Evaluation of Subpopulation Process Comparison Techniques for Process Mining

VICTOR-ADRIAN ALECU, University of Twente, The Netherlands

Process Mining (PM) has emerged as a vital discipline at the intersection of business process management and data analysis, offering insights into organisational processes from event log data. From this perspective, Conformance Checking (CC) is a key component, which checks if real-life executions align with intended process models, thus helping to maintain compliance and operational integrity. If the analysis is instead directed towards disjoint subgroups of a data set, subpopulation process comparison (SPC) focuses CC on identifying the procedural similarities and differences between such segments. This paper aims to serve as a comprehensive comparative study, focusing on how various CC methods can be utilised to conduct subpopulation process comparison, in order to address existing literature gaps. By examining the unique characteristics of these methods, assessing their differences, and exploring their collective efficacy in real-world applications, this study seeks to enhance the understanding of SPC. The goal is to evaluate the comparative strengths and weaknesses of these methods and investigate potential synergies when applied concurrently.

Additional Key Words and Phrases: Process Mining, Subpopulation Process Comparison, Conformance Checking, PM^2

1 INTRODUCTION

From its inception in the late 1990s, Process Mining (PM) has solidified itself as a leading discipline in the fields of business process management (BPM) and data analysis. Originating from the need to extract actionable knowledge from large amounts of event log data produced by various business systems, this field of study bridges the gap between traditional business process analysis and modern data-centric techniques [1].

Conformance checking (CC), a fundamental aspect of PM, verifies whether the actual execution of a process adheres to a predefined, "reference" business model [2]. This procedure plays an important role for organisations aiming to ensure compliance, enhance operational efficiency, and minimise deviations that could lead to inefficiencies or risks [41].

As shown in previous works [39, 42, 47, 52], a useful application of CC methods is the analysis of subgroups within event logs via subpopulation process comparison (SPC). This process entails identifying distinct groups or segments within a data set and comparing their process executions. By examining how these subpopulations adhere to, or deviate from, each other's process models, analysts can gain insights into specific areas that may require targeted interventions or optimisations.

This study aims to contribute to the existing body of work regarding the applicability of conformance checking for subpopulation process comparison. The main goal is to provide a comprehensive evaluation of existing CC methods within the SPC context, highlighting differences and potential synergies. To meet this objective, the remainder of the paper is structured as follows. Section 2 presents the problem statement and research questions guiding this study, while section 3 covers the necessary background information related to process mining. Section 4 describes the methodology of the study and how each sub-research question was addressed, while the results are showcased in section 5. Finally, section 6 discusses potential limitations of the research along with avenues for future work, while section 7 concludes the paper.

2 PROBLEM STATEMENT

With conformance checking being considered one of the three main disciplines of process mining (along with process discovery and enhancement) [2] it has experienced considerable development since its inception. As proof, various research efforts have been conducted over the years to develop faster and more accurate process comparison models. Studies have either introduced new CC approaches [9, 25, 30, 35] or augmented existing techniques [13, 14] in order to tackle larger and more complex cases. Most often, these advancements are evaluated in the traditional context of conformance checking – primarily focusing on their ability to compare event log data against a "global" reference model. Hence, their applicability in subpopulation process comparison is not as popular of a validator. This particular use case involves assessing the behaviour of various sub-groups within the original data set, presenting potentially different requirements than the traditional approach. Methods effective in general conformance checking may not adequately address the complexities or capture the nuanced variations across subpopulations, possibly leading to less accurate or insightful results when applied to this more segmented analysis.

Additionally, most literature introducing CC methods seem to present them as standalone solutions, often only comparing them against the existing standards of conformance checking. Broader evaluations considering multiple approaches seem to be relatively rare, the literature review associated with this study managing to find only one such work [14].

2.1 Research Question

This study aims to bridge the identified literature gaps by providing an examination of the unique characteristics of various CC methods, assessing how they differ, and identifying their comparative strengths and weaknesses in the context of subpopulation process comparison. These goals have been formalised in the following research question:

RQ: How can an SPC-focused evaluation of conformance checking methods be conducted, and what insights can be derived from it?

This overarching research question will be explored through two focused sub-research questions:

SRQ1: What are some of the current CC methods used in Process Mining, and what unique characteristics and functionalities do they offer?

TScIT 41, July 5, 2024, Enschede, The Netherlands

SRQ2: What are the comparative strengths and limitations of the identified CC methods in the context of subpopulation process comparison, particularly when applied to complex event logs derived from real-world scenarios?

These sub-research questions are addressed through an exploratory study and a practical experiment, respectively. Before presenting these parts of the research, the following section will provide background information on the fundamentals of process mining, necessary for understanding the concepts discussed further in this paper.

3 BACKGROUND

Integral to most PM operations are the data sources they interact with, often organised in the form of an event log. Since this study utilised data sources following the XES standard [53], the following description will refer to event logs formatted as such. Nevertheless, the main differences between PM log standards are more structural rather than conceptual [3], so the following definitions should be reasonably universal.

Event logs capture the sequence of activities within an organisation's process, giving insights into the operational process flow [4]. Entries in a log, referred to as traces (or cases), represent a single instance of a process, such as handling a request or executing a transaction. Every trace comprises a sequence of events, where each event corresponds to a specific activity. Additionally, events can be accompanied by attributes which describe the action in more detail: time of occurrence, resources involved (e.g., personnel or equipment), and other relevant data which may impact the analysis. Such attributes can also be present at the trace level, where they characterise the entire process flow instance.

Another type of input common to PM procedures is the process model. These structures condense workflow-related information, either obtained from stakeholder analysis or mined from an event log, into graphical representations. Following the definition of [22], such models are designed to aid communication among stakeholders, presenting the process flow in a structured and approachable manner. Hence, regardless of the modelling language (BPMN, Petri Nets, etc.), the "ordering of activities is modelled by describing causal dependencies" [5]. Every unique event of the log has an associated node in the graph, while the event sequences describe the arcs between vertices.

Conformance checking, a specialised branch of process mining, analyses the aforementioned event logs and process models with the goal of highlighting their similarities and differences [2, 6]. This traditional CC use case focuses on comparing "observed behaviour", represented by the log, against "modelled behaviour", illustrated in the model constructed through stakeholder consultation. By contrasting the two, conformance analysis can identify deviations [9, 14, 25, 30], find bottlenecks [42] or help repair models that do not accurately reflect reality [6].

When the evaluation is conducted on event logs or models which represent subsets of the data set, conformance checking is focused towards the particular use case of SPC. The subgroups may embody traces from a particular department, type of process, or population demographic. Subpopulation process comparison involves comparing these distinct segments in order to uncover variations and patterns that may not be apparent when examining the entire dataset as a whole. This assessment is achievable through an adjustment of CC methods, where instead of comparing against a reference model, the two subgroups are cross-evaluated. By treating the event log of one subpopulation as the "observed" behaviour, while its counterpart is considered the "modelled behaviour", CC can reveal how closely each group adheres to the processes of the other.

To better frame the SPC use case, the following example is given. A log containing operations of a hospital might have its traces describe the treatment path associated with a particular patient. These traces could be split according to the treatment or diagnosis (possibly marked via trace attributes), with flu patients being organised into one subpopulation and appendicitis patients into another. CC techniques can be later applied to compare these subpopulations, identifying differences in the control flow (different treatment paths), duration (lengthier/shorter treatment), or resources (medical equipment used).

