Proceedings of the 2023 Winter Simulation Conference
C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, eds.

CLINICAL PATHWAY CLUSTERING USING SURROGATE LIKELIHOODS AND REPLAYABILITY VALIDATION

William Plumb
Alex Bottle
Giuliano Casale
Alex Liddle

Imperial College London
South Kensington
London, SW7 2AZ, UK

ABSTRACT

Modelling clinical pathways from Electronic Health Records (EHRs) can optimize resources and improve patient care, but current methods for generating pathway models using clustering have limitations including scalability and fidelity of the clusters. We propose a novel pathway modelling approach using Maximum Likelihood (ML) data clustering on Markov chain representations of clinical pathways. Our method is calibrated to produce clusters with low inter-cluster variability across the pathways. We use machine learning with Stochastic Radial Basis Functions (SRBF) kernels for surrogate optimization to handle non-convexity and propose an incremental optimization method to improve scalability. We also define a methodology based on novel replayability scores to help analysts compare the fidelity of alternative clustering results. Results show that our ML method produces clusters that have higher fidelity in terms of replayability scores than k-means based clustering and in capturing queueing contention, which is important for bottleneck identification in healthcare.

1 INTRODUCTION

Electronic health records (EHR) provide a wealth of information on clinical conditions and pathways, from which useful knowledge can be extracted. However, pathways are complex, noisy, and diverse from patient to patient. Many techniques have been developed to prize valuable insights, from process mining to the use of machine learning methods (Aspland et al. 2021). Policymakers and healthcare professionals can leverage these insights to improve patient care and optimize resource allocation.

We identify two main difficulties in current methods for clustering clinical pathways. Firstly, EHRs can be large and complex with heterogeneous pathways. Clustering methods for EHRs must be scalable, robust to noise, and capable of dealing with missing data. However, as we show in the paper, traditional methods face limitations across all these areas, such as returning erratic clustering results even on modest problem instances. Thus, the fidelity of existing methods can be a limiting factor in real-world studies.

Secondly, limited research work is available in the literature on the validation of results obtained from clustering. Understanding the meaning of a clustering algorithm result and comparing it with results obtained through other methods can be challenging and prone to errors. There is a need for scoring techniques that can help quantify the fidelity gap of a candidate clustering compared to the original trace data.

A reason behind these difficulties is that EHRs are routinely populated as part of the care provided, recorded by administrators and healthcare staff, producing large complex data sets. Complex pathways for patients thus differ on multiple levels, such as varying clinical protocols or coexisting conditions and treatment responses, making it challenging to model them. The two difficulties are also intrinsically linked. The richer and more complex the dataset is, the more it can exacerbate the accuracy of traditional

validation metrics, making it harder for them to provide a comprehensive understanding of the model fidelity. Developing novel validation techniques to study the fidelity of clustering results is thus important because clustering patients based on traditional distance metrics does not automatically imply that the resulting simulation, based on clustered data alone, will be close to one run using the original dataset.

This paper demonstrates progress in addressing the clinical pathway clustering problem by proposing a Maximum Likelihood (ML) data clustering approach to clinical pathways modelling. Leveraging advances in machine learning, the ML objective function is optimized using surrogate optimization techniques, making the approach computationally much more tractable. Recent research has shown the applicability of surrogate modelling in complex systems to reduce computational time in searching for results (Choi and Karumanchi 2022; Shen and Shoemaker 2020). Surrogates have not yet been applied to clustering clinical pathways. Additionally, surrogate approaches are typically used to estimate statistical model parameters, whereas in our study, we dynamically compute likelihood values using Markov chains constructed iteratively by aggregating recorded pathways within pathway clusters.

Yet, once supplied with a clustering result, it is difficult for analysts to decide if the clusters have high fidelity or not. Previous studies have mainly used either medical experts or face validity to corroborate that a model is representative of real-world hospitals and health systems. Other statistical methods of cluster validation such as Silhouette scores or Davies-Bouldin index, albeit popular, are unsuitable to pathway clustering due to the assumption that clusters are dense and well-separated (Rousseeuw 1987; Davies and Bouldin 1979), which we also observe on our experiments.

