variable states (e.g. `read` or `written`) and enforce that transitions in the model update variable assignments. By dropping this last requirement in xDPN, exogenous data variables can be updated during execution.

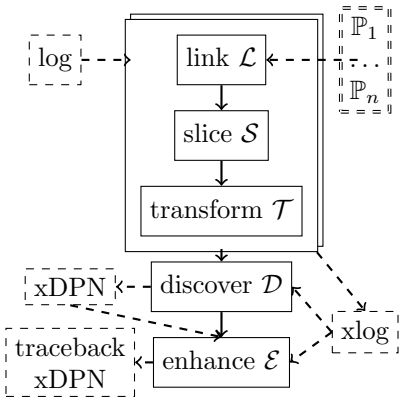


Figure 1: The xPM framework.

4. Method

In this section, we present our approach for using exogenous data in process mining. First, we propose a framework for process mining with exogenous data. After that, we present a process enhancement technique that utilises the framework to visualise trends in exogenous data (exogenous behaviour) based on process outcomes. Finally, we present a procedure to rank such visualisations and allows users to find where exogenous trends are similar between process outcomes.

4.1. A Framework for Process Mining with Exogenous Data (xPM)

In this section, we introduce xPM, our framework for process mining with exogenous data (see Figure 1). xPM takes as input an event log and an exo–description ($\mathcal{U}_{exo}$). xPM applies several *determinations*, quadruples of $(\mathbb{P}, \mathcal{L}, \mathcal{S}, \mathcal{T})$, and creates as output an *exogenous-annotated* log (xlog). In each determination, $\mathbb{P} \in \mathcal{U}_{exo}$ is an exo–panel, $\mathcal{L}$ is a linking function which connects a trace and an exo–panel to an exo–series, $\mathcal{S}$ is a slicing function which creates sub-series (*slice*) of an exo–series for each event, and $\mathcal{T}$ is a transformation function, which transforms a slice into one or more event attributes.

xPM applies each determination to annotate events with transformed attributes, creating an *exogenous-annotated* log (xlog). From the xlog, a discovery function $\mathcal{D}$ can discover an xDPN. In the last step, an enhancement function $\mathcal{E}$ aligns an xlog with a process model to visualise differences in process executions. After enhancement, we can *trace back* from the exogenous variables to an exo–series for fine-grain analysis. In this paper, we illustrate xPM using time series of continuous measurements. However, the xPM framework does not limit the type of time series that can be used. In Section 4.1.1, we outline an exemplar scenario used to describe xPM.

4.1.1. Exemplar Scenario

In this exemplar scenario, we are studying how ambulance triage occurs and what influences if a patient is treated on-site, transported to a hospital or handed over to another ambulance. In Table 1 we present a snippet of this event log (i.e. the endogenous data). Events have the

Table 1: Exemplar event log for ambulance triage.

$\mathcal{C}(e_{\mathrm{i}})$	e_{i}	$\Gamma(e_{\mathrm{i}})$	$\lambda(e_{\mathrm{i}})$	$e_{\mathrm{i}\,\pi,\mathrm{suburb}}$	$e_{\mathrm{i}\,\pi,\mathrm{ambulance}}$	$e_{\mathrm{i}\,\pi,\mathrm{incident}}$
1	1	18-01-2022-02:25	request	Runcorn		15
1	2	18-01-2022-02:30	assign		A	
1	3	18-01-2022-03:00	arrival			
1	4	18-01-2022-04:00	pickup			
1	5	18-01-2022-04:30	delivery	Brisbane		
2	6	18-01-2022-06:18	request	Runcorn		17
2	7	18-01-2022-06:20	assign		A	
2	8	18-01-2022-06:40	arrival			
2	9	18-01-2022-06:55	pickup			
2	10	18-01-2022-07:10	hand-over		B	
2	11	18-01-2022-07:35	delivery	Brisbane		

Table 2: Exemplar of a tublarised exo–description. All timestamps occur on 18-01-2022 in this example.

exo–panel	exo–series id	attributes	exo-measurements
weather	1	{suburb: Runcorn }	$\langle -5^{03:00}, -2^{04:00}, 3^{05:00} \rangle$
weather	2	{suburb: Brisbane}	$\langle 0^{03:00}, 2^{04:00}, 5^{05:00} \rangle$
heart rate	4	{ambulance: A, incident: 15}	$\langle 95^{03:15}, 97^{03:20}, 110^{03:25} \rangle$
heart rate	6	{ambulance: B, incident: 17}	$\langle 105^{06:55}, 110^{07:00}, 115^{07:05} \rangle$
heart rate	7	{ambulance: A, incident: 17}	$\langle 120^{07:15}, 122^{07:20}, 125^{07:25} \rangle$

following variables; `suburb` is the destination of the ambulance, `ambulance` is the identifier for an ambulance, and `incident` is the identifier for an ambulance request. After arriving onsite, paramedics may observe the patient to determine if the patient needs to be transported and if so, they may also consider traffic, weather or the nearest ambulance with a senior paramedic. An analyst may want to understand how these factors influence how paramedics determine the best way to transport a patient.

To consider exogenous data in this scenario, we will use an exo–description containing two exo–panels: the temperature in each suburb and the heart rate of the patient. In Table 2 we present a snippet of exo–series from different exo–panels. Using xPM, we can relate these exo–series with traces, to understand how paramedics observe these factors and how process behaviour changes based on exogenous data.

4.1.2. $\mathcal{L}$ – Linking

Our framework starts by linking exo–panels to traces, as shown in Figure 2. The act of linkage is described in a linking function ($\mathcal{L}$), and if a link exists, finds an exo–series for the trace. For example, a linking function could connect an ICU admission to the workload of an attending nurse, to laboratory results, or to weather conditions surrounding the hospital.

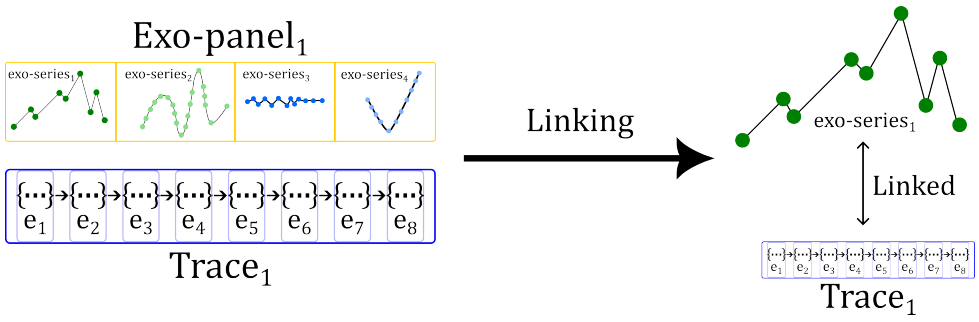


Figure 2: An illustration of linking. Given a trace and an exo–panel, the linking function returns a linked exo–series for the given trace.

Formally, $\mathcal{L}$ takes a trace $\mathrm{T} \in \mathbb{T}$ and an exo–panel $\mathrm{P} \in \mathbb{P}$, and finds an exo–series, i.e $\mathcal{L} : (\mathbb{T} \times \mathbb{P}) \rightharpoonup \mathcal{U}_{es}$. To illustrate $\mathcal{L}$, we revisit our exemplar scenario. In this scenario, we have two exo–panels, temperature and heart rate. A suitable $\mathcal{L}$ for temperature is $\mathrm{L}_1 \in \mathcal{L}$, showing the suburb attribute from events ($\mathrm{T}_{\Pi,\mathrm{suburb}}$) is used to find a matching exo–series. A suitable $\mathcal{L}$ for heart rate is $\mathrm{L}_2 \in \mathcal{L}$, showing the ambulance and incident variables from events ($\mathrm{T}_{\Pi,\mathrm{incident}}$, $\mathrm{T}_{\Pi,\mathrm{ambulance}}$) is used to find a matching exo–series.

$$\mathrm{L}_1(\mathrm{T},\mathrm{P}) = es \iff es \in \mathrm{P} \wedge es_{\#_{\mathrm{suburb}}} \in \mathrm{T}_{\Pi,\mathrm{suburb}}$$

$$\mathrm{L}_2(\mathrm{T},\mathrm{P}) = es \iff es \in \mathrm{P} \wedge es_{\#_{\mathrm{incident}}} \in \mathrm{T}_{\Pi,\mathrm{incident}} \wedge es_{\#_{\mathrm{ambulance}}} \in \mathrm{T}_{\Pi,\mathrm{ambulance}}$$

In case $\mathcal{L}$ finds two or more exo–series for a trace, $\mathcal{L}$ must merge these into a single exo–series. Merging time series is not trivial and will require a thorough understanding of exogenous data and domain knowledge. For example, we have two heart rate sensors monitoring a patient, but they are not time-synchronised and need to be handled within the linking function to find the truthful exo–series.

4.1.3. $\mathcal{S}$ – Slicing

A slicing function ($\mathcal{S}$) associates a sub-series of an exo–series (*slice*) with each event in a trace. After applying a slicing function, some events may be associated with an empty slice, some are associated with the same slice, and others have their own slice (as seen in Figure 3).