4 METHODOLOGY

The methodology which will guide this study will be the one proposed by [23], called "Process Mining Project Methodology" (PM^2). Compared to other Data Science methodologies like CRISP-DM [27] or SEMMA [28], which are high-level and provide little guidance for process mining specific studies, PM^2 has been specifically designed for process mining research. The more focused approach, along with the clearly defined stages with particular inputs/outputs, make this methodology a suitable candidate for this PM-specific research. However, the most important benefit of PM^2 is that its structure illustrates a clear differentiation between the initial, more theoretical, "Initialization" phase and subsequent practical "Analysis iterations" (see Figure 1). This distinction aligns effectively with the two sub-research questions. SRQ1 focuses on the theoretical exploration of existing conformance checking and subpopulation process comparison methods, while SRQ2 involves a practical comparative evaluation on real-world data. In order to better address SRQ1, the initialisation phase has been expanded with a supplementary "Exploration" stage. This addition aims to accommodate the research required for identifying and classifying the SPC and CC methods this study aims to evaluate. Furthermore, the final stage of "Process Improvement & Support" was skipped. The main reason is that this research was not conducted in cooperation with any stakeholder nor aimed at finding solutions to particular issues.

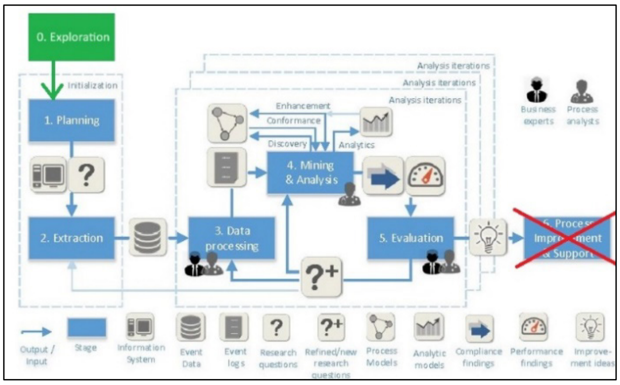

Fig. 1. Structure of the adapted PM2 Methodology

4.1 Exploration

Before evaluating the suitability of conformance checking methods in regards to subpopulation process comparison, an exploration of existing approaches was conducted. It consisted of a literature review aiming to identify the most popular CC methods, as well as emerging techniques which have shown potential in the SPC use case.

This section provides an examination of the explored CC approaches, highlighting their unique characteristics, differences, and potential applicability within the context of subpopulation process comparison. For a more concise overview, Figure 2 presents a Venn Diagram of the analysed CC methods, organized according to the approaches they employ, and Table 1 summarizes relevant information for each method.

4.1.1 Token-based replay

Historically, the field of conformance checking in process mining started with the introduction of token-based replay by [50]. These early methods relied on simulating the replay of individual traces from the event log onto the model with the use of tokens. Such tokens would mark each step taken through the replay, while also keeping track of "missing" or "additional" steps. After the replay, several similarity measures would be offered such as fitness or appropriateness. The former measures how much log traces match valid execution paths in the process model, while the latter indicates how accurately the process model describes the behaviour in the log.

Even though they tend to perform well on contemporary CC cases, traditional token replay approaches have been shown to typically overestimate fitness values. As [8] states, such methods cannot handle cases of models and logs which don't fit well, especially in the presence of "invisible" or "duplicate" activities. In these more complex scenarios, the early token-based replay techniques require additional heuristics and state space exploration.

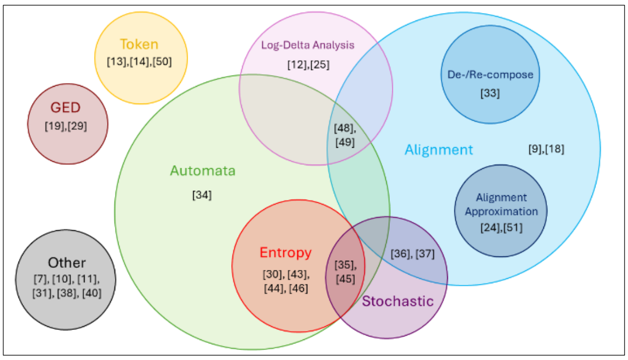

Fig. 2. Venn Diagram of investigated CC methods

Considering the above predicament, token-based approaches could be considered unsuitable for subpopulation process comparison, given that no guarantee can be made regarding the similarity of subgroup processes. Nevertheless, subsequent advancements have reinvigorated this approach, making it not only a suitable candidate for traditional conformance checking, but possibly also for SPC. Works like [13, 14] present an enhanced version of token replay which introduces a dedicated pre-processing step as well as certain caching techniques, thus solving the issues of the previous iterations. These studies have also shown the performance advantages this approach has over other CC methods such as the ones based on alignments or automata. Still, this comparison has been done on the classic CC use case, with SPC being a yet unexplored context.

Overall, considering the improvements in fitness and performance brought by later works, token-based replay is a suitable candidate for evaluation in the subpopulation process comparison scenario. One drawback of this approach, however, is its relative lack of explainability associated with the software tool implementing the method (PM4Py). With only numerical similarity measures being provided, this technique could benefit from the visualisation mechanisms of other CC approaches such as alignments or GED-based.

Table 1. Overview of investigated CC methods

Ref.	Description	Approach	Output	Software
[50]	Replays traces on a Petri Net keeping count of consumed/produced tokens	Token Replay	Token-related metrics Fitness	ProM 5
[13,14]	Augmented version of [50] with pre-processing and activity caching	Token Replay	Token-related metrics Fitness	PM4Py
[19,29]	Computes # of insert/delete operations necessary to get from one model to another	GED	# insert/delete operations Model highlighting deviations	BPMNDiffViz
[12,25]	Compares event structures using an error-correcting Synchronised Product	Log-Delta	Behavioural difference statements (natural language)	Apromore
[30,44,46]	Measures the similarity between the languages of automata derived from model or event log	Entropy	Precision & Recall (Exact/Partial)	Entropia
[45]	Measures avg. # of bits used to compress log traces using the relative likelihoods induced by the model	Entropy, Stochastic	Entropic Relevance	Entropia
[35]	Compares entropy of SDFAs derived from input to the entropy of a third, conjunction SDFA	Entropy, Stochastic	Stochastic Precision and Recall	Entropia
[36,37]	Trace alignments over stochastic models.	Stochastic	Similarity score	ProM 6
[34]	Projects both model and log on subsets of activities, evaluating reflection of captured behaviour	Automata	Fitness, Recall, Precision	ProM 6
[49]	Compares reachability graph and automata using error-correcting Synchronised Product	Automata, Alignment	Optimal alignments and natural language statements	Apromore
[48]	Optimizes [49] using S-components obtained from decomposing the original input	Automata, Alignment	Optimal alignments and natural language statements	Apromore
[9,18]	Aligns event log and model in order to construct "best matching instance"	Alignment	Fitness, Precision, Generalization, and per-trace or per-log projections	ProM 6
[33]	Split process model into sub-nets, perform alignment, recompose results	Alignment	Fitness and Raw Fitness Cost	ProM 6
[24,51]	Alignment approximation	Alignment	Fitness, Precision, Generalization	ProM 6

4.1.2 Alignment-based approaches

With the advent of token-based replay methods, the field of conformance checking expanded towards finer-grained and more explainable approaches. Certain deviations might be more costly than others, thus needing to be penalised more harshly and also, possibly be easier to identify. To this end, alignment-based approaches have been introduced by [9], employing a shortest-path algorithm to align event log traces against the process model, aiming to construct the "best matching instance". Such alignments were based on an initial cost function provided by the user, allowing them to set the "weight" of certain deviations, thus guiding the conformance checking. Upon aligning all the traces, this method specifies two main similarity measures, which quantify how much of the log behaviour is captured by the model (fitness) and how well the model allows only the necessary behaviour, avoiding underfitting (precision). Additionally, measures such as generalisation can help to estimate the extent to which unobserved, but likely possible behaviour is explained by the model. The presented optimisations, finer-grained similarity measures, and the introduction of per-trace as well as per-log mechanisms to visualise deviations (such as alignment matrices showcased in [18]) have solidified alignment-based approaches as the "status quo" of conformance checking.