To enhance clustering fidelity of patient pathways, we instead generalize work on trace-driven replayability scores as a quantitative measure to validate clusters and pathways (Prodel et al. 2018). We propose a new set of scores that incorporates more information than existing scores, such as penalties on large cluster sizes, which can discriminate the less common and rare pathway transitions in the clinical paths. Moreover, our replayability scores are directly applicable to Markov chain representations of clinical pathways. The application of these replayability scores to our ML methods and k-means based clusters, the latter generated using ordinary and embedded representation of the pathways, shows that our ML method produces clusters that have the highest fidelity in capturing queueing contention, which is a key factor for bottleneck identification in clinical pathways. **Providing scalable and accurate clustering can form a foundation of for future simulation research into complex clinical pathways from EHRs.**

The remainder of the paper is as follows. Section 2 collates relevant work on the research topic. We set out our research problem in Section 3. Our novel approach is outlined in Section 4 followed by experiments and results in Section 5. Section 6 presents a discussion followed by a conclusion in Section 7.

2 RELATED WORK

Genetic algorithm (GA) is a popular optimization technique that mimics the process of natural selection and genetics to search for the optimal solution to a given problem (Golberg 1989). While GA can be effective at finding the optimal solution for some problems, including clustering, it has several limitations. GA can be computationally expensive, requiring a large number of evaluations of the objective function. Additionally, GA can get stuck in local optima and may not explore the search space efficiently.

To overcome these limitations, recent work has seen a surge of interest in surrogate optimization methods (Queipo et al. 2005). Surrogate models are used to approximate the objective function, which reduces the computational cost of evaluating the objective function. Particularly useful when approximating a noisy or expensive objective function, as they can make predictions about the optimal solution without requiring a large number of evaluations of the actual objective function.

We adopt in our methodology the Stochastic Radial Basis Function (SRBF) surrogate optimization method for the global optimization of expensive functions (Regis and Shoemaker 2007). The SRBF method is a global optimization algorithm that aims to minimize an expensive function $f(x)$, where x is a vector of input variables. The algorithm constructs an approximation of $f(x)$ using a radial basis function (RBF) model, which is defined as $f(x) \approx m(x) = \sum_{i=1}^{N} w_i \varphi(||x - x_i||)$ where N is the number of sample points,

Table 1: Summary of Related Work

Paper	Validation Method	Surrogate	Structural Clustering	Pathway Output	Replayability Scores
Zhang et al. (2015)	Support, Counts, Visualization		✓	✓	
Funkner et al. (2017)	Silhouettes, Davies-Bouldin index.		✓	✓	
Funkner et al. (2022)	Calinski-Harabasz, Davies-Bouldin index	✓	✓	✓	
Lu et al. (2019)	F1 and Visualization		✓	✓	
Huang et al. (2015)	Visualization and Cluster Homogeneity		✓		
Prodel et al. (2015)	Replayability and Visualization		✓	✓	✓
Prodel et al. (2018)	Replayability and Visualization		✓	✓	✓
De Oliveira et al. (2020)	Replayability and Visualization		✓	✓	✓
Elbattah et al. (2018)	Within-Cluster Sum of Squares, PCA			✓	
This paper	**Replayability scores and queueing error**	✓	✓	✓	✓

w_i are weights, φ is an RBF function, and $||x - x_i||$ is the Euclidean distance between x and x_i. The SRBF method begins by selecting a number of sample points at random. These points are then utilized to construct the RBF model. The algorithm then uses the model to predict the minimum of $f(x)$, which is used to select a new sample point. This selection is based on an acquisition function that balances the exploration of new regions of the search space with the exploitation of known promising regions. The most commonly used acquisition function in the SRBF method is the Expected Improvement (EI), which is defined as $EI(x) = E[\max(f_{min} - m(x), 0)]$ where f_{min} is the current best minimum value of $f(x)$. The algorithm selects the next sample point as the one that maximizes the expected improvement. After selecting a new sample point, the algorithm updates the RBF model using the new point and repeats the process until convergence is achieved. The convergence criteria can be based on a maximum number of iterations or a convergence threshold. **We select** the SRBF method **as the advantages are its efficiency and scalability. SRBF also** forms the basis for MATLAB's *surrogateopt* solver, which we have adopted throughout the paper (López 2014).

