





Stochastic-Aware Comparative Process Mining in Healthcare

Tabib Ibne Mazhar¹, Asad Tariq¹, Sander J.J. Leemans¹, Kanika Goel²,
Moe T. Wynn², and Andrew Staib³ 

¹ RWTH Aachen, Germany

² Queensland University of Technology, Brisbane, Australia

³ Princess Alexandra Hospital, Brisbane, Australia

Abstract. Evidence-based innovations are critical in optimising the delivery of healthcare services. Process mining aims to provide healthcare stakeholders with insights, derived from historical data recorded in hospital information systems, to optimise healthcare processes. Healthcare processes are well-known for their complexity and control-flow variations are inherent in patient pathways undertaken by different patient cohorts. Comparative process mining can reveal insights from studying the differences between healthcare processes to better understand best-practice patient pathways. In this paper, we take a design science approach to redefine an existing method for process comparison (PCM). Where PCM considers predominantly the control-flow perspective, we extend this method with the stochastic perspective, that is, how likely a particular pathway is for certain patient cohorts, to obtain the Probabilistic Process Comparison Method (P²CM). Furthermore, we further automate the method. Concretely, we introduce new, stochastic-aware, methods for sub-dividing process behaviour into cohorts based on trace attributes or other trace features, methods for focusing the comparative analysis on specific pairs of interesting cohorts, and provide a new method for in-depth comparison of process differences. The approach is evaluated using three real-life healthcare datasets, of which one case study is conducted with a domain expert from an Australian hospital.

Keywords: process mining · healthcare · comparative process mining

1 Introduction

Healthcare is a field that is confronted with widespread challenges, which require process improvement to be an integral part of the system. Data-informed innovations are important to make healthcare better and efficient [1,2,3]. New methods can assist healthcare organisations to rapidly adapt their processes to changing needs. Healthcare organisations around the world recognise the need to continually put efforts to improve their clinical as well as administrative processes. Healthcare organisations rely heavily on hospital information systems

which support clinical and administrative processes, and record executed process steps in process execution data [4]. This data consists of sequences of process steps (activities) executed for patients, hospital stays, etc. (cases).

Process mining is a family of techniques and methods, which can assist in answering questions that are crucial to improving processes in healthcare organisations. One area of process mining focuses on comparing groups of cases (*cohorts*) of a process. In such comparative process analysis, processing of different cohorts is compared, which may lead to insights into the process-based differences between cohorts - if processing is expected to be similar, e.g. leading to the identification of best practices, and into process-based similarities - where differences are expected [5,6]. The insights from such comparative process analysis can then be leveraged to optimise the processes involved. When comparing processes, several perspectives can be identified: the *control-flow* perspective entails the activities that can be performed in a process and their organisation into pathways, while the *stochastic* perspective describes how likely activities, pathways and behaviour in processes are [7]. In comparative process mining analysis, both perspectives may be beneficial: the control flow perspective may indicate that, for instance, a rework loop is possible, however without knowledge of the stochastic perspective that will indicate how likely that rework loop is, it remains unclear what the impact on the process of the rework loop is. A little-executed rework may be part of normal operating procedure, while an often-executed loop may pose a threat to process performance. Thus, a comparison of both perspectives may be beneficial in process comparison to optimise optimisation efforts [1].

Several techniques have been proposed to compare different parts of a process with one another, however applying them effectively in practice requires highly similar processes [5]: benefits have been shown to be derivable from the same (or, supposedly similar) process being executed in different settings. In some literature, such a setting of highly similar processes was known, for instance, comparing fulfilment processes in different geographic regions [8] and building permit processes in different municipalities [9]. To compare two processes with one another, several techniques can be applied [10,11]. However, if a single process is to be considered, a sub-division into variants (or, in log terms, *cohorts*) is necessary first. Several techniques have been proposed to identify cohorts from event logs [6,12].

To assist with applying the combination of these techniques, in [5] a generic method was proposed, the Process Comparison Methodology (PCM). However, as we detail in Section 2, PCM does not consider the stochastic perspective, and is highly manual with little automated support. As such, there is no method that takes an event log file as an input and identifies cohorts as output, along with visualisations of similarities and differences between the cohorts.

Given this gap, our problem statement is that we would like to have a method with which analysts can compare sub-processes for stochastic processing differences. In this paper, we use a design science [13] approach to extend the PCM method with stochastic awareness, operationalise PCM in a systematic manner,

and provide further guidance on how the techniques can be applied. We refer to this new method as the Probabilistic Process Comparison Method (P²CM). We evaluate our updated method twofold: using two real-life data sets, we validate the applicability of the method and using a case study, we validate the usefulness of the method in practice.

The remainder of this paper is organised as follows: Section 2 discusses related work. Section 3 details the research design. Section 4 introduces the updated method, Section 5 discusses the evaluative case studies, and Section 6 concludes the paper.

2 Related Work

In this section, we discuss related literature and derive our design objectives.

Process Mining in Healthcare. While PCM can be applied to datasets from any domain, in this paper we focus on healthcare processes as these processes typically consist of many variants, and as domain experts are keen to understand how the different patient cohorts pass through a hospital. Healthcare processes are characterised as complex and inherent to significant variations [14]. These variations can be a result of the differences in which the patient pathways proceed in a hospital. Process mining has the potential of uncovering details related to the execution of processes and has been used in healthcare. The potential has been explained in literature reviews [15] and a research agenda paper that highlights various opportunities and challenges [2]. In [16], the authors reviewed a pool of articles to understand how process mining has been applied to clinical pathways. The papers were classified in three categories, (i) discovery of actual execution pathways, (ii) analyse variants of execution pathways, and (iii) improve execution pathways.