Nevertheless, the state space exploration required to find the best-matching alignment instance can prove computationally expensive, especially for large event logs or complex process models. This has been a generally accepted drawback of alignment-based methods [9, 14] with efforts being conducted to either employ more feasible algorithms or to simplify the problem. Among the latter, this study has identified two notable approaches, the first of which is introduced in [33]. The referenced paper presents a divide-and-conquer approach to CC, which decomposes the event log and process model into sub-structures. These smaller constructs are then aligned, with the results from each operation being recomposed into a "global" alignment, describing the original input. Such decomposition-based methods have been validated for large event logs and models. In these cases, the overhead introduced by de- and re-composing the input is mitigated by the overall performance improvements associated with computing smaller-scale alignments. Nevertheless, this approach is presented as situational, considering that not all models are easily decomposable. Additionally, even if decomposition is possible, the appropriateness of the initial segmentation has a great influence on the accuracy of the returned similarity measure.

The second alignment-derived technique is introduced by [24, 51] which propose an approximation of the similarity measures provided by traditional alignment methods. This approximation is computed by having the process model generate a series of simulation traces (given probabilities of event sequences) which are then compared against the traces of the input event log. Although a fast alternative to traditional approaches, this method proves to be imprecise when confronted with logs containing repetitive or highly variable behaviour.

Taking into account the status that alignment-based methods hold within the field of CC, they serve as natural candidates for an experimental evaluation for SPC. Their importance is further validated by their finer-grained perspective, coupled with high levels of explainability offered by the mature software implementation in ProM. Finally, the existence of relatively feasible alternatives which deal with large and complex inputs is also a notable advantage.

4.1.3 Graph Edit Distance

Another family of methods with a high degree of explainability is the one based on Graph-Edit-Distance, notable representatives being the works by [19, 29]. The approaches proposed by these studies treat the process models (either obtained from the input or mined from the event log) as mathematical graphs. Every unique event of the log has an associated node in the graph, while the event sequences describe the arcs between vertices. Based on these, a structural similarity measure is derived from the number of insert/delete operations on nodes/edges necessary to transform one graph into the other. This explicit identification of discrepancies lends a high degree of explainability to the technique, [29] showing how the associated software tool (BPMNDiffViz) highlights the differences directly on the process model.

Even though works like [39, 42, 47, 52] have shown the efficacy of GED-based methods in identifying discrepancies between subpopulations, comparison with metrics from other CC approaches is unfeasible. Since most similarity measures are expressed as ratios between 0 and 1, they cannot be easily matched against the integer-based scoring of GED approaches. While potential solutions to normalize the GED metric exist, such transformations (along with evaluations of their accuracy) are beyond the scope of this study.

4.1.4 Automata

Providing an alternative perspective to CC, automata-based techniques base their evaluation on state machines and the regular languages they define.

One such automata-based approach is presented by [34], which introduces the Projected Conformance Checking framework, along with an associated ProM extension. This method provides three similarity measures - fitness, precision, and recall - to assess the similarity between an input event log and process model. By projecting both the model and the log onto subsets of activities, it evaluates how well the model reproduces and restricts the behaviour observed within these subsets. This subdivision simplifies the CC task and helps pinpoint specific points of deviation and compliance in the model's behaviour. Still, the accuracy of the output heavily depends on the optimal selection of subsets, with potential trade-offs between precision and recall if certain behaviour in the process model is not well captured.

A different way of approaching automata-based conformance checking is showcased by [49]. This method departs from traditional similarity measures in favour of more descriptive outputs, supported by an extension of the Apromore software. The proposed technique combines automata (derived from the input log and model) into an error-correcting partial synchronised product. Such a construct is then used to extract either a set of optimal alignments between the two inputs, or a series of natural-language behavioural difference statements. Possible examples are "Activity X was planned in the model but skipped in the log", or "The model shows a choice between activities A and B, while the log enforces activity A". Presented as a more stakeholder-friendly output, natural-language statements can lend the method a higher degree of explainability. Nevertheless, difficulty in handling models with high levels of concurrency or loop structures presents a considerable drawback for this approach.

Providing an extension which overcomes the stated limitations, [48] introduces a divide-and-conquer perspective to automata-based CC. This addition decomposes the model into smaller automata (S-components), improving manageability of combinatorial state explosion inherent to models exhibiting high concurrency. It is presented as an overall more flexible and scalable approach, although one which might benefit from tighter heuristics to improve accuracy on certain types of process models.

Despite not having been evaluated from the perspective of SPC, automata-based methods show considerable potential in being a highly explainable conformance checking solution. Even though not all of them offer a quantifiable similarity measure, the natural-language alternatives could serve as useful complements to numerical values provided by more traditional approaches.

4.1.5 Behavioural Alignment (Log-Delta analysis)

Continuing with the concept of natural language outputs, there have been certain papers approaching this area of conformance checking from an alternative direction. Works such as [12, 25] set aside automata-based evaluations in favour of a different data structure, known as an "event structure". This construct is a directed acyclic graph where nodes represent event occurrences sharing a common history, being obtained from lossly encoding an event log or a process model. After encoding, event structures are aligned via an error-correcting partial synchronised product, which helps to identify all behavioural differences. Subsequently, these deviations are formalised into natural-language statements, similar to the automata-based approach of [48, 49], both methods being supported by Apromore. Additionally, besides control-flow evaluations, log-delta analysis can also examine the frequency with which event sequences occur. This alternative perspective might prove useful in the case of SPC, further validating the approach.

Yet, similar to the aforementioned automata-based methods, behavioural alignment does not compute a similarity score for the evaluated input. This fact distances log-delta analysis from being a "standalone" conformance checking method, presenting it as better suited to be an auxiliary to more mature alternatives. Besides, [12] has shown that it is best employed for smaller-scale, simpler models, as the approach does not scale well nor is it particularly good at handling models with high levels of concurrency or cyclic behaviour.

4.1.6 Entropy (Information Theory)

Information Theory has also been identified as a relatively popular approach to conformance checking. The literature review has encountered various methods, most of which focus on either the concept of entropy or stochastic conformance. Among the former, there are two main areas of development, with the first being represented by [30, 44, 46]. These works introduce the concept of behavioural quotients for comparing system behaviour, which help define precision and recall values between recorded executions and process models. Such quotients are obtained from the comparison of languages defined by automata (derived from the input log and model) and help characterise the behaviour allowed by these structures. While [46] is optimised to work with logs that only showcase behaviour allowed by the process model, [44] introduces an adaptation which works with sub-traces of the event log. Hence, a finer-grained analysis is permitted through the examination of partially matching process instances. Still, an even more focused approach is presented by [30], which allows the analyst to define an acceptable level of dissimilarity between the log and the model. This parameter represents the number of "skips" allowed when checking whether or not a model accepts the process flow indicated by a trace.

Taking into consideration the accompanying studies, the 3 referenced techniques have proven to be feasible for the traditional CC use case, given their flexible approach and well-defined similarity measures.

The second identified area of CC development regarding entropy is "entropic relevance". Introduced by [45], it focuses on the average number of bits necessary to compress traces from a log using the relative likelihoods induced by a process model. This is possible with the usage of stochastic automata, which record probabilities for the sequences of events they define. Since it also takes into consideration event frequency, this approach could be particularly useful in SPC tasks, allowing for a more granular evaluation of relevant subgroup event logs.

Overall, the aforementioned entropy-based methods present a new perspective towards CC, considering not only the control flow, but also the frequency of events/ activities. Because of this, they might present a finer-grained way of quantifying process dissimilarity, which could be particularly applicable in the case of SPC. Yet, improvement can still be made when it comes to explaining the obtained results, as the software supporting these methods is a command-line tool (Entropia) which offers only numerical values [43].

4.1.7 Stochastic Conformance (Information Theory)

Capitalizing on the idea that event logs are inherently stochastic, the field of CC has seen developments which fully focus on this alternative perspective. For a start, works like [35] presents the concepts of stochastic precision and recall. These similarity measures are derived from comparing the entropy of two stochastic automata (obtained from the input log and model) against the entropy of a third automaton, obtained from the conjunction of the first two. This way, both commonalities and dissimilarities between the inputs are identified and quantified whilst also taking into account the probabilities of involved sequences and events. Even though more focused in scope than the previously mentioned entropic methods, this approach suffers from similar explainability issues. As Entropia is the main software tool in this case as well, the provided precision/recall values are not accompanied by any motivation for the scoring nor visualizations.