There are several other Response Surface Methods (RSMs) that are commonly implemented in literature, these include Kringing (Kleijnen 2009), Artificial Neural Networks (ANNs) (Zhang et al. 1998), Support Vector Machines (SVM) (Chang and Lin 2011) and Gaussian Process Regression (GPR) (Huang et al. 2006). One example of the Gaussian process is shown in a surrogate-assisted model reduction for stochastic biochemical reaction networks (Singh and Hellander 2017). Whereas Yi et al. (2017) uses a combination of Kringing, ANNs and SVM. Few works consider the intersection of using surrogates and clinical pathway clustering, but significant research has been conducted in both individual fields. Table 1 shows recent works in clinical pathway clustering that we discuss below.

Zhang et al. (2015) uses a heuristic approach to extract patient pathways from one-dimensional sequences that represent multi-dimensional clinical pathways. However, our alternative approach aims to provide a more comprehensive representation of clinical pathways by using ML and surrogate optimization. Funkner et al. (2017) proposes a method that converts sets of patient journeys into templates using a multiple objective function to find the best-performing silhouette score and locate a representative pathway. **Further work utilises a surrogate to predict the performance of potential clusterings Funkner et al. (2022).** However, our work proposes using surrogates to search for the clustering itself, given their positives in large search spaces. Lu et al. (2019) presents a method that uses sequences to cluster patients by defining scores to check traces against sequence patterns, we propose the use of unsupervised methods to find clusters and validates them using trace-based metrics. Papers Prodel et al. (2015), Prodel et al. (2018), De Oliveira et al. (2020) demonstrate the use of process mining in clinical pathway analysis to improve understanding of a system, identify inefficiencies, and inform process improvements. While they suggest using a single process model to represent a system, we propose clustering patient pathways to produce sub-population process models that offer more granularity and reveal unique patterns within each group. Clustering can also inform process improvements and resource allocation by segmenting the system into meaningful groups.

Our approach utilizes the replayability concept of process models, but with an alternative representation that allows for adaptive clustering.

3 PROBLEM STATEMENT

A *clinical* pathway is represented by a set of ordered *events*. Each patient experiences a unique set of sequential *events*. These *events* hold information on the activity undertaken. Our work takes the following assumptions on the structure of the pathways. The first is that clinical events can only occur one at a time and not in parallel, which means that the patient pathway is a series of sequential events without any forks or splits. This assumption simplifies the model and allows for easier analysis of the patient pathway. The second assumption is to model events in discrete time, i.e., each time step corresponds to the progression to the next clinical event; this assumption may be readily relaxed by coupling each step with a duration. Finally, we assume that patients can start their pathway with any event rather than with a specific one.

Medical events span the integrated healthcare system, from GPs to hospitals. Consider a set N of medical events $N = \{n_1, \ldots, n_{|N|}\}$ where N is the set of all possible nodes. Each event is represented by an encounter code that describes the type and reason for the clinical event. As a patient can have multiple events in a visit, we use the notion of *encounter mapping* to represent them as a single episode or event. For example, a patient might suffer a cardiac arrest which outputs multiple events on the same day. These might include events for cardiac arrest, acute myocardial infarction, echocardiography and many more procedures. The encounter mapping will reduce this to one encounter that represents the events, in this example it would be summarized to a single event called cardiac arrest.

We define a set of pathways or *routes* R. Each pathway r represents a sequence of clinical events that occur in a specific order. We preserve the timestamps of the events to maintain the logical dependencies between them. A pathway r can be represented as a sequence of nodes $r =< n_1, \ldots, n_s, \ldots, n_e >$, where n_s is the node visited by the patient at step s and n_e is the end node in the pathway. We hypothesize that subsets of R can be identified through clustering to locate similar pathways.

Patients can undertake multiple different pathways r throughout their lifespan. To locate clinically specific pathways, a timeout threshold is created to break patient records into sub-pathways. Should the transition between 2 events breach the timeout threshold, the pathway will finish, and a separate pathway will begin. This threshold allows the starting node to be any stage of the patient clinical pathway, depending on the specific clinical scenario. Similarly, the endpoint of the route is determined by identifying the next clinical event that occurs after the time threshold. This approach avoids the need for extensive manual mapping of clinical start or end events and provides a standardized method for determining the route endpoints.