As noted, one of the key areas of use of process mining is variant exploration. In [17,1], the authors used process mining to understand the similarities and differences between practices of different hospitals, but this comparison was done manually. Identifying differences between groups of pathway executions using process variant analysis can help to identify areas of potential improvement. Specific challenges related to process variant analysis exist. For example, comparing processes from a resource perspective, checking for compliance, and finding adverse events were mentioned in [2,18]. Despite growing interest in comparing healthcare processes, [2] identified the need for algorithms and methods that provide detailed explanations on the differences between process “variants” as a key challenge. This brings forth our first design objective:

DO1: A method that allows comparative analysis of process-based differences in cohorts of a single process.

Comparative & Stochastic Analysis. To compare multiple event logs with one another, a cross-comparison method has been proposed that first discovers a process model for each event log, and measures the differences between the model and each other event log in a cross-product setting [9]. This method, used in PCM [5] as well, is susceptible to the trade-offs that are present in process

discovery, and consider the stochastic perspective only partially, as the discovered models only consider the control-flow perspective. To compare two event logs with one another, approaches have been proposed based on transition systems [11,10] and fingerprints [19]. Furthermore, [20] and [9] both cross-compare two event logs: [20] by means of quality measures and [9] by means of deviations. Furthermore, in [21], predicted future process models can be compared. However, these techniques do not consider the stochastic perspective explicitly, which is essential to spot e.g. that exceptional behaviour is much more likely in one part than in the other, and assume that the two to-be compared event logs are known. Therefore, for our method, we specified the following design objective:

DO2: A method that compares the stochastic behaviour of two processes.

Process Comparison Methodology (PCM) PCM [5] has been proposed as a method to support comparative process mining. PCM consists of 5 consecutive phases: (1) in the first phase, the data must be extracted from information systems and pre-processed into the XES event log format [22]. Furthermore, in this step a trace attribute is selected to divide the event log (the α attribute). (2) in the second phase, the event log is divided into sub-logs, and an initial selection of these sub-logs is made, such that this selection will enable the answering of business questions and satisfy the goal of the comparative analysis. (3) in the third phase, suitable pairs of sub-logs (cohorts) are selected for comparison. (4) in the fourth phase, the selected pairs of sub-logs are compared to obtain detailed process-based differences. (5) the fifth phase involves reporting the relevant and impactful differences to the process owner.

In [5], the PCM method was applied to a non-healthcare case study, using semi-automated techniques and visualisations for phases 2, 3 and 4. However, most of the mentioned techniques utilised in PCM [5] only take the control flow – which steps are executed – into account, but only implicitly and unpredictably considering the stochastic perspective – how likely pathways are. Furthermore, considering the phases in detail, the alpha-attribute is chosen in phase 1, but little guidance is provided on how this attribute can be chosen, and data-supported automation that may aid analysts is limited. These details resulted in the following design objective:

DO3: A method that combines guidance for users with automated recommendations derived from data.

3 Research Design

We adopt a design science approach [23] and follow the six phases as described in [24].

(1) Problem Identification. Prior literature conveys that comparative process mining is important, in general, and in the healthcare sector in particular, to visualise the similarities and differences between processes. There is a need to develop comparative process mining techniques and methodological guidance to assist healthcare organisations in identifying potential improvements.

(2) Definition of Design Objectives. The overall objective is to propose a process comparison method that takes an event log file as an input, groups similar cohorts together, and provides visualisation of similarities and differences among the cohorts. Three design objectives (DOs) motivated by the related work in 2 are as follows:

- DO1 A method that allows comparative analysis of process-based differences in cohorts of a single process;
- DO2 A method that compares the stochastic behaviour of two processes;
- DO3 A method that combines guidance for users with automated recommendations derived from data.

(3) Design and Development. The design objectives identified in Step 2 were used to design and develop the P²CM method. The proposed method leverages the overall structure of PCM and extends several steps with stochastic awareness, provides more automated techniques, and provides more guidance for users of the method. As such, the P²CM method can be seen as an enhanced version of PCM. The six steps of the P²CM method are detailed in Section 4.

(4) Demonstration. To apply the P²CM method in practice, we implement scripts for the new alpha attribute selection technique and the new comparative process visualisation algorithm.

(5) Evaluation. To evaluate the P²CM method, we use the evaluation framework presented by [25]. Two ex-post evaluation strategies are used. First, an experimental controlled experiment [23] is conducted by applying the P²CM method to two real-life publicly available event logs with the objective of assessing the applicability of the method. Second, we perform an observational case study [23] with the emergency department of a healthcare organisation - the Princess Alexandra Hospital, Brisbane, Australia. The objective was to assess the usefulness of the method, i.e., whether the method we propose can be used to unearth meaningful insights for stakeholders. The findings are presented in 5. The ex ante evaluation [25] of these design objectives - to, for instance, validate their applicability in practice or their appropriateness with focus groups - is not within the scope of this paper, but would be an interesting area of further research.

(6) Communication. This manuscript is a means of sharing the new P²CM method. All introduced techniques have been implemented, and their source code is publicly available.

4 Artifact: P²CM

In this section, we introduce our new method, the Probabilistic Process Comparison Method (P²CM), which extends and instantiates the PCM framework. As by DO3, we aim to automate the steps as much as possible, we slightly change the overall structure of the PCM method, described in Section 2: we denote the selection of the alpha attribute in its own phase, as this step can be automated.