Striking a balance between precision and explainability [36], and later [37], introduce an alternative approach to stochastic CC, in the form of Earth-Mover's stochastic conformance (EMSC) plugin for ProM. Through the use of trace alignments over stochastic Petri Nets, this method is able to provide a fine-grained distance measure, which takes into account frequencies of sequences. Additionally, it offers visualizations and summaries that help to better illustrate the relationship between compared processes.

The EMSC approach quantifies the minimal changes needed to align an event log with a model from both a process-flow and stochastic perspective. It subsequently provides visual mappings of these alignments to better illustrate the underlying stochastic relations and deviations. Furthermore, as [37] illustrates, the additional frequency-based perspective that this method brings to conformance checking validates it as a suitable and viable alternative for the SPC use case.

4.1.7 Other

Besides the main conformance checking techniques reviewed in the previous subsections, the exploratory study has identified a number of other approaches. For brevity reasons, they are not presented as thoroughly as the methods above either because they are too limited in scope for a practical evaluation or do not have any software tool readily available. Nevertheless, they contribute to painting the diverse landscape of conformance checking and might motivate future studies to experiment with them in the context of SPC.

For instance, the concept of footprints capturing the causal and concurrency relations between activities has been encountered several times during exploration. Two notable CC applications of this approach are in the form of footprint matrices [7] and footprint dictionaries for BPMN replay [40].

From another perspective, papers have introduced more algebraic approaches for evaluating conformance. One such method is presented in [11] as a similarity measure computed using operations on process matrices derived from the compared process models. Another technique is introduced by [10], which calculates a distance metric using binary branch vectors obtained from encoding process models into dependency graphs, block trees, and binary trees subsequently.

Finally, the last reviewed methods have a broader approach towards conformance checking, focusing more on the structure of the process rather than anything else. This is done by either applying a mapping function directly on Petri Nets symbolising the process model [38], or on their T-P/P-T tables defining relationships between places and transitions [31].

Taking everything into consideration, the field of conformance checking is a vast one, showcasing a plethora of approaches for evaluating log-model similarity. These range from precise and fine-grained to more explainable and stakeholder-friendly. The literature review demanded by SRQ1 has identified these methods and associated software implementations, classified them according to approach, and identified their comparative strengths and weaknesses from a conceptual standpoint. The next step of this research is to select a number of appropriate candidates from this pool and evaluate them in a practical SPC scenario using real-world event logs.

4.2 Planning

Upon completion of the initial exploration of available CC methods, the next step of the study would involve a practical evaluation mimicking a real-world operational scenario. Following the guidelines of PM^2, an initial experiment planning was conducted. This took into account the business processes on which conformance checking will be performed, the methods applied, and the technical platform supporting them.

For a start, on account of the resource and time limitations associated with this study, the choice of experiment data prioritized two aspects: the event log needed to exhibit easily identifiable and distinct subpopulations, and it needed to be readily available. This way, data extraction and segmentation would be straightforward, allowing for the dedication of more efforts towards the actual application and evaluation of SPC methods. To this end, works such as [17, 39, 42, 47, 52] have shown the potential of medical-related data sets to exhibit well-defined and comparable subpopulations. This aspect determined the study to follow a similar path and conduct SPC on such a type of record. Ultimately, the chosen data set was the real-life event log taken from a Dutch academic hospital, described in BPIC 2011 [21]. As [17] shows, this log consists of 1143 cases (150291 events) describing the treatment paths for a heterogeneous mix of patients diagnosed with various types of cancer, at different stages of malignancy. Hence, a clear division of the subpopulations was available either via the diagnostic code which described each trace, or via the urgent/non-urgent or patient age subdivisions available for each type of diagnostic.

Considering that the research question guiding the practical experiment was already defined in **SRQ2**, the associated stage of the planning was simplified. The last objective before the extraction phase was to pinpoint which conformance checking methods would be evaluated. Considering their prevalence and "status quo" within the CC field, alignment-based methods [9] were a natural first choice. Next, the potential and "competitive" performance of enhanced token-based replay approaches [14] also justified their inclusion. Finally, the alternative perspectives offered by entropic [44] and stochastic [36, 37] methods raised veritable interests regarding their applicability in the SPC context.

Another argument in favour of the chosen approaches was their availability and validation through mature software tools such as: ProM [20], PM4Py [15], and Entropia [43]. Complementing the experiment's technical platform would be process discovery and visualisation tools like Disco [26] or Apromore [32].

4.3 Extraction

The principal objectives of the extraction phase were to determine the scope of the practical experiment, identify suitable subpopulations and set feasible SPC goals within the available time frame. Using the **LogVisualiser (LogDialog)** plugin of ProM, together with the **Map** and **Statistics** functions of Disco, the identifiable subpopulations of the event log were analysed and compared. Hence, the subpopulations described by the diagnosis codes M16 (ovarian cancer) and M14 (cancer of the corpus uteri) were selected for the experiment. They proved to be suitable candidates, considering their size and difference in exhibited behaviour, especially if they were further split based on the age of the patients. An empirical evaluation showed that the ±55 and ±60 age splits for M16 and M14 respectively yielded the most balanced subpopulations. These divisions were both optimal in terms of number of traces as well as observed behaviour. For brevity reasons, the main body of the paper will describe only the experiment conducted on the M16 subpopulation, with an optional appendix dedicated to M14 (The appendix has been excluded from the TScIT41 submission due to the page limit).

Hence, the scope of the evaluation was narrowed down to analysing how a collective application of the chosen CC methods can help in comparing the processes of the ≤55 and >55 subgroups of the M16 cancer diagnostic. Additionally, the experiment aims to identify what are the comparative strengths and weaknesses between these approaches and if they exhibit any potential synergies.

4.4 Data processing

Having identified suitable subpopulations for the experiment, the next step involved separating them from the original event log and applying any necessary filters. Besides making the analysis more computationally feasible with the available hardware (Presented in Section 6.2), the filtering aimed to uniformise the data sets. This allowed the "Mining" stage to extract information which best describes the process flow, minimising the effect of outliers.

Since the diagnosis code was an attribute associated with each trace, the required subpopulation (M16) was obtained using the **Filter Log on Trace Attribute Values** ProM plugin. Using Disco's **Statistics** screen, the distribution of trace length throughout the log was visualised, ranging from 1 to 1684 in an exceptional case. In order to remove outlying traces the **Filter Notebook (Filterbook)** ProM plugin was utilised to extract only the traces which had between 6 and 100 events. This way, the majority of traces exhibiting common behaviour were kept, removing the ones which either were potentially incomplete (having 1-2 events) or looped multiple times through the common behaviour sequence.

4.5 Mining and Analysis

With the relevant data sets extracted and filtered, there remained one more step before the selected CC methods could be applied and compared. Even though approaches like Earth-Mover's Stochastic Conformance [37] require only event logs, the other identified methods need both an event log, as well as a process model (in the form of a Petri Net) for evaluation.

To this end, the **Mine with Inductive Visual Miner** (IvM) ProM plugin was used to extract Petri Nets from the subpopulation event logs. The Inductive Miner algorithm was chosen due to its high level of fitness and good balance of quality forces (when compared to other PM algorithms) [16]. An important aspect of the process discovery stage is that the chosen subpopulation contained a majority of heterogeneous traces (only 2/99 traces being identical). That is, even though the treatment paths they described did share common sequences of events, taken in their entirety they were distinct from one another. One of the main reasons was the high number of unique or very low frequency events present in the log, 80 out of the 154 event classes occurring less than 4 times (or 1%) in the entire data set. Consequently, the IvM sliders for activities and paths were set to 0.325 for both the ≤55 and >55 demographics. This allowed the extracted models to capture the activity sequences shared by the majority of traces, filtering out low-frequency events and associations which mostly occurred once or twice in the entire event log. Since a large share of "unique" behaviour was excluded, the mined models would exhibit a transition which skipped the entire process flow (thus describing traces with 0 events). These "big skips" were subsequently removed as it was discovered they would skew the CC analysis, as most non-conforming traces could just utilise the skip in order to be accepted by the model.