To represent a patient pathway, we express it via a transition matrix. Transition matrix T_r, is the counts of transitions moving from event n_i to event n_j producing the element $T_{r,i,j}$. Hence, $T_r = [T_{r,i,j}]$, $1 \leq r \leq R$, $1 \leq i, j \leq N$, represents the transition matrix for pathway r, where each element is the transition between node i and j. In the baseline methods we compare with, we apply clustering methods directly using the transition matrix T_r. Aggregating transition matrices across a cluster will give a single transition matrix T_G. The transition matrix can be converted from representing count of transitions to probabilities by dividing each value by the sum of the row value. We use P_G as the probability matrix, for a cluster group G.

4 METHODOLOGY

To create likelihood-based pathway clustering, we proposed a method using a Maximum Likelihood (ML) function to calculate the probability of a pathway being represented by a cluster. The likelihood for a pathway r is calculated as follows.

$$L(r, P_G) = \log \theta_{r_1} + \sum_{i=2}^{n} \log(P_{r_{i-1}, r_i}) \tag{1}$$

Here, θ is a one-dimensional vector of initial state probabilities for the Markov chain that models the clustered pathways, and we take the probability θ_{r_1} of the starting node r_1 from θ. P_G is an $N \times N$-dimensional probability matrix of the Markov Chain representing the cluster G, created from aggregating all pathways within the cluster. We aggregate the log of the starting probability with the sum of log probability given from P_G of the transitions in r.

Each cluster has a subset of the pathways set R, which we will call R_g for cluster g. After calculating each pathway r likelihood within their cluster, we aggregate the scores. We want the clustering to group together pathways that have the least variability, hence we also add a penalty score. Using the Squared Coefficient of Variation (SCV), where the SCV of a random variable X is $SCV = Var(X)/E(X)^2$. SCV weighted by a parameter W, to penalize the likelihood based on the degree of variability between routes and their lengths. SCV is calculated for each cluster where the value is the variance of the pathway lengths divided by the square of its mean pathway length. Equation (2) shows the ML objective we maximize.

$$ML(R,P) = \sum_{g=1}^{|G|} \sum_{i=1}^{|R_g|} L(R_{g_i}, P_g) + W \cdot \log\left(SCV(R,P)\right) \tag{2}$$

where $L(R_{g_i}, P_g)$ is computed using Equation (1). The $SCV(R,P)$ value is computed by averaging over the SCV score of each cluster.

4.1 Likelihood Maximization

The ML optimization program searches combinations of pathway groupings that maximize Equation (2). In particular, the decision variables are the cluster memberships, which are integer variables. Based on the current membership vector, the optimization program recomputes the R_g sets for each cluster and the associated value of $ML(R,P)$. This, in general, poses a high computational demand, due to the need to regenerate at each iteration the probability matrix P_g of the Markov chain underpinning the aggregate clinical pathway of each cluster g.

Our experiments adopt different likelihood maximization approaches to reduce computational demand. In the basic approach, we maximize likelihood using GA. To aid GA in finding a global optimal, we supply an initial population given by the k-means solution, which can be rapidly obtained. As highlighted in Section 2, GA has limitations including computational demand, which limits the scalability of the method for use with healthcare datasets with more than a few tens of pathways. Throughout, the GA-based clinical clustering method is simply referred to as ML.

Thus, we explore the use of machine learning based optimization using SRBF kernels for surrogate optimization to counteract the limitations of GA such as computational cost and non-convexity. The initialization of the SRBF-based surrogate optimization solver requires a representative sample proportional to two times the number of pathways we wish to cluster before it can search for an optimal solution. Henceforth, the SRBF-based clinical clustering method is referred to as SurrML.

One limitation of the surrogate approach is that needs to learn the response surface, which requires a considerable number of random samples on larger problems. Moreover, the method can have large memory requirements with just a few hundred pathways. Hence, we propose a third approach to optimizing the ML where we sequentially assign memberships for a subset of the pathways, while keeping the other memberships fixed. This method, referred to as Incremental ML, for short SurrML-INCR, divides the pathways into fix sized groups and conducts the ML optimization on each group sequentially. Initially, the first group of patients is clustered. Then, the cluster memberships of the first group are fixed and pathways of the second group are added and clustered, maximizing the likelihood for the new group. This iteratively proceeds until all groups are initially clustered. After the first complete iteration, it is possible to cycle back to earlier groups, fixing in a similar fashion the cluster memberships of all groups not presently considered, until a maximum number of iterations is reached. In this way, each ML optimization changes memberships

only for a fraction of patients in the original dataset. Using smaller datasets resulting from this approach significantly increases the speed in calculating the clusters, while at the same time requiring less memory.