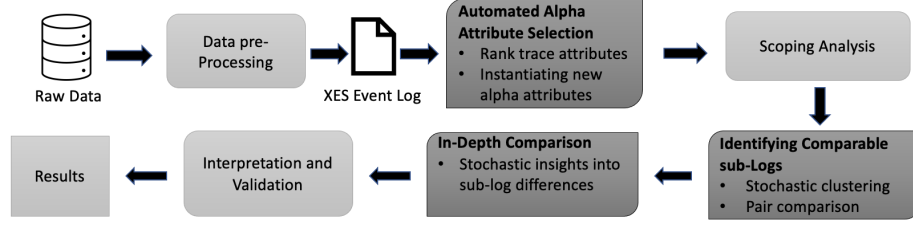


Fig. 1: P²CM. The darker steps are new or different from PCM.

Fig. 1 provides a visual overview of P²CM. Furthermore, we change the following phases to explicitly consider the stochastic perspective (DO2) and further automate them (DO3): we introduce the new phase 2 (selecting the alpha attribute), we change phase 4 (identifying comparable sub-logs) and change phase 5 (in-depth comparison).

4.1 Assisted Alpha Attribute Selection

The alpha attribute plays a significant role in distinguishing between process variants. However, identifying a suitable alpha attribute in an event log requires domain expertise and a good understanding of the process. To assist the analyst while minimising domain expert input, we rank the trace attributes in an event log by feature importance as a guide to the user. We propose two machine learning techniques, one based on unsupervised learning (ID-K), which groups similar data points without any dependency on a target variable and the other on supervised learning (ID-R), which combines decision trees for classification based on a target variable. Each method returns a graph of the relative importance of each trace attribute in an event log. Both techniques may indicate the importance of an attribute; an attribute indicated by both provides an even stronger indication.

ID-K: k-Means Clustering. Clustering groups data into clusters based on their similarity in certain features. For our analysis, we start with an XES event log, and consider the trace attributes. Numeric, boolean and categorical features are considered, while unique identifiers, timestamps and free-text comments are not considered. These latter categories are inherently unsuitable as alpha attributes, as they do not sub-divide the traces of the log into clearly defined and understandable sub-logs. To transform data into a format that machine learning algorithms can process, factorisation is used for categorical and boolean attributes⁴, which transforms this data into enumerated or categorical values.

Once the data has been transformed, the next step is determining the optimal number of clusters through the Elbow method [26], which allows a user to select the appropriate number of clusters by visualising the within-cluster sum

⁴ <https://pandas.pydata.org/docs/reference/api/pandas.factorize.html>

of squares (WCSS)[27]. The inflection point or "elbow" in the plot, where increasing the number of clusters would not significantly lower the WCSS, shows a levelling out of the inter-cluster variability. For instance, Fig. 3 shows the elbow graph for one of our evaluations. In this graph, the elbow is at number of clusters = 2, as after that there is no significant decrease in the WCSS.

Then, the k-means clustering algorithm is applied to the selected trace attributes. The contribution of each attribute in segregating the traces into different groups is studied and a plot of the relevant importance of each feature is returned. This is done by calculating the significance of each feature for each cluster based on the magnitude of the weight of the feature in the centroid vector.

ID-R: Random Forest Classifier In a supervised learning setting, we assess the influence of each trace attribute on a target variable. Here, we consider the length of a trace as the target variable, while the training features are the attributes from the log data. We want the outcome to factor in the effect of length of a trace as the count of activities carried out for a case may have interesting reasons, so we extract the count of activities per trace and add it as the target feature in the data set. We choose to utilise an ensemble classifier, specifically the Random Forest classifier, which combines multiple decision trees to enhance the model's overall performance [28]. It operates by training multiple decision trees on randomly selected subsets of the data and then averaging the predictions of all the trees to make a final prediction. This method is robust to high variance and outliers.

In a random forest classifier, the importance of attributes is calculated using the mean decrease impurity (MDI) method [28], which calculates the total reduction of Gini impurity that each attribute provides across all the trees in the forest. The attribute importance is then determined by averaging the reduction in impurity across all trees that use the attribute.

4.2 Identifying Comparable sub-Logs

The identification of comparable sub-logs is the next phase of the analysis. Sometimes the alpha attribute may have hundreds or thousands of values and comparing them one-on-one creates $n * (n - 1) / 2$ comparisons. To reduce the potential number of comparisons, and thereby limit domain expert involvement (DO3), we introduce a new method, consisting of ranking, filtering and clustering. The method follows several steps. First, it ranks the values of the alpha attributes based on their count and takes the top-most frequent ones, based on a user-provided parameter. Second, we reduce the n^2 comparison space by clustering sub-logs based on their similarities with other sub-logs.

In order to take the stochastic perspective into account, we use the Earth Movers' Stochastic Conformance Checking (EMSC) [7] to obtain a sub-log vs sub-log similarity score table. We used the stochastic perspective instead of the control flow perspective to show the likelihood of following a pathway by similar patient cohorts (DO2). Given two sub-logs, EMSC will compute a score that

is 1 if the two logs have the same stochastic behaviour, and 0 if the stochastic behaviour of the two sub-logs is completely different.

In the table of similarity scores, each sub-log has a pairwise similarity score between 0-1 with every other sub-log. After creating this table, we need to find the optimum number of clusters, for which we apply the Elbow method. Next, clustering is applied to the vectors of EMSC scores, to identify clusters of sub-logs.

Then, the sub-logs to compare are to be chosen, which is a step that inherently requires domain expertise. Nevertheless, the clustering provides guidance in two ways: pairs of sub-logs in different clusters are likely to differ in stochastic behaviour, while pairs of sub-logs are less likely to differ in stochastic behaviour. The former can be used to study the differences between process cohorts – e.g. to perform auditing –, while the latter can be used to study commonalities – e.g. to identify best practices.