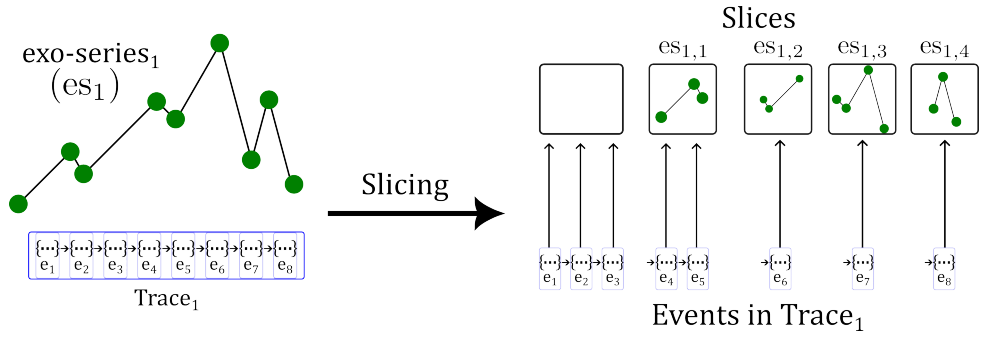


Figure 3: An illustration of slicing, where given trace$_1$ and exo–series$_1$, some events are associated with the same slice (e_4, e_5), some have their own slice (e_6, e_7, e_8), and others are associated with an empty sequence (e_1, e_2, e_3).

Examples of how a slicing function may create slices include selecting all exo–measurement before a patient undergoes a procedure, all exo–measurement in the hour before a procedure, or all exo–measurement between major interventions.

Formally, a slicing function $\mathcal{S}$ is applied to each event in a trace. It takes an event and an exo–series, to return a sub-series of the exo–series, i.e. $\mathcal{S}\colon (\mathcal{U}_E \times \mathcal{U}_{es}) \to \mathcal{U}_{es}$. Thus, creating a slice of relevant exogenous data from an exo–series, for each event. Domain knowledge will often inform the choice of slicing functions; thus, assisting the automatic selection of a slicing function is an interesting area of further research. Such options for automating slicing could consider the frequency of measurements, autocorrelation with a time lag, or investigating the amount of variance within an exo–panel. Another approach could be to consider research around databases or OLAP operations [53, 54] to automate or provide complete functions for slicing time series data.

To illustrate $\mathcal{S}$, we revisit our exemplar scenario. In this scenario, the two exo–panels have different durations of relevance, for temperature longitude trends are crucial. Thus, a suitable $\mathcal{S}$ for temperature is $\mathrm{S}_1 \in \mathcal{S}$, where slices are created using a two-hour window of exo–measurements (before and after each event). In contrast, for heart rate, short-term changes are

significant. Therefore, a suitable $\mathcal{S}$ for heart rate is $\mathrm{S}_2 \in \mathcal{S}$, where slices are created from a 15-minute window of exo–measurements before each event.

$$\mathrm{S}_1(e, es) = \left\langle m^t \mid m^t \in es \wedge (\Gamma(e) - 1 \text{ hour}) \leq t \leq (\Gamma(e) + 1 \text{ hour}) \right\rangle$$

$$\mathrm{S}_2(e, es) = \left\langle m^t \mid m^t \in es \wedge (\Gamma(e) - 15 \text{ minutes}) \leq t \leq \Gamma(e) \right\rangle$$

4.1.4. $\mathcal{T}$ – Transformation

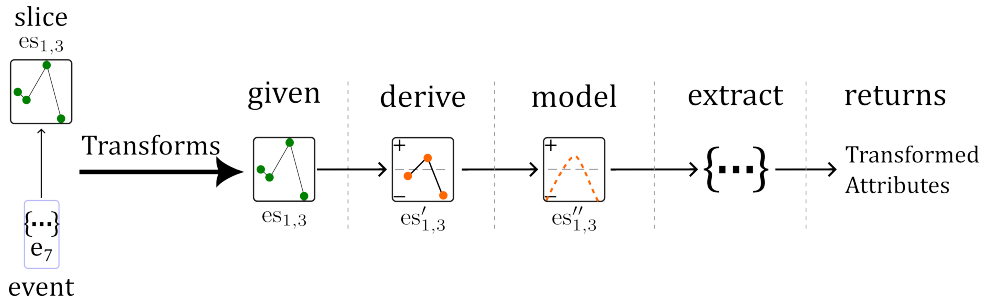


Figure 4: An illustration of transformation, where given e_7 and a slice of an exo–series $\mathrm{es}_{1,3}$, the transformation derives the velocity of a slice ($\mathrm{es}'_{1,3}$), models the derivation ($\mathrm{es}''_{1,3}$), then extracts notable features from the modelling as transformed attributes.

Next, we transform the slice for existing process mining techniques and future techniques, as seen in Figure 4. The result of a transform[1] for a slice, is a set of one or more event attributes. This transformation is handled through a transformation function ($\mathcal{T}$), which takes an event and an associated slice, to create event attributes (*transformed attributes*) for the notable features of a slice. Formally, we denote a transformation function as, $\mathcal{T} : (\mathcal{U}_E \times \mathcal{U}_{es}) \rightarrow 2^{\mathcal{U}_D \times \mathcal{U}_V}$. Examples of transformations may include whether an extreme measurement exists, the number of peaks and lows, or whether the trend line of a slice was increasing or decreasing over time.

We have identified three forms that a $\mathcal{T}$ can take: (i) $\mathcal{T}$ can return a single attribute; such a transformation might return the minimum, maximum or mean of a slice; (ii) $\mathcal{T}$ can return a set

[1]Where we refer to transformation, any aggregation beyond min, max and mean can be considered as well.

of attributes; e.g. coefficients of the nth Taylor polynomial of the slice; (iii) $\mathcal{T}$ can be recursive to meet either case (i) or (ii). Such a transformation finds the nth derivative of a slice, then applies any previously mentioned examples, to meet the form of either (ii) or (iii).

To illustrate $\mathcal{T}$, we revisit our exemplar scenario. In this scenario, the interpolation is vastly different between exo–panels and each requires a unique transformation function. A suitable $\mathcal{T}$ for temperature is $\mathcal{T}_1 \in \mathcal{T}$, showing a slice being transformed into an attribute that states if extreme weather occurred. A suitable $\mathcal{T}$ for heart rate is $\mathcal{T}_2 \in \mathcal{T}$, showing a slice being transformed into many attributes, such as the velocity of a heart rate and if a projected heart rate pattern will go outside guidelines.

$$\text{abnormal}(es) \equiv \exists_{m^t \in es} -10℃ \geq m^t \vee m^t \geq 40℃$$

$$\mathcal{T}_1(e, es) = \{(\text{`extreme'}, \text{abnormal}(es))\}$$

$$\text{velocity}(es) = \left\langle \left(\frac{m_{i+1} - m_i}{t_{i+1} - t_i} \right)^{t_{i+1}} \middle| m_i^{t_i} = es(i) \wedge m_{i+1}^{t_{i+1}} = es(i+1) \wedge 1 \leq i < |es| \right\rangle$$

$$\text{avg}(es) = \sum_{m^t \in es} \frac{m}{|es|}$$

$$\text{dng}(m) \equiv -5 \geq m \vee m \geq 5$$

$$\mathcal{T}_2(e, es) = \{(\text{`velocity'}, \text{avg}(\text{velocity}(es))), (\text{`danger'}, \text{dng}(\text{avg}(\text{velocity}(es))))\}$$

In this scenario, we have created two determinations for xPM to enact, $(\text{temperature}, \text{L}_1, \text{S}_1, \mathcal{T}_1)$ and $(\text{heart rate}, \text{L}_2, \text{S}_2, \mathcal{T}_2)$. Access to exogenous data is facilitated after xPM has applied all determinations to an event log, such that some events are annotated with (i) exogenous slices and (ii) transformed attributes. We refer to such an event log as an *exogenous-annotated* log (xlog).

4.1.5. $\mathcal{D}$ – Discovery

This step uses the newly created xlog to discover how process behaviour may vary based on the available exogenous data. A discovery function ($\mathcal{D}$) is applied to the xlog to achieve this goal. Given the plenitude of process discovery techniques and process notations in literature,

both the slices and transformed attributes can be used to investigate process behaviour using existing techniques.

One example of process discovery techniques that could use transformed attributes is decision mining, which finds insightful business rules for decision points. Examples of such techniques are [55] and [56]. However, researchers should be careful about assumptions that these techniques make about the nature of data variables. One assumption made is the constant data placement within process executions, which suits endogenous data variables but may have unintended consequences when considering exogenous data variables. Another approach for process discovery techniques could be to use the lower-level information stored in the slices. Using the trends stored within slices, techniques could extract unique process behaviour associated with the changes within the slice. The description of process behaviour based on exogenous influences will require new process notations and execution semantics but may lead to valuable insights.

4.1.6. $\mathcal{E}$ – Enhancement

In the final step of our framework, the enhancement step $\mathcal{E}$, uses both endogenous and exogenous data in an xlog and a process model. The enhancement step uses this information to visualise the process and the exogenous influences on the process. One approach for techniques could be to investigate process activities linked to exogenous changes and use explainable machine learning techniques [57, 58, 59] to explain why variance occurs. Another approach could be to use the exogenous data, find cohorts of traces with similar contextual settings, and then visualise differences in process executions between cohorts. Furthermore, techniques describe trends within slices on process models to show differences between process outcomes.