4.2 Replayability Scores

In this work, we focus on evaluating the replayability of clusters derived from clinical pathways described by Markov chains. Previous process mining studies have applied replayability analysis to processes in healthcare (Prodel et al. 2015; Prodel et al. 2018; De Oliveira et al. 2020). The published metrics are implemented using process modelling variables, our methods use an alternative representation, namely probability matrices and transition pairs. Modification of all previous metrics are required as well as additional metrics that reflect our clustering problem.

We introduce six new Replayability Scores (RS) to be used with individual pathways r and with probability matrices of the clustered Markov chains. The final metric for each RS is obtained by averaging the scores for each cluster before creating a weighted average across all clusters. RS1 is the percentage of transitions for a single pathway represented in the cluster probability matrix P_G.

$$RS1(P_G, r) = \frac{1}{|r|} \sum_{i=1}^{|r|-1} 1(\langle (r_i, r_{i+1}) \rangle \in P_G) \tag{3}$$

Here $1\langle (r_i, r_{i+1}) \rangle \in P_G)$ represents a binary score of 1 if the transition in the pathways r_i to r_{i+1} is represented in the clustered probability matrix P_G.

RS2 is a measure representing the percentage of events played, modified by a node skipping penalty. A variable $\eta(P_G, r)$ is a count of the number of events in the pathway that are not featured in the cluster, scaled by the penalty weight β.

$$RS2(P_G, r) = \left(RS1 - \beta \frac{\eta(P_G, r)}{|r|} \right)^+ \tag{4}$$

Using $|r|$ as the total number of transitions in pathway r. RS1 does not include a penalty for skipping events, the addition of the skipping penalty in RS2 will help distinguish if the pathways in a cluster are not well represented. RS3 goes through the full route r and averages the probabilities of each of the transition r_i to r_{i+1}, given from the aggregated cluster probability matrix P_G. This metric penalizes trace transitions that have a low probability of occurring in the cluster.

$$RS3(P_G, r) = \frac{1}{|r|} \sum_{i=1}^{|r|-1} \log(P_{G_{i,i+1}}) \tag{5}$$

The next score, RS4, corrects RS3 using a term with weight γ that accounts for the number of nodes represented in the cluster but not visited in the trace. We refer to this count of unused nodes as $\theta(P_G, r)$. This will show where clustering methods place the majority of pathways into one group, as the cluster will lack the specificity to individual pathways. In addition, we adapt RS4 to incorporate RS1 instead of the probability and apply the penalty, calling the resulting term RS5.

$$RS4(P_G, r) = RS3 - \gamma\theta(P_G, r) \tag{6}$$

$$RS5(P_G, r) = RS1 - \gamma\theta(P_G, r) \tag{7}$$

Lastly, we create a discrete time Markov chain using the probability matrix and simulate an artificial pathway, named ρ, with the same length and starting probability as the pathway r. RS6 counts of the number of times each node is visited with the artificial pathway and compare the values against the individual pathway. This score will assist understanding if the probability of the cluster can create an artificial pathway similar to

the actual pathway. $|n_r|$ is the unique number of nodes visited by trace r with n_{r_i} representing the number of visits to node r_i.

$$RS6(T_G, r) = \frac{1}{|n_r|} \sum_{i=1}^{|n_r|} 1 - \frac{|n_{r_i} - n_{\rho_i}|}{max(n_{r_i}, n_{\rho_i})} \tag{8}$$

The main advancement of our proposed metrics to the scores presented by Prodel et al. (2018), is the application of probability matrices instead of evaluating events. As previous scores have been used for process mining, the focus has been on evaluating single process models. These previous scores are not applicable to clustering as the results using these create misleading results not suitable for modelling. **The introduced replayability scores are created for generic pathway clustering, with each score providing differing information on the cluster validity.** To validate our methods, we provide an average across all replayability scores as an overall RS score. **Clinical application can have tailored replayability scores based from clinical expertise to assist validation without the need for manual face validity checking.**

5 EXPERIMENTS

We implement multiple baselines to verify that our ML method provides an improvement to clustering of clinical pathways compared to simpler techniques such as k-means. The first result is a baseline result which places all pathways into one cluster, named *All1Cluster*. The second baseline method is a straightforward application of k-means clustering to the vectorized transition matrices, denoted as *Kmeans*.