Upon successful discovery of the process models describing each subpopulation's control flow, the identified CC methods (listed in Section 4.2) were applied via their associated software tools. The ProM extensions **Replay a log on Petri Net for Conformance Analysis** and **Compute Earth-movers' stochastic conformance (log-log)** were used for alignments and EMSC respectively. Alternatively, the PM4Py method **fitness_token_based_replay** was used for the CC approach bearing the same name. For entropic based conformance, the command line tool Entropia was used, with the parameters **-pmp** (partial matching precision) and **-pmr** (partial matching recall). Important to note is that, for alignments, all discrepancies have been uniformly penalised with a score of 1.

The results of each evaluation, along with the input logs and models, can be found in the author's GitHub[1].

4.6 Evaluation

The last phase of the iterative analysis component of PM^2 involved evaluating the results obtained from the application of the identified CC methods. This process explored what insights their collective usage could extract and how they differed and/or synergized with one another.

5 FINDINGS

5.1 Overview of analysed subpopulations

Upon completion of the preprocessing and mining steps, listed in the previous section, the event logs and associated process models describing each subpopulation have been refined. Hence, conformance checking was conducted on two subpopulations describing the treatment path for ovarian cancer in the 19-55 and 56-83 age groups. Table 2 gives an overview of the two subpopulations along with some relevant quantity metrics.

Table 2. Overview of the ≤55 and >55 subpopulations for the M16 diagnosis code

	≤55	**>55**
Cases	41	35
Events	539	908
Activities	113	110
Avg. events / trace	13	26
Min. events / trace	5	3
Max. events / trace	27	28

Before applying any CC methods, a naïve analysis of the observed metrics indicates that the process flow for the ≤55 subpopulation seems simpler than the one for the >55 demographic. Shorter on-average traces along with nearly twice as few total events in the log might suggest a simpler overall approach in the treatment of ovarian cancer for younger individuals.

Additionally, further visual analysis was conducted on the process models derived from the two event logs. From this step, it was discovered that the two subpopulations share a substantial common treatment path, both in activities and their causal relations. However, a notable difference was observed towards the end of the process flow: where the >55 subpopulation accepted only a specific sequence of events, the ≤55 demographic allowed for a more "liberal" approach. As can be seen in Figures 3 and 4, a strict sequence of activities that is enforced by the >55 model is replaced in the ≤55 model by a branching structure. This section makes every event, along with the order in which they are executed, optional. Besides, there were also deviations in how the process models allowed "skips" of this stretch of common behaviour. These observations can help corroborate the idea that the treatment path for the younger subpopulation is, in principle, simpler, allowing for greater liberty in execution and shorter process flows.

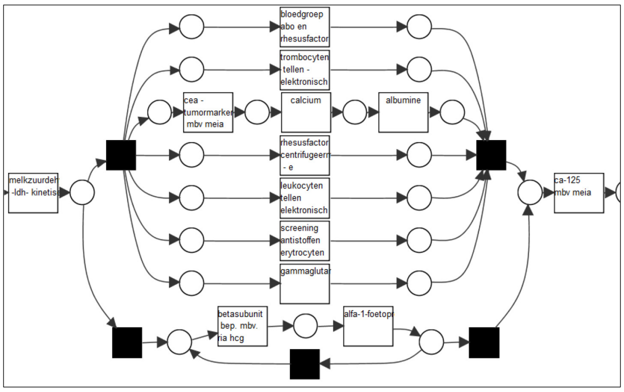
melkzuurdeh -ldh- kinetis
bloedgroep abo en rhesusfactor
trombocyten tellen - elektronisch
cea - tumormarker mbv meia
calcium
albumine
rhesusfactor centrifugeerr - e
leukocyten tellen elektronisch
screening antistoffen erytrocyten
gammaglutar
betasubunit bep. mbv. ria hcg
alfa-1-foetopr
ca-125 mbv meia

Fig. 3. Branching structure of the ≤55 years model for the M16 diagnosis code

[1] https://github.com/Victor-Adrian/HospitalLogSPCModels

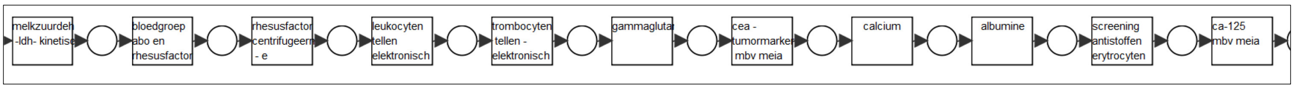
melkzuurdeh -ldh- kinetisc
bloedgroep abo en rhesusfactor
rhesusfactor centrifugeern - e
leukocyten tellen elektronisch
trombocyten tellen - elektronisch
gammagluta
cea - tumormarker mbv meia
calcium
albumine
screening antistoffen erytrocyten
ca-125 mbv meia

Fig. 4. Strict event sequence of the >55 years model for the M16 diagnosis code

Nevertheless, these claims could only be validated through further investigation, using the identified conformance checking methods. The following sections aim to identify the insights provided by each approach, along with any interactions between them.

5.2 Trace alignment

Considering its high level of explainability on both a general as well as a more granular level, trace alignment was the first CC method applied on the event logs. This approach helped to form a solid first impression, yielding considerable information regarding the commonalities and discrepancies between the two evaluated subpopulations.

For a start, the alignment of the >55 log with the ≤55 model produced a general fitness score of 0.82, while the reverse scenario generated a score of 0.33, as shown in Table 3. These results indicate that the model for ≤55 is more "permissive", allowing more behaviour than its counterpart for the older subpopulation. A principal reason for this significant difference could be that the branching structure of ≤55 (Figure 3) does permit the specific sequence of events described by >55, however, the opposite is not true. The "strict" order enforced by the process flow for the older demographic does not accommodate the deviations/skips in treatment path more common to the younger subgroup. Additionally, ≤55 allows for a shorter procedure which skips the common stretch of sequential behaviour between the two models, replacing it with a 5-event flow. This option is not represented in the >55 model. Since the simpler treatment path is fairly common for the younger subpopulation, this discrepancy further lowers the fitness score between the ≤55 log and >55 model.

Table 3. Similarity measures of the applied CC methods for the M16 diagnosis code

Metric	Log > 55, Model ≤ 55	Log ≤ 55, Model > 55
Global trace fitness (Alignment)	0.82	0.33
Percentage of fit traces (Token replay)	0	0
Average trace fitness (Token replay)	0.88	0.88
Log fitness (Token replay)	0.93	0.87
Partial matching precision (Entropy)	0.77	0.83
Partial matching recall (Entropy)	0.92	0.91
Similarity (EMSC)	0.36 (Log – Log)	0.36 (Log – Log)

Looking beyond the numerical scores, the log and model projections provided by alignments further support the claim that the ≤55 model is more permissive. These visualisations indicate how the strict treatment for the >55 demographic is indeed permitted by the branching structure of ≤55. Still, there are two notable differences between the two subpopulations. Firstly, the process flow for the older patients does allow certain 2-3 event traces not reflected in the ≤55 model These are most likely a result of skips to the treatment path catered specifically towards this subpopulation (as the younger patients also have a specific way of skipping it, not reflected by the >55 model). Secondly, there are certain ordering mismatches regarding the last 3 events of the models. Here the >55 model is more "liberal" and allows skips, while the ≤55 model enforces a strict order.

Overall, trace alignment proved to be a robust conformance checking method even in the context of subpopulation process comparison. The general overview provided by the global fitness scores seems to have coupled well with the more fine-grained analysis provided by log and model projection. These constructs helped to greatly enhance the method's explainability, allowing the experiment to pinpoint specific areas of deviations between the two models.