4.3 In-Depth Comparison

Using the techniques of sections 4.1 and 4.2, we get the alpha attributes and the sub-logs that we want to compare. In this section, we present a new visualisation technique of in-depth comparison between two sub-logs which we call the Visual Process Comparator (VPC). VPC takes a log (the complete, not-subdivided log), L and two sub-logs of that log, L_1 and L_2 . At first, a directly follows graph (DFG) is made from L . In a DFG, every node represents an activity and the edges describe the relationship between the activities [29]. To increase understandability, and to avoid clutter and spaghetti models, we first filter the edges: given a percentage parameter set by the analyst, we remove all edges that are below that threshold. If the log is not very complex, we can choose a threshold of 0. Then we check how many of those connections are present in L_1 and L_2 and use only those. Second, we visualise the differences between L_1 and L_2 on this filtered DFG.

We denote $L_1(a \rightarrow b)$ and $L_2(a \rightarrow b)$ as the frequency of the DFG edge from a to b in L_1 and L_2 respectively. For a particular node a , $\sum_{(a,b') \in DFG} L_1(a, b')$ and $\sum_{(a,b') \in DFG} L_2(a, b')$ denote the summation of all the edge frequencies out going from that particular node for L_1 and L_2 respectively.

The below formula is used for showing the relative frequency difference D :

$$D_{(a \rightarrow b)} = \frac{L_1(a \rightarrow b)}{\sum_{(a,c') \in DFG} L_1(a, c')} - \frac{L_2(a \rightarrow b)}{\sum_{(a,c') \in DFG} L_2(a, c')}$$

To indicate the importance of an edge, we scale the width according to its relative appearance in both sub-logs varying between a width of 0.5 when the edge is present in neither L_1 or L_2 to 1.5 when the edge is the only outgoing edge of that node. The below formula is used for width calculation,

$$width_{(a \rightarrow b)} = \frac{L_1(a \rightarrow b)}{\sum_{(a,c') \in DFG} L_1(a, c')} + \frac{L_2(a \rightarrow b)}{\sum_{(a,c') \in DFG} L_2(a, c')} + 0.5$$

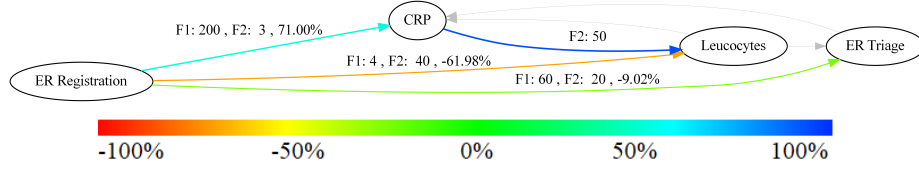


Fig. 2: Example of the Visual Process Comparator.

The edges are coloured so that the user can see the differences instantly. We use the HSV colour scale for the graph. The colour ranges from red (0° , 73%, 96%) to blue (200° , 73%, 96%) and all the colours in between based on D (see Fig. 2). Grey colour indicates the edges are neither in L_1 nor L_2 .

Besides that, the frequency of each edge of each sub-log (denoted as F1 and F2) and the relative frequency differences in percentage between two sub-logs, are also shown on each edge.

For instance, consider Fig. 2. The thin grey edges indicate they are only present in the L but not in L_1 and L_2 and the green edge indicates near 0 relative difference. $CRP \rightarrow Leucocytes$ edge is deep blue as it only present in L_2 . $ER\ Registration \rightarrow CRP$ has a teal colour as the relative difference is 71% and $ER\ Registration \rightarrow Leucocytes$ has relative difference of -61.98% and the colour is orange here.

5 Evaluation

In this section, we describe the twofold evaluation we performed to verify the applicability and usefulness of the method and its implemented tool support.

5.1 Applicability

As a first evaluation, we assess the applicability of P^2CM by applying it to two real-life healthcare data sets that are publicly available. The aim is to illustrate that P^2CM can be applied to real-life event logs and may lead to insights into the differences in (stochastic) process behaviour of cohorts within a single process with minimal domain experts' input.

Sepsis. Sepsis, a condition characterised by the body's harmful response to infection, is a frequent cause of severe illness and death worldwide [30]. The dataset https://data.4tu.nl/articles/dataset/Sepsis_Cases_-_Event_Log/12707639 consists of 1050 patient cases recorded between 2013 and 2015 and includes diagnostic test results, patient demographic information and organisational information. We apply P^2CM to understand the diagnostic journey of sepsis patients and identify factors that may affect patient outcomes.

Data pre-processing. The event log has 5 distinct release types, i.e. patient discharges. After applying a filter to exclude cases that lacked any release activity, as these cases were deemed incomplete and lacked a definitive end activity,

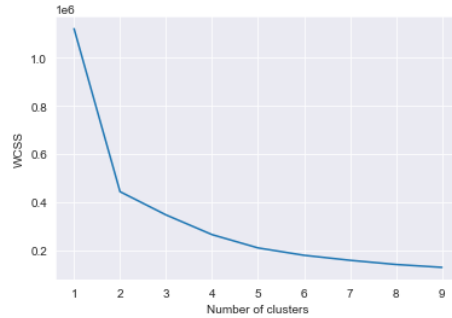


Fig. 3: The Elbow method on our Sepsis analysis.

Table 1: Alpha attribute influence.