Advancements in AI embedding have offered alternative representations in a learned latent space. One relevant method associated with graphical representation and embedding is Graph2Vec (G2V) (Narayanan et al. 2017). G2V learns structural similarities between sub-graphs, turning a clinical pathway of transition pairs into a one-dimensional vector in the learned latent space. The embedded representation should hold more structural information, such as node similarities, within the output vector. As an additional baseline, we use Graph2Vec representation of pathways with k-means to assess how an AI embedding approach can cluster our clinical pathways. This method is referred to in the results as *G2V-Kmeans*. Our proposed methods of ML using GA, surrogate and iterative surrogate are termed *ML, SurrML* and *SurrML-INCR*.

5.1 Validation Dataset

Our data for validating our methodology is generated by Synthea (Walonoski et al. 2018), which is an open-source software tool that generates synthetic patient data for research purposes. The tool is designed to create a diverse patient population with different medical histories, demographics, and clinical profiles. **The reader should refer to Synthea's paper and documentation for more details**.

We implement a basic set of filters and data cleansing steps to pre-process the data. 1) Data reduction to focus on clinical events that take place. 2) Timeout function to locate sub-pathways of EHRs that span many years. 3) Limiting medical code mapping to ensure detailed event relationships are kept.

Selecting a timeout of 180 days to create sub-graphs of clinical pathways, we generate a set of paths. We assume, with some domain knowledge, that using half a year as a threshold will prevent unnecessarily long clinical pathways of multiple medical conditions. Using the set of pathways, we create three Case Studies (CS) to evaluate our clustering and validation:

- CS1: Manually selecting three event classes that we expect can be clustered. Selected conditions for the classes are presentation for pain, a stroke and admission for surgery. Pathways within each class can differ in length and routes based off simulated pathways generated in Synthea. We randomly selected 10 pathways, which do not have more than one of the selected conditions, from each of these event classes such that we achieve a total of 30 pathways. The data for this case study has an average pathway length of 90.27 transitions, with 72 unique nodes across the pathways. An average of 14.8 unique transitions are made by each pathway.

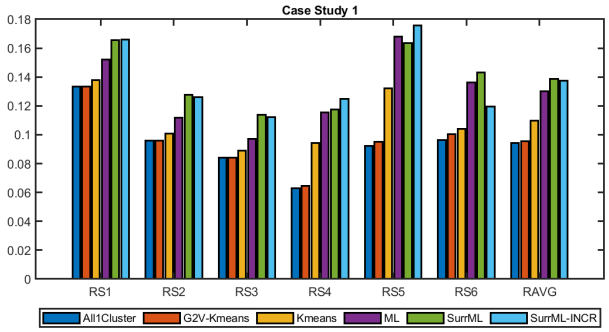


Figure 1: Case Study 1 Replayability Scores

Table 2: Case Study 1 Clusters (30 pathways)

Method	C1	C2	C3
Kmeans	26	1	3
ML	9	12	9
G2V Kmeans	1	29	0
SurrML	20	5	5
All1Cluster	30	0	0
SurrML-INCR	10	11	9