We note that process mining techniques in this step are likely to rely on process conformance techniques that produce an alignment between a process model and an event log. An (optimal) alignment is a mapping between observed process behaviour occurring in reality (traces/events) and a process model such that the deviation between is minimised [5]. While data-aware align-

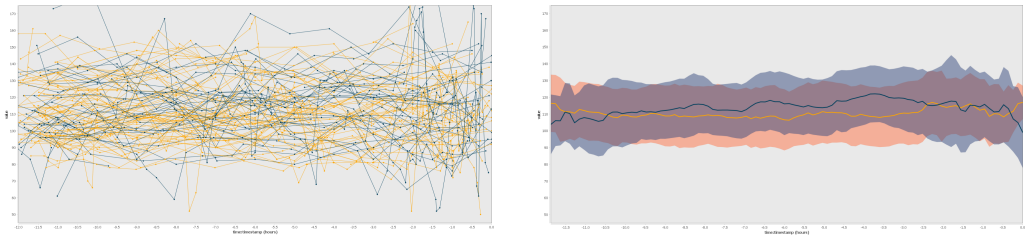
ments exist (e.g. [4, 60]), [4] only considers the writing of variables by transitions and [60, 4] correct the data written by transitions using Integer Linear Programming. However, exogenous data should not be adjusted in conformance techniques as it occurs outside the internal process execution. Further work is needed to generalise these techniques for usage outside business processes, as assumptions introduced in these techniques about the nature of data variables may have negative consequences when applied to exogenous data.

4.2. Explorative Exogenous Series Analysis (EESA)

We refer to our enhancement technique for xPM as Explorative Exogenous Series Analysis (EESA). This technique visualises changes in exogenous sources seen at a point within a process model. This technique builds upon the xPM framework by visualising slices. Naively, one could visualise the slices by using a line for each slice observed, as seen in Figure 5a. Identifying trends or similarities at a glance may only be possible when the number of slices is relatively low or when the variance between slices is low. Our technique overcomes this drawback and presents the observed trends regardless of the number of slices used, as seen in Figure 5b. At a glance, users can see the variance between slices and the median trend between slices in an exogenous source between activities in a process model. However, in this paper, we focus our contribution on using continuous measurements within slices.

In this section, we present the steps to handle slices and generate a visualisation. These steps start with collecting slices to form observations, constructing a shared time axis, interpolation, and visualisation. In section 4.2.1, we present a variation of these steps that allows for multiple process outcomes to be considered when visualising slices. Finally, in section 4.3, we present an automatic analysis procedure for ranking a collection of EESA visualisations.

EESA uses a collection of observations (O) as an input data source. Each observation (e, s, A, X), is a quadruple of an event (e), a slice (s), the activity label of e ($A = \lambda(e)$) and the exogenous data source (X). Each observation used as input to our technique must have the



(a) Raw Slices

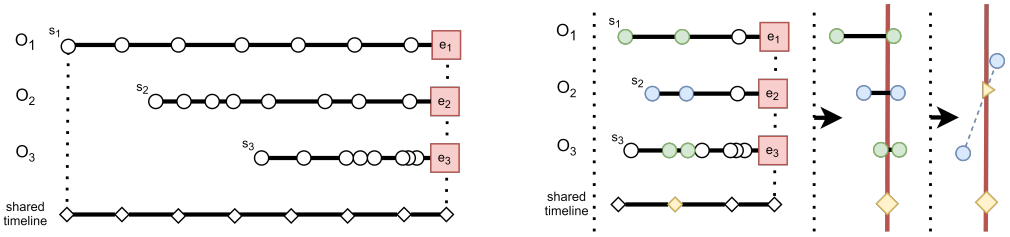
(b) Overlapping EESA Visualisation

Figure 5: A comparison of visualisation techniques, one using the exogenous slices as-is (left), and another showing the generalisation of trends seen in raw slices (right). The usage of colour in each figure, is to highlight that a slice is associated with a particular process outcome.

same A and X, to ensure that all slices are comparable when visualised. Before visualisation, each s is resampled across a shared time axis (on relative time from e) using linear interpolation to ensure that a consistent view is presented. Finally, we interpolate slices for each point on the shared time axis and plot the median value and one standard deviation on each side of the median value.

EESA's first step is to construct a shared time axis covering the longest timespan seen in observations. To construct this time span, we consider all slices (s) in relative time difference from their event (e), which ensures that at time `00:00:00`, we always observe the event. Then we find the oldest measurement recorded, across all slices (s) and the latest measurement. Using these two extreme measurements, EESA constructs a timeline ($TL \subseteq \mathcal{U}_\Gamma$) for resampling. This timeline has intervals ($i \in TL$) that are evenly spaced between these measurements, as seen in Figure 6a. The number of intervals to consider for a given timeline is left to the analyst to choose. Depending on the amount of variance within slices or frequency of measurements, more or less distance between intervals may be preferred. However, for the following section, we assume that 100 intervals exist within the timeline and automating this choice is left for future work.

The second step of EESA uses $i \in TL$, to find interpolated values at i for all s in the given



(a) How we construct a timeline using relative time from the aligned event e in an observation and the slice s. Measurements are noted as white circles.

(b) For a given point on the shared timeline (i), we select the leftmost and rightmost data points in slices (highlighted circles) and interpolate.

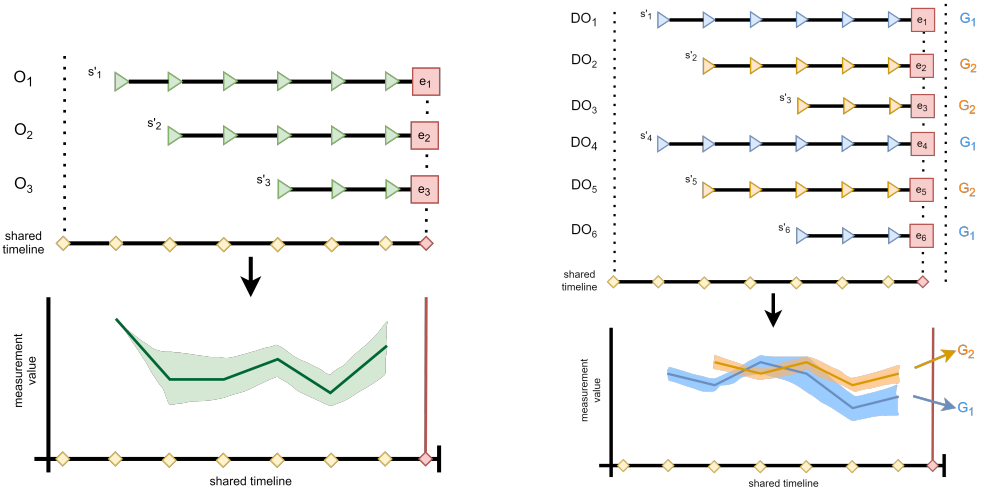
Figure 6: Construction of shared time axis (left) and interpolation of values for shared time axis (right)

observations. For a given i and s, we find the nearest earlier measurement and the nearest older measurement in s. Then using linear interpolation, we find a value at i, as shown in Figure 6b. If we cannot find a measurement for either of the points, we do not record a value from this s at i. This step generates a mapping that returns all possible interpolated values at the given i.

The final step of EESA collects statistics needed for each $i \in TL$ to create the visualisation. The visualisation consists of a line and a shaded area, the median is plotted as a line, and one standard deviation (-/+ from the median) is plotted as a shaded area. We chose the median as it is less prone to outliers, shows an actual value rather than a summary value, and is a robust measure of the central tendency [61]. These choices allow an EESA visualisation to show the evolution of slices, and where variation in measurements would likely occur across slices, as seen in Figure 7a.

4.2.1. Overlapping EESA

This section presents a variation of the EESA visualisation, where multiple trends can be presented within a single visualisation on the same axis. Our variation, an overlapping EESA visualisation, characterises each observation, allowing trends to be compared within graphs rather than between graphs, where the shared time axis and measurement range could influence



(a) How an EESA visualisation uses interpolated values for a given i to plot the median trend (solid) and one standard deviation (shaded).

(b) How an overlapping EESA visualisation uses group identifiers (G_i) to create comparable trends.

Figure 7: Demonstrations of EESA visualisation process from slices (top) to generated outcome (bottom)

interpretation. This variation adds a discrimination process of observations, a recursive step around the last step of the EESA procedure, whereby observations are extended to a quintuple, referred to as a distinguished observation (DO).

A distinguished observation is $DO = (e, s, A, X, G)$, where (e, s, A, X) is an observation, and G is a set of group identifiers for the observation. If G is limited to a size of one, this signals that DO cannot be used with replacement. However, when G is not limited to a size of one, the group identifiers are less distinctive, and DO can be used with replacement across groups. The identification of group identifiers can vary, as the process model, process behaviour, and exogenous data could inform how to characterise an observation and the number of possible groups. These group identifiers are then used to create subsets of the original collection of interpolated values, whereby each subset generates a separate EESA visualisation on the same graph.

To find group identifiers for a O, a discrimination function is introduced before the visualisation step in EESA. This function takes a process model, a partial process execution, and all data associated with the execution and finds a set of group identifiers for a O, thus creating a DO. Applying the discrimination function to each O, we create a collection of DO. Next, we created subsets of DO, where all DO have a single group identifier in common. Then each subset generates an EESA visualisation. Finally, all generated EESAs are overlapped in a single visualisation, thus creating an overlapping EESA, as seen in Figures 5b & 7b.