5.3 Enhanced token-based replay

Following up on the insights uncovered by trace alignment, enhanced token-based replay was conducted. The main objectives were to both try and validate the initial findings, as well as potentially uncover new ones.

As shown in Table 3, no traces fit perfectly between the two subpopulations (i.e., fully accepted by both models). However, the 88% average trace fitness indicates the fact that process flows are largely similar, with only small deviations between them. Most probably, these discrepancies originate from the branching structure associated with the ≤55 group and model-specific skips of the common stretch of sequential behaviour.

This difference in scoring could indicate a potential synergy between alignments and token-based replay. Where the former helped to better identify discrepancies (illustrating how the two process models differ in the behaviour they allow), the latter seems more effective at highlighting overarching commonalities between the subpopulations. By offering similar scores for both perspectives (log for ≤55 - model for >55 and its opposite), token replay seems to steer away from the notion that one subgroup has a more "permissive" model than the other. Instead, it supports the idea of the two models being largely similar, albeit with certain exceptions.

Nevertheless, the low explainability factor of this method, providing only numerical values, does attach a certain degree of opacity to the conclusion. This makes it, at least in this scenario, a better fit as an auxiliary to more robust methods (like alignments) rather than a principal form of CC.

5.4 Entropic partial matching precision/recall

The next CC method applied on the two subpopulations was the one based on entropic partial matching precision and recall. These measures are obtained through comparing the languages accepted by the DFAs derived from the input log and models, hence providing another "global" overview of conformance.

The first measure, precision, measures how well traces of the model are represented in the event log, while its counterpart, recall, indicates how well traces in the event log are represented in the model. Considering this, the values for the Log >55, Model ≤55 case (Table 3) suggest that the younger subpopulation may have a more permissive treatment path. Still, this idea is contradicted in the Log ≤55, Model >55 context, the results for the two cases seemingly opposing each other.

In essence, the relatively high precision and recall values suggest that the evaluated subpopulations are fairly similar. Still, the identified discrepancy, where cross-evaluation yields contradicting results, raises uncertainty concerning the validity of entropic measures in this particular scenario. Considering the event log's high level of trace heterogeneity, the great range of exceptional behaviour could have an adverse effect on the permissibility of derived DFA languages. Hence, further research might be necessary to explore the applicability of such methods in cases with more homogeneous subpopulations.

5.5 Earth-Mover's Stochastic Conformance

Finally, the last evaluated conformance checking method was also the one which provided the most granular perspective. Besides analysing the treatment paths' process flow, EMSC also took the frequency of event sequences into account. Due to this, the provided measure from Table 3 seems to contradict the values from all other approaches. With an approximate 0.5 difference from the other similarity measures, the 0.36 score paints the two subpopulations as notably dissimilar.

Further investigation of the stochastic log and model projections reveals the motive for this discrepancy. Even though regular alignments have shown that the ≤55 model does indeed permit the strict event sequence enforced by the >55 demographic, that specific choice of treatment is not as popular with younger patients. By looking at the log projections one can see that the 5-event skips of the ≤55 model occurs in 60% of cases. The fact that this relatively popular treatment path is not fully supported in the >55 model might be an important factor influencing the lower similarity value. Additionally, stochastic analysis also penalises the differences between the previously identified strict event sequence (>55) and branching structure (≤55). All deviations allowed in the younger subpopulation naturally have a frequency of 0% in the older group (since they're not allowed), hence decreasing the final score.

Overall, stochastic conformance proved to be an effective addition to the SPC scenario, helping uncover an additional layer of analysis that was "hidden" from the other evaluated CC methods. Still, the limitation in scope which provides this advantage also makes the approach better suited as an auxiliary to more "robust" methods like alignments. Additionally, its alignment-based visualisations in log and model projection also promote a synergy between these two methods, based on practicality as well as explainability.

6 DISCUSSION

Considering the impact that the overall research context might have had on the study, this section provides an exploration of any potential limitations or factors which might have influenced the results. Additionally, possible avenues for future development are also identified and discussed.

6.1 Limitations

A notable potential influence over the experiment's results is the lack of domain-specific knowledge, i.e., cancer treatment paths and procedures. Since the purpose of SRQ2 was to evaluate the practical capabilities of the identified CC methods, the hospital data set was chosen accordingly. Mainly, because it offered a suitable scenario for such an analysis, not because it represented a problem needing a solution. Consequently, research focused more on observing how different CC approaches highlight and quantify process flow differences rather than on the domain-specific factors behind these deviations. Furthermore, limited time and resources set a solely data-oriented scope for the evaluation, rather than an expert-oriented or hybrid one.

The performance of the CC analysis might have also been affected by the quality and general nature of the data set. As presented in Section 3.5, the subpopulations displayed a high degree of trace heterogeneity, ultimately impacting the accuracy of mined models. Although such logs can be common in real-life scenarios, the lack of suitable, readily available data sets for SPC, along with time constraints, hindered result validation on more defined, homogeneous subpopulations.

Finally, hardware limitations demanded a higher degree of abstraction in processing logs and mining models, to make CC computation time feasible. The experiment was conducted on an Intel i7-11800H 16 CPU @2.30 GHz with 32GB RAM, leading to the relatively small size of compared subpopulations.

6.2 Future research

Throughout the study, various avenues for further development have been identified, either to validate the current results or expand upon them, using different perspectives.

For example, as highlighted in the previous subsection, a more expansive investigation can be conducted on higher-performance hardware. This would allow the conformance checking to be done on larger and more diverse subpopulations. Additionally, the same evaluation could be carried out on different datasets, exhibiting ranging levels of trace homogeneity. This way, the synergies and differences between the tested CC approaches can be verified in multiple scenarios.

The evaluation can also be expanded to contain more conformance checking methods identified in answer to SRQ1, broadening its scope. Additionally, future works could evaluate CC approaches also from a resource and time perspective, since the current study only considered the process flow.

7 CONCLUSION

This study evaluates the applicability of four conformance checking approaches in the particular use case of subpopulation process comparison. After conducting a literature review of available CC methods, a selection has been made considering both their relevance in the field of process mining, as well as their potential in the SPC context. The chosen techniques were: trace alignments, token-based replay, entropy-based quotients, and Earth-Mover's Stochastic Conformance.

These methods have been used to assess the similarity of subpopulations describing ovarian cancer treatment paths for two age demographics (≤55 and >55). The research identified several comparative strengths and weaknesses between the identified approaches, as well as certain synergies. Considering the global perspective and high levels of explainability, trace alignments have shown to be among the most advantageous of the four, offering the most "complete" perspective of the two subpopulations. Nevertheless, token-based replay provided a better description of the common behaviour shared by the subpopulations, since alignments focused more on highlighting discrepancies. Additionally, stochastic conformance enhanced the granularity of the analysis, offering a novel frequency-based perspective. For this experiment, entropic approaches proved unconvincing, providing results which contradicted themselves to a certain extent, showing the need for further evaluation.

Overall, the findings from this research underscore the effectiveness of trace alignments, token-based replay, and Earth-Mover's Stochastic Conformance in the SPC context. The identified synergies indicate both the strength of concurrent applications of such methods, as well as the need for further, more expansive studies covering other CC approaches in broader operational scenarios.

ACKNOWLEDGEMENTS

I would like to extend my gratitude to my supervisor, Dr Faiza A. Bukhsh, for her insightful guidance, feedback, and encouragement throughout the course of this research.

USE OF AI ASSISTANCE

During the preparation of this work, the author used ChatGPT4 for general-purpose assistance in revising the document. This was comprised of grammar checks, review of academic tone in sentences, and finding context-appropriate synonyms. ChatGPT was not used for content generation but solely for providing review and feedback. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the content of the work.