Sepsis			MIMIC			PAH		
Feature	ID-R	ID-K	Feature	ID-K	ID-R	Feature	ID-K	ID-R
Age	0.23	0.76	icd_title	0.51	0.15	Time on Ramp	0.10	0.13
Diagnose	0.33	0.30	chiefcomplaint	0.08	0.16	Primary Diagnosis	0.86	0.08
SIRSCritHeartRate	0.03	0.16	acuity	0.44	0.01	Location after Triage	0.44	0.07
SIRSCritTachypnea	0.04	0.16	heartrate	0.07	0.15	Consultation Type	0.03	0.06
Release_type	0.04	0.08	temperature	0.01	0.14	Departure Destination	0.11	0.05

a total of 777 cases remained. We added a new trace attribute to the event log denoting the release type. The event log contains trace attributes such as **Case ID**, **Age**, **Transition**, **Organization**, **Activity Count**, **Diagnose**, and **Diagnostic Tests**. After removing the non-contributing trace attributes (case identifier, comments and timestamps, see Section 4.1), we have 26 trace attributes for our analysis. We removed the cases with missing values for these attributes and this filtered event log has 729 cases.

Assisted Alpha Attribute Selection. We applied the ID-K and ID-R methods to select the alpha attribute from the selected 26 trace attributes and got their respective relevant importance of each feature as output, using the output of the Elbow method in Fig. 3). The relevant importance of the top 5 attributes of both methods is shown in Table 1. It can be observed that ID-K and ID-R both returned **Age** and **Diagnose** as the most important features and for our analysis, we have considered both **Age** and **Diagnose** as candidate alpha attributes.

Scoping analysis. In the selected candidate alpha attributes, **Diagnose** has more than 100 distinct values, and we take the top ten most frequent ones choose C,B,E,H,G,D,K,R,Q and S, as a result we have 10 sub-logs; one for each selected diagnosis (the diagnoses are anonymised; knowledge of them is not necessary for P²CM). For **Age**, we partition the values into 10-year periods, resulting in eight sub-logs and they are 0-20, 20-30, 30-40, 40-50, 50-60, 60-70, 70-80, 80-90.

Identifying comparable sub-logs. In this step, we apply our new method (see Section 4.2) to obtain the log vs log comparison scores. We find the number of optimal cluster is 3 by using the elbow method on the scores. After that, we use k-means clustering to find {C,B,E,H,D,K and, R} in cluster 0, {S} in cluster 1 and {G,Q} in cluster 2.

After applying VPC (Section 4.2) on the **age** attribute sub-logs, obtaining the log vs log comparison scores and using the Elbow method (Section 3), we get 2 clusters. Then using k-means clustering, we find 0-20, 20-30 in cluster 0 and 30-40, 40-50, 50-60, 60-70, 70-80, 80-90 in cluster 1.

When comparing the sub-logs of **B** and **S** (see Fig. 4a), we instantly notice a significant number of bright red edges, which indicates that **B** has a lot more edges. The edges **IV Liquid** \rightarrow **IV Antibiotics** and **LacticAcid Triage** \rightarrow **Admission NC** (not shown) have relative differences -36.99% and -52.81% which means for sepsis **S** patients' treatment, this paths play an important role.

Analysing the **Age** attribute of sub-log 0-20 vs 30-40 (see Fig. 4b), we see that around 50% of the edges are only present in the sub-log 30-40. **IV Liquid** \rightarrow **IV Antibiotics** has a percentage difference of -39.08 percent which indicates that for patients' age between 40-50, this path is more important than any other paths. When the sub-logs were compared using VPC, it became clear that the treatment paths for sepsis patients varied significantly and that IV fluids were preferable to IV antibiotics for patients aged 40-50. Overall, P²CM provided insights into the differences between sub-logs of healthcare pathways with minimal domain expert input.

MIMIC MIMIC-IV-ED is a database of emergency department (ED) admissions at the Beth Israel Deaconess Medical Center between 2011 and 2019, which contains vital signs, triage information, medication reconciliation, medication administration and discharge diagnoses of around 425 000 ED stays. The ED is a

resource limited environment and human care is rationed to provide the best possible patient care [31]. MIMICEL is an event log derived from MIMIC [31].

Data pre-processing. We selected the event attributes `temperature`, `heartrate`, `resprate`, `o2sat`, `sbp`, `dbp`, `pain`, `acuity`, `chiefcomplaint` and `icd_title` and lifted them to the trace level. Other attributes that had a high percentage of null values or were identifiers, timestamps or comments were dropped; 10 trace attributes were used further. Out of the initial 448 972 cases, 436 737 cases remained after filtering traces with missing values.

Assisted alpha attribute selection. We utilised both ID-K and ID-R; the top five results are shown in Table 1. The two most important features from ID-K are `icd_title` and `Acuity`, and for ID-R are `icd_title` and `chiefcomplaint`; we selected the common one `icd_title` as the alpha attribute.

Scoping analysis. Here, we find comparable sub-logs based on the `icd_title` attribute, which has more than a thousand values, and it is not possible to compare these all one-on-one. So we select the top ten most frequent values and generate ten event logs by filtering the event log based on these values.

Identifying comparable sub-logs. Then we create ten event logs from the MIMICEL event log based on these ten values. By using our proposed technique for identifying comparable sub-logs, we obtain the log vs log comparison scores. The Elbow method indicates using 3 clusters. We then apply k-means clustering to find Pneumonia, unspecified organism, Altered mental status unspecified, Fever, unspecified in cluster 0, ALCOHOL ABUSE-UNSPEC, Alcohol abuse with intoxication unspecified in cluster 1 and Unspecified abdominal pain, CHEST PAIN NOS, Chest pain unspecified, ABDOMINAL PAIN OTHER SPECIED, HEADACHE in cluster 2.

Next, we compare two DFGs - they are ALCOHOL ABUSE-UNSPEC vs Alcohol abuse with intoxication, unspecified and HEADACHE vs Altered mental status, unspecified. As ALCOHOL ABUSE-UNSPEC and Alcohol abuse with intoxication, unspecified are in the same log with very similar diagnosis name, we are interested in understanding the main differences between them. HEADACHE and Altered mental status, unspecified are in two different clusters, and thus we also like to observe the main differences between them.