4.3. Highlighting Points of Interest

This section presents an automatic macro-analysis to complement EESA visualisations that provides micro-analysis of process executions. The analysis' goal is to find where distinctions exist between process outcomes across a process model using a collection of EESAs. To highlight where the most/least amount of distinction exists, we rank all EESA visualisations and assign a rank based on the amount of distinction seen. This analysis focuses on the following properties; the shape of the trends should be considered (not just the magnitude of change), usability on differing lengths of time series, and guarantees the amount of distinction. Our solution for ranking uses a scaling function to make EESA visualisations cross-compatible and Dynamic Time Warping (DTW) [62, 63] for ranking.

Our solution for ranking takes a collection of EESA as input, then for each EESA, computes a measure of distinctiveness and assigns a rank. The lower the assigned rank, the higher measure of distinctiveness between median trends observed in the visualisation. In our solution, we only measure overlapping EESAs, and all other visualisations are given a pseudo rank, the highest possible rank plus one $(n + 1)$. For example, given the collection of EESAs in Figure 9, where three EESAs are overlapping, our analysis would give each of these a numbered rank based on their distinctiveness. Then, all remaining EESAs are given the highest possible rank, which is $3 + 1$, as the last overlapping rank was three.

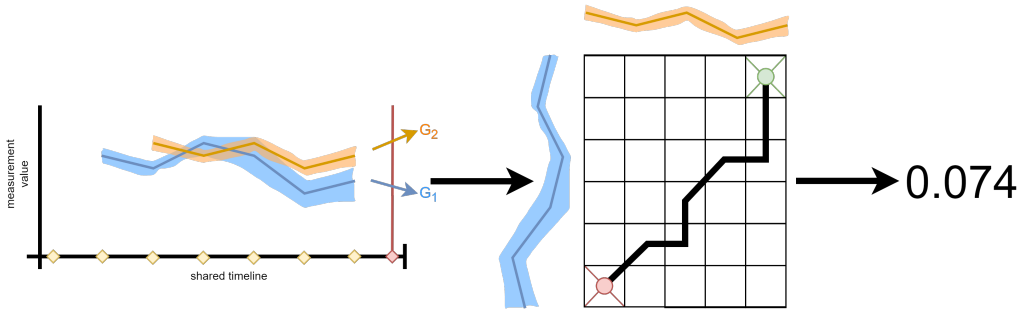


Figure 8: Measurement of an overlapping EESA graph (left), where a dynamic time warping matrix is computed between median trends (middle) then the summarisation of the path produces a measurement (right).

Formally, given a process model and an xlog, we create a collection of EESA visualisations from the transitions in the process model using the slices in an xlog. Using this collection, we start our analysis by making two subsets, one for overlapping EESAs and another for non-overlapping. For each overlapping EESA, our analysis extracts the median trend for each G_i, scales each median trend and computes a summation of DTW between medians in a pairwise manner. After each overlapping EESAs has been measured, we sort this subset such that the first element is the one with the largest summation of DTW. Finally, ranks are assigned based on this ordering for overlapping EESA $(1, 2, 3, ..., n)$, while EESAs in the second subset are given a pseudo rank $(n + 1)$.

Analysts may be interested in different ends of our ranking system, depending on the discrimination function and the analysed process. For example, using our ranking system, analysts could use the discrimination function to consider if slices agreed, or disagreed, with the associated precondition. Then using our ranking, analysts could consider the relatively highly distinctive graphs (near rank 1), the relatively least distinctive graphs (near rank n), or rank n+1 for non-overlapping EESAs to understand how exogenous data could inform decision making.

We have used DTW to compute the total difference between trends for the following reasons.

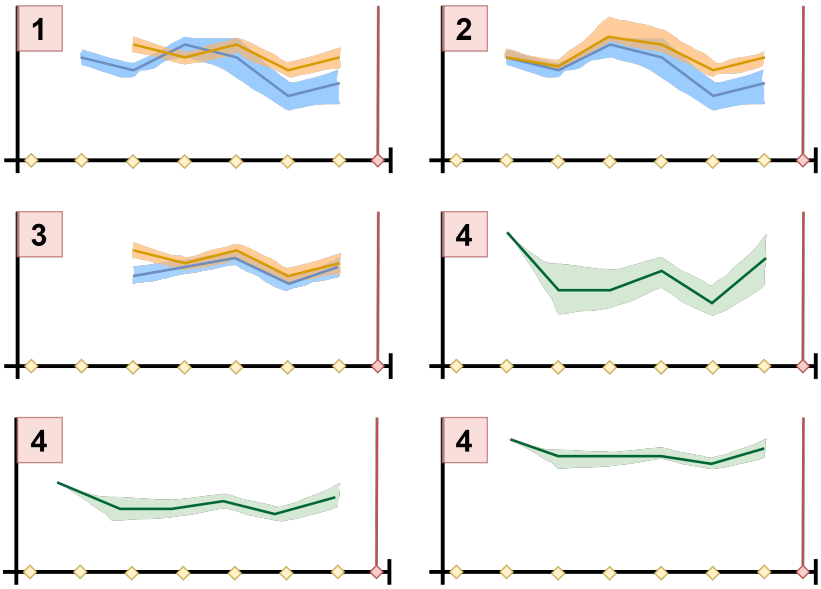


Figure 9: Example of ranking EESA visualisations, rank is denoted in red box, in top left. All non-overlapping EESA are ranked at $n + 1$ (4) while overlapping EESA are ranked based on distinctiveness between outcomes (1-3).

Firstly, the method is elastic, meaning that the two trends under comparison do not need to be equal in length. Second, the method compares the similarities between shape and time, so differences will be noted when peaks and lows occur differently. Finally, the method computes a mapping between trends through a matrix and finds the path with the least amount of resistance that can be adapted for both numerical and discrete cases, as seen in Figure 8. Furthermore, we have used min-max scaling on trends before computing DTW, as such exogenous sources with different magnitudes/measurements will be comparable through our ranking.

5. Evaluation

In this section, we evaluate both of our contributions, xPM and EESA. In section 5.1, we evaluate the model quality of models discovered by xPM. In section 5.2, we illustrate the influences of exogenous data on clinical outcomes using EESA and our ranking procedure for

a real-world healthcare setting. In both our evaluations, we used a publicly available dataset, MIMIC-III [27]. This dataset contains anonymised data about forty thousand patients who stayed in critical care units at a large tertiary care hospital between 2001 and 2012.

We have implemented an MVP interface for our framework in Java[2], which uses the Weka[3] package for machine learning, JFreeChart[4] for visualisation, and several ProM[5] plug-ins libraries. The interface was implemented as a plug-in to the ProM framework (an open-source project for process mining).

5.1. Exogenous Influences on Model Quality

This section evaluates the influence of exogenous data on the quality of discovered xDPNs, using two event logs from the MIMIC-III dataset. To evaluate this influence, we will investigate the capacity to find preconditions using decision mining techniques and the effects on model quality. Our initial proposition was that including exogenous data would increase the possibility of finding data-driven rules, as more data attributes would be available to techniques. However, the situational nature of the data will affect model quality due to a lack of consistency across traces, as a single rule will be more unlikely to cover the whole event log.

5.1.1. Procedure

We evaluate model quality in three settings: only endogenous event attributes, only exogenous event attributes (transformed attributes from xPM) and using both. First, we use (i) an event log with endogenous data attributes (*endo*). Then, we use (ii) an event log with exogenous attributes where endogenous attributes have been removed (*exo*). Finally, we use (iii) an event log with both endogenous and exogenous attributes (*endo+exo*). We applied the outlined xPM configuration below to produce exogenous attributes to compare with the endogenous case (i).

[2]https://github.com/AdamBanham/ExogenousData

[3]https://www.cs.waikato.ac.nz/ml/weka/

[4]https://www.jfree.org/jfreechart/

[5]https://www.promtools.org/doku.php

Our instantiation of xPM is as follows:

- $\mathcal{L}$ For each exogenous data set, a linking function was defined that links data sets to the patient of the trace and that occurred during the admission.
- $\mathcal{S}$ We included two slicing functions. Let $\langle e_1 \dots e_n \rangle$ be a trace. Then, for event e_i the first slicing function (S_1) finds sub-time series between events e_{i-1} and e_i, while the second slicing function (S_2) finds the sub-time series between e_1 and e_i.
- $\mathcal{T}$ We included four transformation functions: minimum, average, maximum and the cumulative sum of a Fourier transform [64].
- $\mathcal{D}$ To discover a control-flow model, we applied the Inductive Miner - infrequent [3] with path filtering of 0.25. To discover an xDPN, we applied two Data Petri Net discovery techniques: Mutually Exclusive Decision Tree (`me`) [65] and Overlapping Rules Decision Tree (`or`) [56]. These techniques each take a parameter *min instances* (mi) that sets the minimum level of observed decision point instances that support a clause in a precondition. We repeated the experiment for $mi \in \{0.05, 0.15, 0.25\}$.
- $\mathcal{E}$ was not part of this experiment.

In total, 6 xDPNs were discovered for each variation of the source event log.