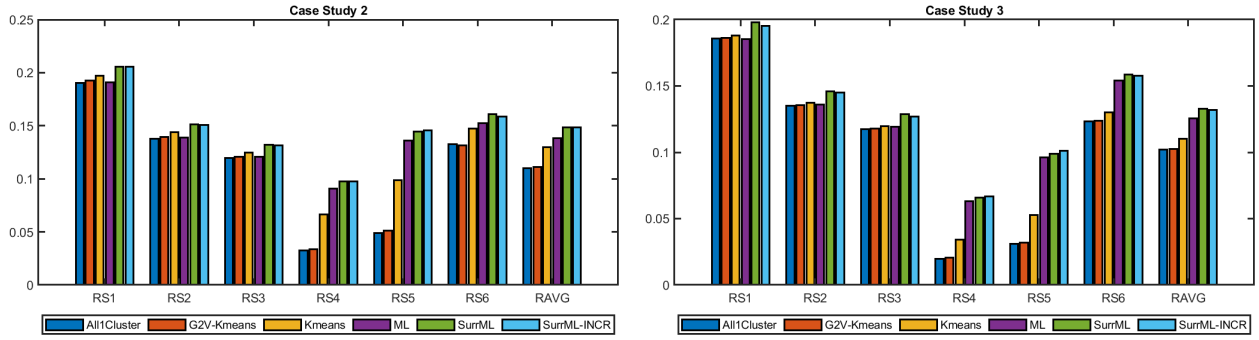
- CS2: We expand the dataset to include more clinical pathways. For case study two, we use 100 random pathways of any specialty given from Synthea to see if reasonable clusters can be found given the additional pathways. As we increase the number of pathways, the average pathway length drops to 37.94, with 100 unique nodes visited across the pathways. An average of 11.13 unique transitions are made by each pathway.
- CS3: To assess scalability, we increase the number of random pathways to 250. The average pathway length further decreases to 34.05 transitions per pathway. The number of unique nodes in the dataset increases to 121 and the average unique transitions per pathway is 9.86.

For CS1 we expect three clusters to be found, we limit the clustering methods to a maximum of three clusters. For CS2 and CS3, the optimal number of clusters is unknown. We set an arbitrary limit of clusters to ten in both case studies. Graph2Vec embedding is trained on a random sample of 5000 pathways generated by Synthea using the following parameters: Output dimensions 128, epochs 10, minimal feature 5, Feature extraction recursions 1, Initial learning rate. 0.025, Downsampling rate 0.0001.

In this experimental setup, several parameters were defined to ensure reliable and accurate results. Firstly, the surrogate model was configured with an SCV penalty weight (W) of 1. Additionally, both the genetic algorithm (GA) and surrogate model were initialized using the output from k-means clustering. The optimal number of clusters in k-means was determined using silhouette scores. For the incremental surrogate model, we employed a fixed iteration size of 30 pathways to enable a more efficient and targeted search for the optimal solution. Calculations of the replayability scores use the parameters; $\beta = 0.5 \cdot \sqrt{|u(r)|}$, $\gamma = 3 \cdot |u(r)|$ where $|u(r)|$ is the number of unique transitions for pathway r.

5.2 Clustering Results

In CS1 involving 30 pathways from three manually selected groups shown in Figure 1, it is clear that both using genetic algorithm and surrogate ML solvers outperform traditional clustering methods across all the replayability scores. More importantly, how our proposed methods separate pathways across possible clusters are realistic. Table 2 shows k-means grouped 26 of the 30 pathways into one group with G2V k-means also grouping the majority of pathways into one cluster. This theme of unbalanced clusters continues into the second and third case study where k-means produces groupings where a substantial proportion of pathways are in one group. The new RS scores produced in this paper give more information on how well a cluster represents each pathway across multiple classes. Scores RS4 and RS5 penalize these majority one cluster groups heavily as they fail to represent the underlying clinical pathways. A key motivation of using surrogate optimization methods over GA is reinforced in our largest case study, where GA becomes caught in a local minimum and ceases to search further. The results of the case studies underline the improvements of using surrogate optimization. Shown in the result figures is the scoring of our proposed incremental surrogate method. It is seen that using incremental optimization is equivalent to



(a) CS2 (b) CS3

Figure 2: Replayability Scores for Case Studies 2 and 3

the complex full surrogate search. Given its lower computational demand, it is scalable to a larger dataset where we can expect a large volume of clinical pathways.

5.2.1 Queueing Prediction Results

Using replayability validation, our methods may be used to inform queueing-based methods to model the healthcare system delays and bottlenecks. To model a healthcare system accurately, there is however a trade-off between computational cost and accuracy. Modelling each individual pathway as its own class of jobs in a queueing model may be very costly or infeasible at even small numbers, such as hundreds of pathways, when the system consists of many nodes. Clustering pathways into aggregated groups alleviates this computational strain should the clustering be accurate.