In-depth comparison.

To the selected pairs, we apply the VPC with a filtering parameter of 80%. When we look at ALCOHOL ABUSE-UNSPEC vs Alcohol abuse with intoxication, unspecified in cluster 1. As the names suggest, there should not be too many differences between them, and the graph validates our intuition. All the

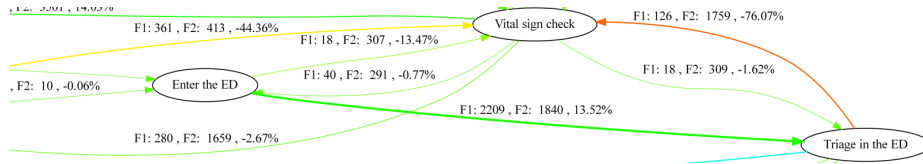


Fig. 5: VPC on HEADACHE vs Altered mental status, unspecified comparison.

edges are green except **Triage in the ED** \rightarrow **Vital sign check**, -43.29%. This shows that, when there is intoxication involved, more patients are sent for **Vital sign check**.

Lastly, **Headache vs Altered mental status, unspecified** Fig. 5 shows considerable differences in edges. Specially, **Triage in the ED** \rightarrow **Vital sign check**, -76.03% indicates **Vital sign check** is an important step when treating patients with **Altered mental status, unspecified**. Overall, using P²CM we were able to study process-based differences with minimal domain expert input.

5.2 Usefulness

The second evaluation entails an application of P²CM in a case study, performed at the emergency department of the Princess Alexandra Hospital in Brisbane, Australia. The corresponding data set contains ED pathways in 2019-2021.

Data pre-processing. The data set was converted to XES, after which the activities related to bed management were removed to focus the analysis. Furthermore, cases with data-type mismatches were removed. The remaining log had 2 329 846 events, 134 846 traces and 48 activities.

Assisted Alpha Attribute Selection. In our study on alpha attribute extraction, we utilised our exclusion criteria to select categorical attributes that were likely to be useful alpha attributes. Our methods ID-K and ID-R both on the pre-selected attributes revealed that **Time on Ramp** and **Primary Diagnosis Snomed Code** were the top 2 most important attributes for segregating the traces into sub event-logs. The relative importance of the top 5 attributes of both methods is shown in Table 1. From the alpha attribute selection, we get **Time on Ramp** and **Primary Diagnosis Snomed Code** (**Primary Diagnosis**) as the alpha attributes.

Scoping analysis. We binned the **Time on Ramp** in six parts based on their frequency, while attempting to keep the bins balanced in their number of traces. From the **Primary Diagnosis** feature, we chose the top 10 diagnoses based on their frequency, this included **Chest Pain**, **Mental Health Problem**, **Abdominal Pain**, **Viral Illness**, **Syncope**, **Back Pain**, **NSTEMI - Non-ST Segment Elevation MI**, **Headache**, **Cellulitis** and **Alcohol Intoxication**.

Identifying comparable sub-logs. The clustering of sub-logs, with 3 clusters identified, was according to expectation: the bins of **Time on Ramp** that were close in value were clustered together. We go through the same process for **Primary Diagnosis** and obtain 3 clusters. Based on the clustering, the domain expert identified two pairs of potential interest: (1) **mental health problem** vs **viral illness**, as an example of within-cluster differences, and (2) **chest pain** vs **NSTEMI - Non-ST segment elevation MI**, as these are medically closely related, but still showed as being in different clusters. Furthermore, we decided to compare sub-logs based on the **time on ramp** attribute clustering, taking (0,1.0] vs (1.0,105.0] minutes as representatives of two different clusters (3).

In-depth comparison & interpretation and validation. We apply the VPC to these three pairs of sub-logs. For the first pair, we compared **mental health problem** and **viral illness**. Some of the procedural differences highlighted in the visualisation were expected by stakeholders, such as the edge from **Triaged at**

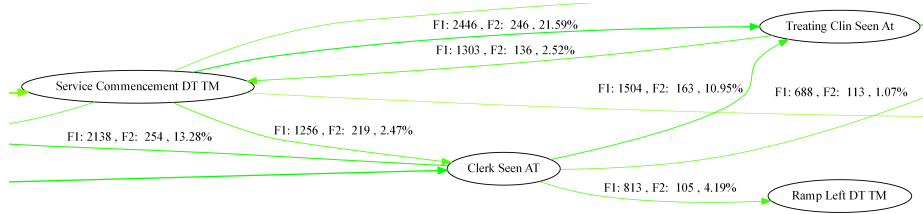


Fig. 6: VPC on Chest Pain and NSTEMI - Non-ST segment elevation MI.

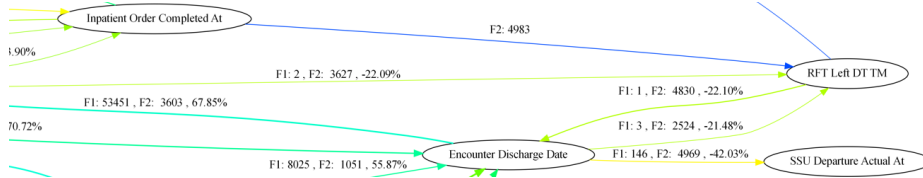


Fig. 7: VPC on time on ramp: (0, 1.0] vs (1.0,105.0] minutes.