We assess the quality of the discovered xDPNs using fitness, precision and determinism. For fitness, we use balanced multi-perspective conformance checking [60]. For precision, we use multi-perspective precision [35]. For determinism, we propose the following measure, which expresses the decision points in the model that are deterministic. That is, the fraction of places in the model, with two or more outgoing arcs (decision points) to transition/s, that have an

associated precondition (i.e. $\Phi(t) \neq \perp$). Formally, let $N = (P, T, F)$ be a Petri net.

$$\text{decision points or } dp(P,F) = \{p \mid p \in P \wedge |p^\bullet| \geq 2\} \tag{1}$$

$$\text{weight or } w(p,F) = \frac{|\{t \mid t \in p^\bullet \wedge \Phi(t) \neq \perp\}|}{|p^\bullet|} \tag{2}$$

$$\text{Determinism or } D(P,F) = \begin{cases} \frac{\sum_{\{p \in dp(P,F)\}} w(p,F)}{|dp(P,F)|} & \text{if } |dp(P,F)| > 0 \\ 1 & \text{otherwise} \end{cases} \tag{3}$$

A value of 1 for determinism implies that all transitions that are involved in choices in the model have preconditions, while a value of 0 indicates that no transition that is involved in a choice has a precondition.

5.1.2. Data

We created two event logs from the MIMIC-III dataset: a log of patient movements (*movements log*) and a log of procedures for respiratory failures (*procedures log*). The extraction scripts for these two event logs can be found in a public repository[6]. The *movements log* captures the movements of patients between ICU wards, and contains 24 271 traces, 290 462 events, 65 activities and 6 endogenous attributes. The *procedures log* describes the procedures that a patient received, and contains 65 traces, 610 events, 34 event classes and 4 endogenous attributes. Both logs use a exo–description including exo–panels for respiratory rate, heart rate, oxygen saturation, and arterial blood pressure measurements. The *movements log* linked to 25 684 680 exo–measurements; the *procedures log* linked to 590 285 exo–measurements.

5.1.3. Results

Table 3 shows the results. The best results for each log appear in **boldface**. When considering the *movements log*, using exogenous data only (exo) does not introduce preconditions in most cases, and henceforth the fitness and precision values are high. In cases where determinism

[6] https://github.com/adambanham/icpm2021

is greater than zero, fitness is very low but precision is competitive. We conclude that for this log, the exogenous data by itself does not suffice. For exo+endo, typically more preconditions are discovered, which lowers fitness and precision (at most 0.11 lower than endo). This decrease in quality is to be expected, as adding more preconditions means that multi-perspective measures will consider more data attributes from the event log, thus increasing the state space on which precision is based.

Table 3: Breakdown of measured model quality for discovered xDPNs across two event logs and three variants. The best measurement for an event log is highlighted in bold. In both event logs, including exogenous data lead to a higher determinism.

			Movements log			Procedures log		
Variant	Discovery	mi	fitness	precision	determinism	fitness	precision	determinism
endogenous	`me`	0.25	0.586	0.644	0.048	0.739	0.395	0.119
		0.15	0.573	**0.656**	0.108	0.739	0.392	0.119
		0.05	0.573	0.644	0.140	0.761	0.418	0.167
	`or`	0.25	0.586	0.657	0.048	**0.785**	0.393	0.131
		0.15	0.583	**0.656**	0.108	**0.785**	0.393	0.131
		0.05	0.587	0.647	0.140	0.761	0.419	0.167
endogenous	`me`	0.25	0.575	0.591	0.079	0.692	0.409	0.155
and		0.15	0.518	0.537	0.156	0.705	0.411	0.155
exogenous		0.05	0.465	0.533	**0.283**	0.717	0.459	0.238
	`or`	0.25	0.586	0.649	0.048	0.731	0.445	0.214
		0.15	0.536	0.559	0.156	0.739	0.430	0.179
		0.05	0.465	0.550	0.259	0.717	0.463	0.238
exogenous	`me`	0.25	**0.654**	0.640	0.000	0.722	0.413	0.143
		0.15	**0.654**	0.623	0.000	0.697	0.436	0.143
		0.05	0.173	0.567	0.222	0.709	0.420	0.214
	`or`	0.25	**0.654**	0.628	0.000	0.701	**0.512**	**0.274**
		0.15	**0.654**	0.639	0.000	0.697	0.425	0.143
		0.05	0.099	0.582	0.198	0.709	0.424	0.214

When considering the *procedures log*, surprisingly, larger values of the parameter mi did not always decrease the number of preconditions found as is to be expected as mi is a support threshold. We suspect that the rather small size of the procedures log and the nature of the overlapping rules (`or`) algorithm is at play here, which after building a first precondition, does not have enough observations left for a second precondition to meet the mi threshold. For this log, using the exogenous data increased the determinism and hence the number of preconditions found (exo+endo and exo vs. endo). Consequently, for endo and endo+exo, fitness goes up with mi for `me` and goes down for `or`, but in the exogenous variant these patterns are not observed. If we consider exo and endo+exo vs. endo, then fitness consistently decreases, precision consistently increases, and determinism consistently increases. We suspect that the preconditions cover a larger fraction of the increased state space than for the *movements log*, which has a greater variety of process execution.

Based on these results, we conclude that the inclusion of exogenous data can lead to greater determinism, a decrease in fitness, and a possible increase in precision. A caveat to this outcome is that the same exogenous data may have more or less impact on different processes. In some cases, exogenous data may tell analysts nothing, while in other domains, exogenous data may be the only relevant part. Furthermore, we did not investigate if the discovered preconditions provide greater insights with exogenous data; we only investigated if using additional data is beneficial to the quality of discovered process models, in a single domain.

5.2. Exogenous Influences on Process Outcomes

In this section, we demonstrate how xPM and EESA visualisations can be used to understand how changes in exogenous data influence process outcomes. In contrast to typical process outcomes from business processes (e.g., a loan application is either accepted or rejected), clinical outcomes in healthcare processes can change throughout a patient's healthcare journey several times or not at all (e.g., a patient may require assistance breathing intermittently during a

hospital stay). Thus, this evaluation uses the dynamic characterisation of slices within EESA to study clinical outcomes. Here we assume that clear distinctions within exogenous data could lead to earlier detection of adverse clinical outcomes. Furthermore, our instantiation of xPM was chosen such that we can focus on the enhancement step.

5.2.1. Clinical Outcomes in ICUs

In this evaluation, we focus on a healthcare scenario where we consider invasive procedures performed on patients admitted to an intensive care unit (ICU) at a hospital. Intensive care units (ICU) are clinical areas that are designed to manage the most critically ill patients. Critically ill patients are those that have failure of one or more organ systems that require advanced supportive therapies. Patients admitted to ICUs will require one or more intravascular accesses (i.e. intravenous cannulae, central venous catheters, arterial monitoring lines) and are typically managed with an array of monitoring and therapeutic devices.

Sepsis is a condition that is defined by serious organ dysfunction associated with a dysregulated host response, and it has been highlighted by the World Health Organization as a global health priority and a major threat to patient safety [66]. Affecting more that 50 million people each year [67], septic shock is sepsis that is associated with hypotension, lactatemia and a range of etiologies, which are most commonly associated with serious infections.

Infections (e.g., severe pneumonia) are not only among the most common reasons for patients to be admitted to ICUs, but also frequently complicate a patient's stay. Growth of microorganisms in cultures of specimens from patients are typically required to diagnose an infection. Bloodstream infections are those where microorganisms are present in the bloodstream of a patient and may be related to infections at any body site such as the respiratory (pneumonia), gastrointestinal, or genito-urinary tracts.

Recognising patients with early or incubating infections and/or predicting those who will develop subsequent infections/sepsis is a major challenge. As a result of their inherent complexity

due to severe illness and the plethora of clinical variables arising from physiologic observations and interventions, usual bedside examination and monitoring of patients will occasionally fail to detect patients at very early stages of infection/sepsis when treatment would be most efficacious. Thus, developing approaches to support decision-making around the early detection of infection/sepsis in critically ill patients is a priority research objective.

To date, technology developments in healthcare centre around the support of clinical decision-making to improve patient care and outcomes [68, 69, 70]. Process mining has recently been used to explore the process of care and the diagnostic trajectory of a disease [71, 25, 10, 26]. However, several key issues arise when technology is employed to support clinical decision making in healthcare, such as rule based systems or Bayesian neural networks [70]. These issues include the quality of data to produce accurate and reliable outputs[69, 70]; the lack of mining techniques that can provide understandable, high-level information[68, 69]; and the ethical responsibility to maintain healthcare professionals trust in those outputs [68, 72, 70]. Therefore, we investigate if our EESA contribution could fill a gap, by providing understandable information to enhance the clarity of a clinical outcome in retrospective analysis.

5.2.2. Procedure

Using this scenario, we will show that the combination of our techniques, process behaviour, and exogenous data can identify, within patient vital sign measurements, clinical outcomes related to a sepsis/infection. Our instantiation of xPM for this evaluation is outlined below.

- $\mathcal{L}$ We used the patient (`subject_id`), hospital admission (`hadm_id`) and ICU (`icustay_id`) identifiers to link all exogenous data sets (except laboratory results) to traces. Laboratory results were linked using patient and hospital admission identifiers.
- $\mathcal{S}$ We included four slicing functions: a slicing function collects all exo–measurements either 2, 4, 6 or 12 hours before an event (i.e. a procedure).
- $\mathcal{T}$ We included a single transformation function that determines if a vital sign increased or

decreased over time. The slice is used to find the line of best fit for all exo–measurements, then returns the gradient of the line as an event attribute for annotation.

$\mathcal{D}$ Instead of applying a discovery technique, we used a hand-made process model. This model used a flower-like pattern to describe the process, and repeated the same process behaviour for different admission intervals (lengths of stay) to help identify different cohorts of patients.

$\mathcal{E}$ This step is the focus of our experiment. We applied our EESA procedure and the ranking procedure, outlined in Sections 4.2 & 4.3. Using these contributions, we looked to identify points of interest between patients associated with sepsis and those that are not.