REFERENCES

[1] Wil van der Aalst. 2016. Process Mining as a Bridge Between Data Mining and Business Process Management. In *Process Mining: Data Science in Action* (2nd ed.). Springer, Berlin, Heidelberg, 447–449. Retrieved from https://link.springer.com/book/10.1007/978-3-662-49851-4

[2] Wil van der Aalst. 2016. Process Mining. In *Process Mining: Data Science in Action* (2nd ed.). Springer, Berlin, Heidelberg, 30–34. Retrieved from https://link.springer.com/book/10.1007/978-3-662-49851-4

[3] Wil van der Aalst. 2016. XES. In *Process Mining: Data Science in Action* (2nd ed.). Springer, Berlin, Heidelberg, 138–144. Retrieved from https://link.springer.com/book/10.1007/978-3-662-49851-4

[4] Wil van der Aalst. 2016. Event Logs. In *Process Mining: Data Science in Action* (2nd ed.). Springer, Berlin, Heidelberg, 128–137. Retrieved from https://link.springer.com/book/10.1007/978-3-662-49851-4

[5] Wil van der Aalst. 2016. Limitations of Modeling. In *Process Mining: Data Science in Action* (2nd ed.). Springer, Berlin, Heidelberg, 25–30. Retrieved from https://link.springer.com/book/10.1007/978-3-662-49851-4

[6] Wil van der Aalst. 2016. Conformance Checking. In *Process Mining: Data Science in Action* (2nd ed.). Springer, Berlin, Heidelberg, 243–275. Retrieved from https://link.springer.com/book/10.1007/978-3-662-49851-4

[7] Wil van der Aalst. 2016. Comparing Footprints. In *Process Mining: Data Science in Action* (2nd ed.). Springer, Berlin, Heidelberg, 263–267. Retrieved from https://link.springer.com/book/10.1007/978-3-662-49851-4

[8] Wil van der Aalst, Arya Adriansyah, and Boudewijn van Dongen. 2012. Replaying history on process models for conformance checking and performance analysis. *WIREs Data Min. Knowl. Discov.* 2, 2 (March 2012), 182–192. https://doi.org/10.1002/widm.1045

[9] A. Adriansyah, B.F. van Dongen, and W.M.P. van der Aalst. 2011. Conformance Checking Using Cost-Based Fitness Analysis. In *2011 IEEE 15th International Enterprise Distributed Object Computing Conference*, August 2011. IEEE, Helsinki, Finland, 55–64. https://doi.org/10.1109/EDOC.2011.12

[10] Joonsoo Bae, James Caverlee, Ling Liu, and Hua Yan. 2006. Process Mining by Measuring Process Block Similarity. In *Business Process Management Workshops*, Johann Eder and Schahram Dustdar (eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 141–152. https://doi.org/10.1007/11837862_15

[11] Joonsoo Bae, Ling Liu, James Caverlee, and William B. Rouse. 2006. Process Mining, Discovery, and Integration using Distance Measures. In *2006 IEEE International Conference on Web Services (ICWS'06)*, September 2006. IEEE, Chicago, IL, 479–488. https://doi.org/10.1109/ICWS.2006.105

[12] Nick R. T. P. van Beest, Marlon Dumas, Luciano García-Bañuelos, and Marcello La Rosa. 2015. Log Delta Analysis: Interpretable Differencing of Business Process Event Logs. In *Business Process Management*, Hamid Reza Motahari-Nezhad, Jan Recker and Matthias Weidlich (eds.). Springer International Publishing, Cham, 386–405. https://doi.org/10.1007/978-3-319-23063-4_26

[13] Alessandro Berti and Wil van der Aalst. 2019. Reviving Token-based Replay: Increasing Speed While Improving Diagnostics. 2019. Aachen.

[14] Alessandro Berti and Wil M. P. Van der Aalst. 2021. A Novel Token-Based Replay Technique to Speed Up Conformance Checking and Process Enhancement. In *Transactions on Petri Nets and Other Models of Concurrency XV*, Maciej Koutny, Fabrice Kordon and Lucia Pomello (eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 1–26. https://doi.org/10.1007/978-3-662-63079-2_1

[15] Alessandro Berti, Sebastiaan van Zelst, and Daniel Schuster. 2023. PM4Py: A process mining library for Python. *Softw. Impacts* 17, (September 2023), 100556. https://doi.org/10.1016/j.simpa.2023.100556

[16] Alejandro Bogarín, Rebeca Cerezo, and Cristóbal Romero. 2018. Discovering learning processes using Inductive Miner: A case study with Learning Management Systems (LMSs). *Psicothema* 30.3 (August 2018), 322–329. https://doi.org/10.7334/psicothema2018.116

[17] R. P. Jagadeesh Chandra Bose and Wil M. P. van der Aalst. 2012. Analysis of Patient Treatment Procedures. In *Business Process Management Workshops*, Florian Daniel, Kamel Barkaoui and Schahram Dustdar (eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 165–166. https://doi.org/10.1007/978-3-642-28108-2_17

[18] Joos C. A. M. Buijs and Hajo A. Reijers. 2014. Comparing Business Process Variants Using Models and Event Logs. In *Enterprise, Business-Process and Information Systems Modeling*, Ilia Bider, Khaled Gaaloul, John Krogstie, Selmin Nurcan, Henderik A. Proper, Rainer Schmidt and Pnina Soffer (eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 154–168. https://doi.org/10.1007/978-3-662-43745-2_11

[19] Remco Dijkman, Marlon Dumas, Boudewijn van Dongen, Reina Käärik, and Jan Mendling. 2011. Similarity of business process models: Metrics and evaluation. *Inf. Syst.* 36, 2 (April 2011), 498–516. https://doi.org/10.1016/j.is.2010.09.006

[20] B. F. van Dongen, A. K. A. De Medeiros, H. M. W. Verbeek, A. J. M. M. Weijters, and W. M. P. van der Aalst. 2005. The ProM Framework: A New Era in Process Mining Tool Support. In *Applications and Theory of Petri Nets 2005*, Gianfranco Ciardo and Philippe Darondeau (eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 444–454. https://doi.org/10.1007/11494744_25

[21] Boudewijn van Dongen. 2011. Real-life event logs - Hospital log. https://doi.org/10.4121/UUID:D9769F3D-0AB0-4FB8-803B-0D1120FFCF54

[22] Marlon Dumas, Marcello La Rosa, Jan Mendling, and Hajo A. Reijers. 2018. Introduction to Business Process Management. In *Fundamentals of Business Process Management*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1–35. https://doi.org/10.1007/978-3-662-56509-4

[23] Maikel L. van Eck, Xixi Lu, Sander J. J. Leemans, and Wil M. P. van der Aalst. 2015. PM2: A Process Mining Project Methodology. In *Advanced Information Systems Engineering*, 2015. Springer International Publishing, Cham, 297–313. https://doi.org/10.1007/978-3-319-19069-3_19

[24] Mohammadreza Fani Sani, Juan José Garza González, Sebastian J. van Zelst, and Wil van der Aalst. 2023. Alignment Approximator: A ProM Plug-In to Approximate Conformance Statistics. September 11, 2023. Utrecht, The Netherlands.

[25] Luciano Garcia-Banuelos, Nick R.T.P. van Beest, Marlon Dumas, Marcello La Rosa, and Willem Mertens. 2018. Complete and Interpretable Conformance Checking of Business Processes. *IEEE Trans. Softw. Eng.* 44, 3 (March 2018), 262–290. https://doi.org/10.1109/TSE.2017.2668418

[26] C.W. Günther and A. Rozinat. 2012. Disco: Discover your processes. 2012. 40–44.