to Treat Nrs, and Service Commencement to Edip date as for many mental health problems, neither Emergency Department treatment nor admission occurs. These patients are rapidly transferred to the Emergency Mental Health Unit. Other differences were **Triaged at** to **Clerk seen at**, which stakeholders indicated may be a missing recording step in the process. For the second pair, we compared **Chest Pain** with **NSTEMI - Non-ST segment elevation MI**, shown in Fig. 6. Again, several differences were expected, **Triaged at** → **Clerk Seen**, which indicates that the administrative step in the middle is often skipped for urgent cases. However, the **Service Commencement** to **Treating Clin Seen** edge was a new insight to the expert, while **Service Commencement** to **Edip Date** may again indicate a recording issue. For the third pair, we compared time on ramp being less than one minute (F1) vs 1 to 105 minutes (F2), shown in Fig. 7. The first observation is that the colours indicate large process-based stochastic differences, as several edges are of teal and yellow colours. These findings suggest both differences that can be explained by differences in the nature of the clinical presentations and their journey through the ED as well as differences related to data recording processes, which may be important in performance reporting.

From our discussion, it is clear that PVC as part of P²CM, based on the alpha attributes **Primary Diagnosis** and **Time on Ramp**, can be effective in showing the stochastic process-based differences between different patient groups.

5.3 Discussion

P²CM presented in this paper takes a single event log as input and provides results between cohorts within that event log. P²CM hence allows comparative analysis of process-based differences (DO1). P²CM is also a stochastic-aware technique (DO2). The key steps of alpha attribute selection and identifying comparable sub-logs and in-depth comparison, which are stochastic to a larger extent

than the original PCM, make P²CM stochastic-aware. In the future, P²CM could be extended in the data pre-processing and scoping analysis steps with explicit consideration of the stochastic perspective. Furthermore, P²CM provides automated techniques that guide users into choosing alpha attributes and comparable sub-logs, and visualises stochastic differences between processes to guide analysts in finding differences or commonalities between cohorts of a process, thus taking a step towards satisfaction of guiding users with automated recommendations (DO3). In the future, it would be interesting to extend automation by refining the comparable sub-logs identification step using heuristics or machine learning to guide analysts further towards potentially notable differences.

The techniques introduced in this paper as part of P²CM use concepts from existing stochastic-aware techniques, but differ in key ways. In [6], an event log is split along trace attribute values to find their values with the largest influence on stochastic behaviour. Our approach extends it with a full method, non-categorical attributes and a visualisation of the actual differences. Finally, we provide several automated techniques that assist analysts in selecting or creating one or more alpha attributes. Suitable pairs of sub-logs to compare are selected in phase 3, however [5] emphasises the need to use similar processes. We extend the method with a stochastic-aware approach that guides analysts in choosing *similar and dissimilar pairs* of sub-logs for comparison. Furthermore, our approach does not require the discovery of process models to perform this selection, which inherently involves certain well-known trade-offs [32]. The process comparison methods in phase 4 of [5] and literature do not focus on the stochastic perspective. We provide a new process discovery technique/visualisation that highlights differences in stochastic behaviour between two sub-logs.

The experiments can be reproduced using the scripts available at <https://github.com/asadTariq666/BPM-Alpha-Attribute-Selection>. The Sepsis data is publicly available, while the MIMIC data is semi-publicly available [31]. For legal/privacy reasons, the data of the Princess Alexandra Hospital cannot be shared.

Several limitations of this work are noted. As the author team applied P²CM themselves, it was not possible to evaluate the ease of use of P²CM objectively: future research is needed to assess this aspect. Second, the experiment covered treatment (Sepsis) and emergency care (MIMIC, PAH), and we see no factors that would prevent applying P²CM to other areas of healthcare. In order to generalise the application of P²CM to other domains, it is important that such cases *have* attributes, or, more in general, sub-processes that can sensibly be compared with one another. In case these sub-processes are already known to domain experts, P²CM might be applicable partially (identifying comparable sub-logs & VPC). Another limitation is that the ordinal encoding used for categorical and boolean attributes may impose an order that can impact k-means clustering. Future research could explore alternative encoding methods like word2vec or one-hot encoding to preserve semantic meaning without introducing implicit order.

6 Conclusion

In health processes, optimisation ideas may be derivable from the comparison of similar but differing processes. In this paper, we applied a design science approach to introduce a method, the Probabilistic Process Comparison Method (P²CM), to satisfy the design objectives of (i) allowing for comparative analysis of process-based differences in cohorts of a single process, (ii) considering the stochastic perspective of behaviour, and (iii) guiding users with automated recommendations derived from data. We showed that P²CM adheres to (i) and (iii), while (ii) is satisfied by the combination of the techniques used in P²CM: in all changed steps, the stochastic perspective is taken into account: most insights obtained were of a stochastic nature, and would have been missed by techniques unaware of the stochastic perspective. Following open-science principles, method and analysis techniques are available to the community and two publicly accessible datasets were used to ensure the reproducibility of our findings.

As further future work, the concepts of process cubes [12] may be applied to expand P²CM to use the structure between attributes to further reason about (hierarchical) relations between attributes, and guide users towards sub-processes with notable differences.