5.2.3. Data

For this evaluation, we use the MIMIC-III dataset [27], to collect ICU admissions. We included all patient admissions to an ICU ward such that: sepsis was not noted at the time of admission to the hospital (noted in `admissions` table), at admission the patient was at least 18 years old and the ICU stay lasted at least 48 hours. These requirements exclude cases of pre-existing sepsis and focus on sepsis that could be associated with the ICU stay.

We did not distinguish between ICU wards for this evaluation, resulting in 11,161 admissions being included using the `admissions` and `icustay` tables. The following procedures were used as activities from the `procedureevents_mv` table: mutli-lumen, intravenous cannula (recorded as gauge 18-22), picc line, invasive ventilation, cordis/introducer, dialysis and arterial line. Each event was split into a start and stop event as these require clinicians involvement.

The exo–description included the following exo–panels representing vital sign measurements of patients: heart rate (beats/min), respiratory rate (insp/min), non-invasive blood pressure (mmHg), blood oxygen saturation (SpO2 %) and body temperature (celsius). These exo–panels were extracted from the `chartevents` table and represent nurses updating the patient's electronic chart.

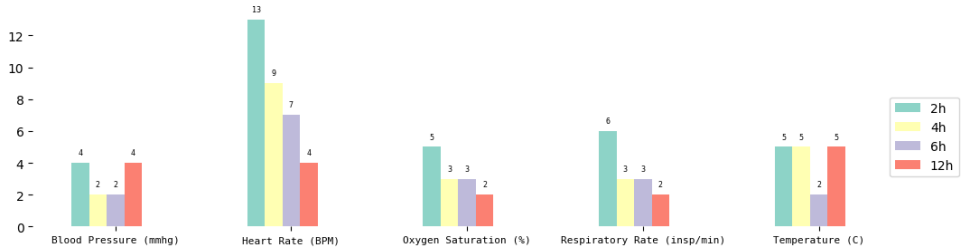


Figure 10: Histogram showing the top-ranked vital signs in EESA visualisations (only considering EESA visualisation where at least 1% of procedures were associated with sepsis).

We also included an exo–panel for laboratory results from the `microbiologyevents` table, specifically blood tests, where a blood culture was tested for the presence of organisms. We interpolated the outcome of these laboratory results to identify when a patient had a sepsis/infection during a ICU admission. In total, across six exo–panels, we extracted 111,312 exo–series and 18,014,496 exo–measurements. Our extraction scripts for both the endogenous and exogenous data from the MIMIC-III can be found in a public repository[7].

5.2.4. Results

In total, 2,240 EESA graphs were produced. In which 2028 EESAs described trends between patients with sepsis and without; hence sepsis was seen in 81.77% of all process activities within admissions. Drilling down to where sepsis occurred to a noticeable degree (i.e. at least 1% of procedures were associated with sepsis), the overall presence dropped to 36.05% of process activities seen in admissions.

The histogram in Figure 10 shows the frequency of basic vital sign measures seen in the top ranked (rank < 90) EESA visualisations, with at least 1% of procedures were associated with sepsis. In which we see the majority of EESAs with a higher level of discrimination occurred when considering the heart rate of the patient. Moreover, drilling down to the length

[7] https://github.com/adambanham/icpm2021

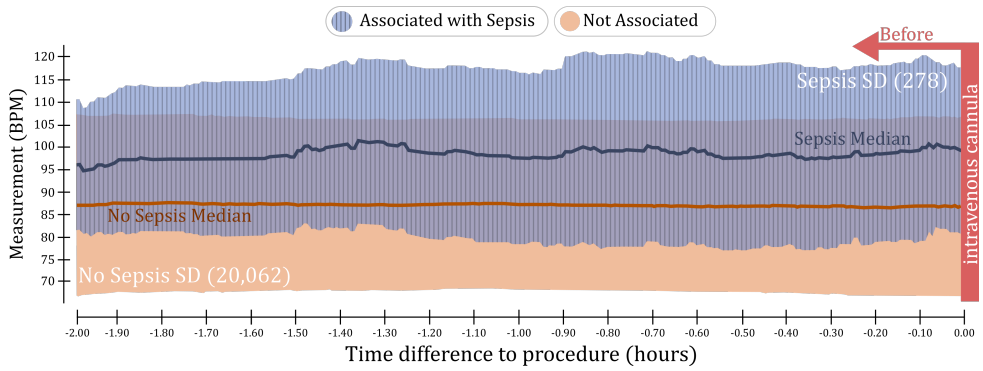


Figure 11: A top-ranked EESA of heart rate measurements linked to 20,342 intravenous cannula procedures, of which, 278 procedures were associated with a patient acquiring sepsis. Patients' heart rate measurements were taken from the two-hour window prior to the patient undergoing the intravenous cannula procedure. All procedures occurred within the first 48 hours of admission.

of observations, we can see that using two hours of vital sign measurement provides the greatest discrimination between sepsis outcomes across all vital sign measurements.

These results demonstrate that this approach can transform clinical data from being clinically uninterpretable (Figure 5a) to interpretable and having potential clinical value given further interrogation (Figure 11 & 12). In Figure 11, we can see that our EESA visualisation can convert 20,340 exo–series into a clinically interpretable visualisation. Furthermore, in Figure 12, our EESA visualisations shows how those patients who develop infection/sepsis may be discriminated based on heart rate values alone.

5.2.5. Discussion

This analysis was primarily based on standard, routinely measured, vital signs. In the course of a given day, a patient admitted to ICU may be expected to have tens to hundreds of intermittent measurements and thousands of continuous measurements. While clinicians do examine trends, typically most clinical decisions are made on "snap shot" intermittent measures. The possibility exists that patients with slow deterioration in continuously measured parameters

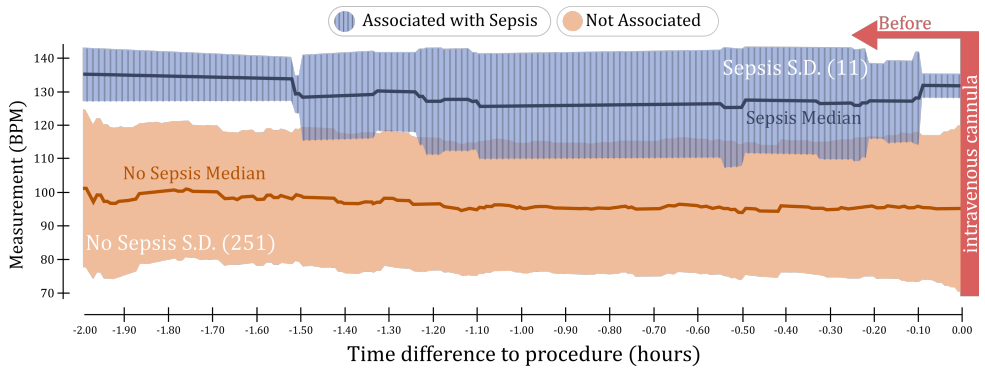


Figure 12: A top-ranked EESA of heart rate measurements linked to 262 intravenous cannula procedures, of which, 11 were associated with a patient acquiring sepsis. Patients' heart rate measurements were taken from the two-hour window prior to the patient undergoing the intravenous cannula procedure. All procedures occurred after the patient had been admitted for 25 days. We observe that heart rate may discriminate between patients associated with sepsis.

(i.e. heart rate) may not be recognised until other measures of worsening (i.e. shock onset) become overt. One of the values of our approach is that it integrates the vast number of often seemingly scattered values into patterns that can be of clinical value for decision making and flag patients for an in-depth bedside review of their clinical status.

While the processed data does not, in its current format (EESA visualisations), predict whether or not the patient will develop infection/sepsis, it is important that techniques can demonstrate the discrimination between these two outcomes. Our approach allows domain experts and process analysts to investigate a broader interpretation of the context, then refine the interpolation in an iterative manner based on data-driven outcomes. However, further studies are needed to see if our approach would help find variables that have greater discrimination for multivariable predictive modelling.

When reviewing the procedure's ranking, we noted that ranking results included sources with vastly different measurement units and variability, as seen in Figure 10. While heart rate was the most common exo–panel included in the top-ranked EESAs, as it does have large variability

naturally, the other exo–panels were evenly present in the histogram. This inclusion of exo–panels reflects our design choice to have min-max scaling occur on trends before computing the DTW distance measurement used for ranking. However, our results may reflect the distinct nature of sepsis, and the effectiveness of this design choice may not be universal. As such, future work is needed to understand if our design choice allows for an unbiased comparison of exo–panels. This also highlights the potential generalisation of our approach.

There are some limitations and assumptions about the presented analysis in this section. We selected process activities (i.e. arterial line, ventilator, insertion of an intravenous cannula) in a somewhat arbitrary fashion to explore the potential use of our approach. Furthermore, we used a very liberal understanding of the process structure (process model used for $\mathcal{D}$), to emphasise the analysis of EESA visualisations. We also assumed that data quality issues had not affected our analysis and that our extraction of vital sign measurements was representative. Finally, we assumed that our identification of patients with sepsis using a blood culture test provided a reliable understanding of this clinical outcome.

Our evaluation suggests that we can provide clarity in an otherwise noisy analysis area for healthcare professionals. Further study is needed to evaluate the feasibility of deploying our approach within a clinical setting. For example, the complexity of an approach could be a reason for non-acceptance. The adoption of an approach may be subject to a detailed cost benefit analysis. It is imperative that insights must be comprehensible by healthcare professionals in order to build trust and acceptance.