[27] Nick Hotz. 2018. What is CRISP DM? *Data Science Process Alliance.* Retrieved May 29, 2024 from https://www.datascience-pm.com/crisp-dm-2/

[28] Nick Hotz. 2021. What is SEMMA? *Data Science Process Alliance.* Retrieved May 29, 2024 from https://www.datascience-pm.com/semma/

[29] S. Ivanov, A. Kalenkova, and Wil M. P. van der Aalst. 2015. BPMNDiffViz: A Tool for BPMN Models Comparison. 2015. . Retrieved May 30, 2024 from https://www.semanticscholar.org/paper/BPMNDiffViz%3A-A-Tool-for-BPMN-Models-Comparison-Ivanov-Kalenkova/2d2b9a39234c673db77fb694319e7e7049e5625b

[30] Anna Kalenkova and Artem Polyvyanyy. 2020. A Spectrum of Entropy-Based Precision and Recall Measurements Between Partially Matching Designed and Observed Processes. In *Service-Oriented Computing*, Eleanna Kafeza, Boualem Benatallah, Fabio Martinelli, Hakim Hacid, Athman Bouguettaya and Hamid Motahari (eds.). Springer International Publishing, Cham, 337–354. https://doi.org/10.1007/978-3-030-65310-1_24

[31] Min-Hsun Kuo and Yun-Shiow Chen. 2012. A Method to Identify the Difference between Two Process Models. *J. Comput.* 7, 4 (April 2012), 998–1005. https://doi.org/10.4304/jcp.7.4.998-1005

[32] Marcello La Rosa, Hajo A. Reijers, Wil M.P. van der Aalst, Remco M. Dijkman, Jan Mendling, Marlon Dumas, and Luciano García-Bañuelos. 2011. APROMORE: An advanced process model repository. *Expert Syst. Appl.* 38, 6 (June 2011), 7029–7040. https://doi.org/10.1016/j.eswa.2010.12.012

[33] Wai Lam Jonathan Lee, H.M.W. Verbeek, Jorge Munoz-Gama, Wil M.P. van der Aalst, and Marcos Sepúlveda. 2018. Recomposing conformance: Closing the circle on decomposed alignment-based conformance checking in process mining. *Inf. Sci.* 466, (October 2018), 55–91. https://doi.org/10.1016/j.ins.2018.07.026

[34] Sander J. J. Leemans, Dirk Fahland, and Wil M. P. van der Aalst. 2018. Scalable process discovery and conformance checking. *Softw. Syst. Model.* 17, 2 (May 2018), 599–631. https://doi.org/10.1007/s10270-016-0545-x

[35] Sander J. J. Leemans and Artem Polyvyanyy. 2020. Stochastic-Aware Conformance Checking: An Entropy-Based Approach. In *Advanced Information Systems Engineering*, Schahram Dustdar, Eric Yu, Camille Salinesi, Dominique Rieu and Vik Pant (eds.). Springer International Publishing, Cham, 217–233. https://doi.org/10.1007/978-3-030-49435-3_14

[36] Sander J. J. Leemans, Anja F. Syring, and Wil M. P. van der Aalst. 2019. Earth Movers' Stochastic Conformance Checking. In *Business Process Management Forum*, Thomas Hildebrandt, Boudewijn F. van Dongen, Maximilian Röglinger and Jan Mendling (eds.). Springer International Publishing, Cham, 127–143. https://doi.org/10.1007/978-3-030-26643-1_8

[37] Sander J.J. Leemans, Wil M.P. van der Aalst, Tobias Brockhoff, and Artem Polyvyanyy. 2021. Stochastic process mining: Earth movers' stochastic conformance. *Inf. Syst.* 102, (December 2021), 101724. https://doi.org/10.1016/j.is.2021.101724

[38] Yahui Lu, Haofei Yu, Zhong Ming, and Hui Wang. 2016. A similarity measurement based on structure of Business Process. In *2016 IEEE 20th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, May 2016. IEEE, Nanchang, China, 498–503. https://doi.org/10.1109/CSCWD.2016.7566040

[39] Francesca Marazza, Faiza Allah Bukhsh, Jeroen Geerdink, Onno Vijlbrief, Shreyasi Pathak, Maurice Van Keulen, and Christin Seifert. 2020. Automatic Process Comparison for Subpopulations: Application in Cancer Care. *Int. J. Environ. Res. Public. Health* 17, 16 (August 2020), 5707. https://doi.org/10.3390/ijerph17165707

[40] Thomas Molka, David Redlich, Marc Drobek, Artur Caetano, Xiao-Jun Zeng, and Wasif Gilani. 2014. Conformance checking for BPMN-based process models. In *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, March 24, 2014. ACM, Gyeongju Republic of Korea, 1406–1413. https://doi.org/10.1145/2554850.2555061

[41] Jorge Munoz-Gama. 2016. Conformance Checking Explained: The University Case. In *Conformance Checking and Diagnosis in Process Mining: Comparing Observed and Modeled Processes.* Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-49451-7

[42] Priya Naguine. 2022. Subpopulation Process Comparison and Bottleneck Analysis: A Case Study of Frozen Shoulder. August 08, 2022. Enschede.

[43] Artem Polyvyanyy, Hanan Alkhammash, Claudio Di Ciccio, Luciano García-Bañuelos, Anna Kalenkova, Sander J. J. Leemans, Jan Mendling, Alistair Moffat, and Matthias Weidlich. 2020. Entropia: A Family of Entropy-Based Conformance Checking Measures for Process Mining. https://doi.org/10.48550/ARXIV.2008.09558

[44] Artem Polyvyanyy and Anna Kalenkova. 2019. Monotone Conformance Checking for Partially Matching Designed and Observed Processes. In *2019 International Conference on Process Mining (ICPM)*, June 2019. IEEE, Aachen, Germany, 81–88. https://doi.org/10.1109/ICPM.2019.00022

[45] Artem Polyvyanyy, Alistair Moffat, and Luciano Garcia-Banuelos. 2020. An Entropic Relevance Measure for Stochastic Conformance Checking in Process Mining. In *2020 2nd International Conference on Process Mining (ICPM)*, October 2020. IEEE, Padua, Italy, 97–104. https://doi.org/10.1109/ICPM49681.2020.00024

[46] Artem Polyvyanyy, Andreas Solti, Matthias Weidlich, Claudio Di Ciccio, and Jan Mendling. 2020. Monotone Precision and Recall Measures for Comparing Executions and Specifications of Dynamic Systems. *ACM Trans. Softw. Eng. Methodol.* 29, 3 (July 2020), 1–41. https://doi.org/10.1145/3387909

[47] Floor Rademaker. 2022. Subpopulation process comparison for in-hospital treatment processes: a case study for sepsis treatment. August 08, 2022. Enschede.

[48] Daniel Reißner, Abel Armas-Cervantes, Raffaele Conforti, Marlon Dumas, Dirk Fahland, and Marcello La Rosa. 2020. Scalable alignment of process models and event logs: An approach based on automata and S-components. *Inf. Syst.* 94, (December 2020), 101561. https://doi.org/10.1016/j.is.2020.101561

[49] Daniel Reißner, Raffaele Conforti, Marlon Dumas, Marcello La Rosa, and Abel Armas-Cervantes. 2017. Scalable Conformance Checking of Business Processes. In *On the Move to Meaningful Internet Systems. OTM 2017 Conferences*, Hervé Panetto, Christophe Debruyne, Walid Gaaloul, Mike Papazoglou, Adrian Paschke, Claudio Agostino Ardagna and Robert Meersman (eds.). Springer International Publishing, Cham, 607–627. https://doi.org/10.1007/978-3-319-69462-7_38

[50] A. Rozinat and W.M.P. van der Aalst. 2008. Conformance checking of processes based on monitoring real behavior. *Inf. Syst.* 33, 1 (March 2008), 64–95. https://doi.org/10.1016/j.is.2007.07.001

[51] Mohammadreza Fani Sani, Juan J. Garza Gonzalez, Sebastiaan J. Van Zelst, and Wil M.P. van der Aalst. 2020. Conformance Checking Approximation Using Simulation. In *2020 2nd International Conference on Process Mining (ICPM)*, October 2020. IEEE, Padua, Italy, 105–112. https://doi.org/10.1109/ICPM49681.2020.00025

[52] Eliza Stel. 2021. Process Comparison for Subpopulations of Patients with Heart Failure. 2021. Enschede.

[53] IEEE Standard for eXtensible Event Stream (XES) for Achieving Interoperability in Event Logs and Event Streams. https://doi.org/10.1109/IEEESTD.2023.10267858