References

1. S. Suriadi, R. S. Mans, M. T. Wynn, A. Partington, and J. Karnon, “Measuring patient flow variations: A cross-organisational process mining approach,” in *APBPM*, pp. 43–58, Springer, 2014.
2. J. Munoz-Gama, N. Martin, *et al.*, “Process mining for healthcare: Characteristics and challenges,” *JBIM*, vol. 127, p. 103994, 2022.
3. D. Duma and R. Aringhieri, “Real-time resource allocation in the emergency department: A case study,” *Omega*, vol. 117, p. 102844, 2023.
4. R. Andrews, K. Goel, P. Corry, R. Burdett, M. T. Wynn, and D. Callow, “Process data analytics for hospital case-mix planning,” *JBIM*, vol. 129, p. 104056, 2022.
5. A. Syamsiyah, A. Bolt, L. Cheng, B. F. Hompes, R. Jagadeesh Chandra Bose, B. F. van Dongen, and W. M. van der Aalst, “Business process comparison: A methodology and case study,” in *BIS*, pp. 253–267, 2017.
6. S. J. J. Leemans, S. Shabaninejad, K. Goel, H. Khosravi, S. W. Sadiq, and M. T. Wynn, “Identifying cohorts: Recommending drill-downs based on differences in behaviour for process mining,” in *ER*, vol. 12400 of *LNCIS*, pp. 92–102, 2020.
7. S. J. J. Leemans, W. M. P. van der Aalst, T. Brockhoff, and A. Polyvyanyy, “Stochastic process mining: Earth movers’ stochastic conformance,” *Inf. Syst.*, vol. 102, p. 101724, 2021.
8. M. L. van Eck, X. Lu, S. J. J. Leemans, and W. M. P. van der Aalst, “PM²: A process mining project methodology,” in *CAiSE*, vol. 9097 of *LNCIS*, 2015.
9. J. C. A. M. Buijs and H. A. Reijers, “Comparing business process variants using models and event logs,” in *BPMDIS*, vol. 175 of *LNBIP*, pp. 154–168, 2014.
10. A. Bolt, M. de Leoni, and W. M. van der Aalst, “A visual approach to spot statistically-significant differences in event logs based on process metrics,” in *CAiSE*, pp. 151–166, Springer, 2016.
11. A. Bolt, M. de Leoni, and W. M. P. van der Aalst, “Process variant comparison: Using event logs to detect differences in behavior and business rules,” *Inf. Syst.*, vol. 74, pp. 53–66, 2018.

12. A. Bolt and W. M. P. van der Aalst, "Multidimensional process mining using process cubes," in *EMMSAD*, vol. 214 of *LNBIP*, pp. 102–116, Springer, 2015.
13. A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," *MISQ*, vol. 28, no. 1, p. 6, 2004.
14. P. Homayounfar, "Process mining challenges in hospital information systems," in *FedCSIS*, pp. 1135–1140, IEEE, 2012.
15. T. G. Erdogan and A. Tarhan, "Systematic mapping of process mining studies in healthcare," *IEEE Access*, vol. 6, pp. 24543–24567, 2018.
16. W. Yang and Q. Su, "Process mining for clinical pathway: Literature review and future directions," in *ICSSSM*, pp. 1–5, IEEE, 2014.
17. A. Partington, M. T. Wynn, S. Suriadi, C. Ouyang, and J. Karnon, "Process mining for clinical processes: A comparative analysis of four australian hospitals," *ACM Trans. Manag. Inf. Syst.*, vol. 5, no. 4, pp. 19:1–19:18, 2015.
18. F. Caron, J. Vanthienen, K. Vanhaecht, E. Van Limbergen, J. Deweerdt, and B. Baesens, "A process mining-based investigation of adverse events in care processes," *HIMJ*, vol. 43, no. 1, pp. 16–25, 2014.
19. F. Taymouri, M. L. Rosa, and J. Carmona, "Business process variant analysis based on mutual fingerprints of event logs," in *CAiSE*, vol. 12127 of *LNCS*, 2020.
20. J. C. A. M. Buijs, B. F. van Dongen, and W. M. P. van der Aalst, "Towards cross-organizational process mining in collections of process models and their executions," in *BPM Workshops*, pp. 2–13, 2011.
21. J. D. Smedt, A. Yeshchenko, A. Polyvyanyy, J. D. Weerdt, and J. Mendling, "Process model forecasting using time series analysis of event sequence data," in *ER*, vol. 13011 of *LNCS*, pp. 47–61, Springer, 2021.
22. G. Acampora, A. Vitiello, B. N. D. Stefano, W. M. P. van der Aalst, C. W. Günther, and E. Verbeek, "IEEE 1849: The XES standard," *IEEE Comput. Intell. Mag.*, vol. 12, no. 2, pp. 4–8, 2017.
23. A. Hevner, S. Chatterjee, A. Hevner, and S. Chatterjee, "Design science research in information systems," *DRISTP*, pp. 9–22, 2010.
24. K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," *MISQ*, vol. 24, no. 3, pp. 45–77, 2007.
25. J. Venable, J. Pries-Heje, and R. Baskerville, "A comprehensive framework for evaluation in design science research," in *DESRIST*, pp. 423–438, Springer, 2012.
26. D. Marutho, S. Hendra Handaka, E. Wijaya, and Muljono, "The determination of cluster number at k-mean using elbow method and purity evaluation on headline news," in *ATIC*, pp. 533–538, 2018.
27. D. J. Ketchen and C. L. Shook, "The application of cluster analysis in strategic management research: An analysis and critique," *SMJ*, vol. 17, no. 6, 1996.
28. L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
29. W. M. Van Der Aalst, "A practitioner's guide to process mining: Limitations of the directly-follows graph," 2019.
30. K. Reinhart, T. Goolsby, *et al.*, "Recognizing sepsis as a global health priority—a who resolution," *NEJM*, vol. 377, no. 5, pp. 414–417, 2017.
31. J. Wei, Z. He, C. Ouyang, and C. Moreira, "Mimicel: Mimic-iv event log for emergency department," *PhysioNet*, 2022.
32. J. C. A. M. Buijs, B. F. van Dongen, and W. M. P. van der Aalst, "On the role of fitness, precision, generalization and simplicity in process discovery," in *CoopIS*, vol. 7565 of *LNCS*, pp. 305–322, Springer, 2012.