6. Conclusion

In this paper, we presented xPM and EESA, an approach for process mining with exogenous data, which analyses the associations between exogenous data and process behaviour. We evaluated our approach from two perspectives; a process mining analyst's perspective, and a clinician's perspective. We demonstrated that our proposed approach can provide automated

analysis, and can find useful insights within exogenous data for processes. Furthermore, our evaluations show that our approach can be used to integrate large amounts of physiologic data and interventions, with resulting discrimination and conversion to clinically interpretable information. This sets the stage for further studies aimed at defining a greater range of variables that may be included in studies to improve the management of patients admitted to ICUs.

We see many possibilities for extension of this work. The semantics of xDPN could be expanded to introduce means for expressing temporal changes in data, so that exogenous data can be presented alongside the process structure. Additionally, new decision mining techniques could be developed which have operators in preconditions that consider temporal sequences instead of merely event attribute values. These extensions, while clearly useful in the medical domain, would also be relevant to other domains such as business processes, transport processes and manufacturing processes. Lastly, many opportunities exist for enhancement techniques to take advantage of our framework, providing as it does, access to new types of data variables (such as slices, or the original exo–series) to provide new types of analyses.

Funding Acknowledgements

Adam Banham's work on this project was jointly funded through an Australian Government Research Training Program Scholarship and a Queensland University of Technology, Centre for Data Science Scholarship.

References

[1] W. M. P. van der Aalst, Process Mining - Data Science in Action, Second Edition, Springer, 2016.

[2] R. Shraga, A. Gal, D. Schumacher, A. Senderovich, M. Weidlich, Inductive context-aware process discovery, in: 1st International Conference on Process Mining (ICPM), IEEE, 2019, pp. 33–40.

[3] S. J. J. Leemans, D. Fahland, W. M. P. van der Aalst, Discovering block-structured process models from event logs containing infrequent behaviour, in: BPM Workshops, volume 171 of *Lecture Notes in Bus. Inf. Process.*, Springer, 2013, pp. 66–78.

[4] M. De Leoni, W. M. P. van der Aalst, Aligning event logs and process models for multi-perspective conformance checking: An approach based on integer linear programming, in: Lecture Notes in Computer Science, Springer, 2013, pp. 113–129.

[5] A. Adriansyah, Aligning observed and modeled behavior, Ph.D. thesis, Mathematics and Computer Science, 2014.

[6] F. Mannhardt, M. de Leoni, H. A. Reijers, W. M. P. van der Aalst, Data-driven process discovery - revealing conditional infrequent behavior from event logs, in: Advanced Information Systems Engineering, Springer International Publishing, 2017, pp. 545–560.

[7] G. Ibanez-Sanchez, C. Fernandez-Llatas, A. Martinez-Millana, A. Celda, J. Mandingorra, L. Aparici-Tortajada, Z. Valero-Ramon, J. Munoz-Gama, M. Sepúlveda, E. Rojas, V. Gálvez, D. Capurro, V. Traver, Toward value-based healthcare through interactive process mining in emergency rooms: The stroke case, International Journal of Environmental Research and Public Health 16 (2019) 1783.

[8] A. Partington, M. Wynn, S. Suriadi, C. Ouyang, J. Karnon, Process mining for clinical processes, ACM Transactions on Management Information Systems 5 (2015) 1–18.

[9] R. C. Basole, H. Park, M. Gupta, M. L. Braunstein, D. H. Chau, M. Thompson, A visual analytics approach to understanding care process variation and conformance, in: Proceedings of the 2015 Workshop on Visual Analytics in Healthcare - VAHC '15, ACM Press, 2015, pp. 6:1 – 6:8.

[10] F. Mannhardt, D. Blinde, Analyzing the trajectories of patients with sepsis using process mining, in: Joint Radar Tracks at the 18th International Working Conference on Business Process Modeling, Development and Support, BPMDS 2017 and the 22nd International Working Conference on Evaluation and Modeling Methods for Systems Analysis and Development, EMMSAD 2017 and the 8th International Workshop on Enterprise Modeling and Information Systems Architectures, EMISA 2017, volume 1859, CEUR-WS, 2017, pp. 72–80.

[11] H. S. G. Caballero, A. Corvo, P. M. Dixit, M. A. Westenberg, Visual analytics for evaluating clinical pathways, in: 2017 IEEE Workshop on Visual Analytics in Healthcare (VAHC), IEEE, 2017, pp. 39–46.

[12] R. A. Q. Neira, B. F. A. Hompes, J. G.-J. de Vries, B. F. Mazza, S. L. S. de Almeida, E. Stretton, J. C. A. M. Buijs, S. Hamacher, Analysis and optimization of a sepsis clinical pathway using process mining, in: Business Process Management Workshops, Springer International Publishing, 2019, pp.

459–470.

[13] E. Rojas, D. Capurro, Characterization of drug use patterns using process mining and temporal abstraction digital phenotyping, in: Business Process Management Workshops - BPM 2018 International Workshops, Lecture Notes in Business Information Processing, Springer International Publishing, 2019, pp. 187–198.

[14] G.-J. de Vries, R. A. Q. Neira, G. Geleijnse, P. Dixit, B. F. Mazza, Towards process mining of EMR data - case study for sepsis management, in: Proceedings of the 10th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2017) - Volume 5: HEALTHINF, Porto, Portugal, February 21-23, 2017, SCITEPRESS - Science and Technology Publications, 2017, pp. 585–593.

[15] H. D. Oliveira, M. Prodel, L. Lamarsalle, M. Inada-Kim, K. Ajayi, J. Wilkins, S. Sekelj, S. Beecroft, S. Snow, R. Slater, A. Orlowski, "bow-tie" optimal pathway discovery analysis of sepsis hospital admissions using the hospital episode statistics database in england, JAMIA Open 3 (2020) 439–448.

[16] A. Najjar, D. Reinharz, C. Girouard, C. Gagné, A two-step approach for mining patient treatment pathways in administrative healthcare databases, Artificial Intelligence in Medicine 87 (2018) 34–48.

[17] A. C. Apunike, L. Oliveira-Ciabati, T. L. M. Sanches, L. L. de Oliveira, M. N. Sanchez, R. M. Galliez, D. Alves, Analyses of public health databases via clinical pathway modelling: TBWEB, in: Computational Science - ICCS 2020 - 20th International Conference, Amsterdam, The Netherlands, June 3-5, 2020, Proceedings, Part IV, Springer International Publishing, 2020, pp. 550–562.

[18] E. Rojas, J. Munoz-Gama, M. Sepúlveda, D. Capurro, Process mining in healthcare: A literature review, Journal of Biomedical Informatics 61 (2016) 224–236.

[19] N. Martin, J. D. Weerdt, C. Fernández-Llatas, A. Gal, R. Gatta, G. Ibáñez, O. Johnson, F. Mannhardt, L. Marco-Ruiz, S. Mertens, J. Munoz-Gama, F. Seoane, J. Vanthienen, M. T. Wynn, D. B. Boilève, J. Bergs, M. Joosten-Melis, S. Schretlen, B. V. Acker, Recommendations for enhancing the usability and understandability of process mining in healthcare, Artificial Intelligence in Medicine 109 (2020) 101962.

[20] A. Guzzo, A. Rullo, E. Vocaturo, Process mining applications in the healthcare domain: A comprehensive review, WIREs Data Mining and Knowledge Discovery (2021).

[21] E. D. Roock, N. Martin, Process mining in healthcare – an updated perspective on the state of the art, Journal of Biomedical Informatics (2022) 103995.

[22] E. Rojas, M. Sepúlveda, J. Munoz-Gama, D. Capurro, V. Traver, C. Fernandez-Llatas, Question-driven methodology for analyzing emergency room processes using process mining, Applied Sciences 7 (2017) 302.

[23] C. Alvarez, E. Rojas, M. Arias, J. Munoz-Gama, M. Sepúlveda, V. Herskovic, D. Capurro, Discovering role interaction models in the emergency room using process mining, Journal of Biomedical Informatics 78 (2018) 60–77.

[24] K. S. Cosby, R. Roberts, L. Palivos, C. Ross, J. Schaider, S. Sherman, I. Nasr, E. Couture, M. Lee, S. Schabowski, I. Ahmad, R. D. Scott, Characteristics of patient care management problems identified in emergency department morbidity and mortality investigations during 15 years, Annals of Emergency Medicine 51 (2008) 251–261.e1.

[25] R. S. Mans, W. M. P. van der Aalst, R. J. B. Vanwersch, A. J. Moleman, Process mining in healthcare: Data challenges when answering frequently posed questions, in: Lecture Notes in Computer Science, volume 7738, Springer Berlin Heidelberg, 2013, pp. 140–153.

[26] J. Munoz-Gama, et. at. al., Process mining for healthcare: Characteristics and challenges, Journal of Biomedical Informatics (2021).

[27] A. E. Johnson, T. J. Pollard, L. Shen, L. wei H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, R. G. Mark, MIMIC-III, a freely accessible critical care database, Scientific Data 3 (2016).

[28] A. Banham, S. J. J. Leemans, M. T. Wynn, R. Andrews, xPM: A framework for process mining with exogenous data, in: J. Munoz-Gama, X. Lu (Eds.), Process Mining Workshops - ICPM 2021 International Workshops, volume 433 of *Lecture Notes in Business Information Processing*, Springer, 2021, pp. 85–97.

[29] E. Batista, A. Solanas, Process mining in healthcare: A systematic review, in: 9th International Conference on Information, Intelligence, Systems and Applications, IISA 2018, Zakynthos, Greece, July 23-25, 2018, IEEE, 2018, pp. 1–6.

[30] M. Rosemann, J. Recker, C. Flender, Contextualisation of business processes, Inter. J. of Business Process Integration and Management 3 (2008) 47–60.

[31] W. M. P. van der Aalst, S. Dustdar, Process mining put into context, IEEE Internet Computing 16 (2012) 82–86.

[32] J. De Smedt, S. K. L. M. van den Broucke, J. Obregon, A. Kim, J.-Y. Jung, J. Vanthienen, Decision

mining in a broader context: An overview of the current landscape and future directions, in: BPM Workshops, Springer, 2017, pp. 197–207.

[33] K. Diba, K. Batoulis, M. Weidlich, M. Weske, Extraction, correlation, and abstraction of event data for process mining, WIREs Data Mining and Knowl. Discovery 10 (2019).

[34] A. E. Marquez-Chamorro, K. Revoredo, M. Resinas, A. Del-Rio-Ortega, F. M. Santoro, A. Ruiz-Cortes, Context-aware process performance indicator prediction, IEEE Access 8 (2020) 222050–222063.

[35] F. Mannhardt, Multi-perspective process mining, Ph.D. thesis, Technische Universiteit Eindhoven, 2018.

[36] A. Senderovich, C. Di Francescomarino, F. M. Maggi, From knowledge-driven to data-driven inter-case feature encoding in predictive process monitoring, Inf. Systems 84 (2019) 255–264.

[37] H. Nguyen, M. Dumas, M. L. Rosa, A. H. M. ter Hofstede, Multi-perspective comparison of business process variants based on event logs, in: Conceptual Modeling, Springer International Publishing, 2018, pp. 449–459. doi:10.1007/978-3-030-00847-5_32.

[38] J. Cremerius, M. Weske, Supporting domain data selection in data-enhanced process models, in: 17th International Conference on Wirtschaftsinformatik, Proceedings. 3, Association for Information Systems, 2022. URL: https://aisel.aisnet.org/wi2022/bpm/bpm/3.

[39] M. Dees, B. Hompes, W. M. P. van der Aalst, Events put into context (EPiC), in: 2nd International Conference on Process Mining (ICPM), IEEE, 2020, pp. 65–72.

[40] S. J. Leemans, K. Goel, S. J. van Zelst, Using multi-level information in hierarchical process mining: Balancing behavioural quality and model complexity, in: 2020 2nd International Conference on Process Mining (ICPM), IEEE, 2020, pp. 137–144.

[41] M. Pesic, H. Schonenberg, W. M. van der Aalst, DECLARE: Full support for loosely-structured processes, in: 11th IEEE International Enterprise Distributed Object Computing Conference (EDOC 2007), IEEE, 2007, pp. 287–287.

[42] S. Schönig, C. D. Ciccio, F. M. Maggi, J. Mendling, Discovery of multi-perspective declarative process models, in: Service-Oriented Computing, Springer International Publishing, 2016, pp. 87–103.

[43] V. Leno, M. Dumas, F. M. Maggi, M. L. Rosa, A. Polyvyanyy, Automated discovery of declarative process models with correlated data conditions, Information Systems 89 (2020) 101482.

[44] S. Montani, G. Leonardi, S. Quaglini, A. Cavallini, G. Micieli, Improving structural medical process comparison by exploiting domain knowledge and mined information, Artificial Intelligence in Medicine 62 (2014) 33–45.

[45] Y. Xue, D. Klabjan, Y. Luo, Predicting ICU readmission using grouped physiological and medication trends, Artificial Intelligence in Medicine 95 (2019) 27–37.

[46] Z. Huang, X. Lu, H. Duan, On mining clinical pathway patterns from medical behaviors, Artificial Intelligence in Medicine 56 (2012) 35–50.

[47] Z. Huang, X. Lu, H. Duan, W. Fan, Summarizing clinical pathways from event logs, Journal of Biomedical Informatics 46 (2013) 111–127.

[48] Z. Huang, Z. Ge, W. Dong, K. He, H. Duan, Probabilistic modeling personalized treatment pathways using electronic health records, Journal of Biomedical Informatics 86 (2018) 33–48.

[49] T. Huang, B. Xu, H. Cai, J. Du, K.-M. Chao, C. Huang, A fog computing based concept drift adaptive process mining framework for mobile APPs, Future Generation Computer Systems 89 (2018) 670–684.

[50] W. M. P. van der Aalst, J. Carmona (Eds.), Process Mining Handbook, volume 448 of *Lecture Notes in Business Information Processing*, Springer, 2022. doi:10.1007/978-3-031-08848-3.

[51] M. De Leoni, P. Felli, M. Montali, A holistic approach for soundness verification of decision-aware process models, in: Concept. Model., Springer, 2018, pp. 219–235.

[52] P. Felli, M. De Leoni, M. Montali, Soundness verification of decision-aware process models with var.-to-var. conditions, in: 2019 19th Int. Conf. on ACSD, IEEE, 2019, pp. 82–91.

[53] E. Gonzalez Lopez de Murillas, Process mining on databases: extracting event data from real-life data sources, phdthesis, Mathematics and Computer Science, 2019.

[54] S. Chaudhuri, U. Dayal, An overview of data warehousing and OLAP technology, SIGMOD Rec. 26 (1997) 65–74.

[55] F. Mannhardt, M. De Leoni, H. A. Reijers, The multi-perspective process explorer, in: BPM Conference Demos, volume 1418, CEUR-WS, 2015, pp. 130 – 134.

[56] F. Mannhardt, M. de Leoni, H. A. Reijers, W. M. P. van der Aalst, Decision Mining Revisited - Discovering Overlapping Rules, in: Advanced Information Systems Engineering, Springer International Publishing, Cham, 2016, pp. 377–392.

[57] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: I. Guyon,

U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30, Curran Associates, Inc., 2017, pp. 4765–4774.

[58] M. T. Ribeiro, S. Singh, C. Guestrin, "why should I trust you?": Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016, ACM, 2016, pp. 1135–1144.

[59] C. Moreira, Y. Chou, M. Velmurugan, C. Ouyang, R. Sindhgatta, P. Bruza, LINDA-BN: an interpretable probabilistic approach for demystifying black-box predictive models, Decis. Support Syst. 150 (2021) 113561.

[60] F. Mannhardt, M. de Leoni, H. A. Reijers, W. M. P. van der Aalst, Balanced multi-perspective checking of process conformance, Computing 98 (2015) 407–437.

[61] R. S. Rajan Chattamvelli, Statistics for Scientists and Engineers, WILEY, 2015.

[62] H. Sakoe, S. Chiba, Dynamic programming algorithm optimization for spoken word recognition, IEEE Transactions on Acoustics, Speech, and Signal Processing 26 (1978) 43–49.

[63] S. Aghabozorgi, A. S. Shirkhorshidi, T. Y. Wah, Time-series clustering – a decade review, Information Systems 53 (2015) 16–38.

[64] D. Griffin, J. Lim, Signal estimation from modified short-time fourier transform, IEEE Trans. on Acoustics, Speech, and Signal Processing 32 (1984) 236–243.

[65] M. De Leoni, W. M. P. van der Aalst, Data-aware process mining: discovering decisions in processes using alignments, in: Proceedings of the 28th Annual ACM Symposium on Applied Computing, SAC '13, Coimbra, Portugal, March 18-22, 2013, ACM, 2013, pp. 1454–1461.

[66] K. Reinhart, R. Daniels, N. Kissoon, F. R. Machado, R. D. Schachter, S. Finfer, Recognizing sepsis as a global health priority — a who resolution, New England Journal of Medicine 377 (2017) 414–417.

[67] L. J. Schlapbach, N. Kissoon, A. Alhawsawi, M. H. Aljuaid, R. Daniels, L. A. Gorordo-Delsol, F. Machado, I. Malik, E. F. Nsutebu, S. Finfer, K. Reinhart, World sepsis day: a global agenda to target a leading cause of morbidity and mortality, American Journal of Physiology-Lung Cellular and Molecular Physiology 319 (2020) L518–L522.

[68] K. Kawamoto, C. A. Houlihan, E. A. Balas, D. F. Lobach, Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success, BMJ

330 (2005) 765.

[69] D. F. Sittig, A. Wright, J. A. Osheroff, B. Middleton, J. M. Teich, J. S. Ash, E. Campbell, D. W. Bates, Grand challenges in clinical decision support, Journal of Biomedical Informatics 41 (2008) 387–392.

[70] J. Varghese, M. Kleine, S. I. Gessner, S. Sandmann, M. Dugas, Effects of computerized decision support system implementations on patient outcomes in inpatient care: a systematic review, Journal of the American Medical Informatics Association 25 (2017) 593–602.

[71] R. S. Mans, M. H. Schonenberg, M. Song, W. M. P. van der Aalst, P. J. M. Bakker, Application of process mining in healthcare – a case study in a dutch hospital, in: Biomedical Engineering Systems and Technologies, Springer Berlin Heidelberg, 2008, pp. 425–438. doi:10.1007/978-3-540-92219-3_32.

[72] A. Holzinger, Interactive machine learning for health informatics: when do we need the human-in-the-loop?, Brain Informatics 3 (2016) 119–131.