To further confirm our methods are the most acceptable for queueing prediction, we evaluate the error in predicted mean queue length values between the clustered and unclustered pathway models when we model the nodes of the pathway Markov chain as queueing stations in a healthcare system. Thus, the probabilities of visiting nodes in the pathway translate into the routing probability matrix of the queueing network. The resulting Markovian queueing network may be analyzed using stochastic analysis techniques, such as the Mean-Value Analysis (MVA) algorithm, which returns mean queue-length predictions (Reiser and Lavenberg 1980). This method gives exact results for the class of models under consideration, and it is in practice equivalent to a steady-state simulation evaluation. We ask the question on how the queueing network model predictions differ from the original trace when data is clustered using the considered approaches. Based on the proposed replayability scores, we expect ML methods to win over k-means based techniques. To answer this question, we generate 100 random queueing networks differing only for the service times of each encounter, which is assumed exponentially distributed with random means. The expectation is that the methods with the largest replayability scores are robust across various scenarios, so a randomized validation of this kind can give insight into the robustness of the clusters irrespective of service time parameters. Table 3 shows the mean error scores for each case study and methods. The mean is averaged across the 100 random queueing networks. Here a lower score represents less error. The table shows our ML methods based on GA and surrogate optimization have the least error when evaluated using MVA.

6 DISCUSSION

To understand the contribution of the proposed methods and validation techniques, we evaluated the clustering with some traditional validation metrics used in related work. The metrics we assessed include log-likelihood, entropy, Silhouette, and Davies-Bouldin scores shown in Table 4. G2V k-means fails to calculate the log likelihood in CS1 due to all pathways bar one being grouped together. All Scores except log

Table 3: Queueing network model errors on 100 random models. The lowest error is marked in bold.

Experiment	Kmeans	ML	G2V-Kmeans	SurrML	All 1 Cluster	SurrML-INCR
CS1	0.86	0.37	0.99	0.59	1.07	**0.36**
CS2	4.74	0.89	9.10	**0.58**	11.26	0.83
CS3	15.41	1.36	19.25	**1.13**	27.16	1.23

Table 4: Evaluation of alternative validation metrics. The best performing method is in bold, showing the difficulty in inferring the best method using traditional metrics. Instead, RAVG identifies the general high performance of SurrML and SurrML-INCR with a high correlation with Table 3.

Dataset	Method	Clusters	Log Likelihood	Mean Entropy	Silhouette	Davies-Bouldin	RAVG
CS1	kmeans	3.000	-272.23	0.045	0.34	1.36	0.11
CS1	G2V-kmeans	2.000	NA	0.0023	**0.63**	**0.36**	0.10
CS1	ML	3.000	**-234.18**	**0.049**	-0.031	3.97	0.13
CS1	SurrML	3.000	-244.96	0.049	0.086	3.35	**0.14**
CS1	SurrML-INCR	3.000	-236.20	0.048	-0.038	3.80	**0.14**
CS2	kmeans	6.000	-480.95	0.050	0.14	1.63	0.13
CS2	G2V-kmeans	2.000	-3475.91	**0.21**	**0.38**	**0.61**	0.11
CS2	ML	10.000	-218.17	0.044	-0.15	5.04	0.14
CS2	SurrML	10.000	-216.54	0.043	-0.093	4.90	**0.15**
CS2	SurrML-INCR	10.000	**-210.06**	0.043	-0.14	4.79	**0.15**
CS3	kmeans	2.000	-3446.12	0.18	0.13	2.79	0.11
CS3	G2V-kmeans	2.000	-7350.02	**0.31**	**0.45**	**0.57**	0.10
CS3	ML	10.000	-517.36	0.073	-0.086	7.46	**0.13**
CS3	SurrML	10.000	-509.26	0.071	-0.083	6.92	**0.13**
CS3	SurrML-INCR	10.000	**-498.27**	0.071	-0.11	6.63	**0.13**

likelihood suggested *k*-means and G2V *k*-means perform best. In reality, they produce almost meaningless pathway groupings. Similarly, purely taking the log likelihood score in smaller datasets, for example in CS1, are similar across all methods. These results can be misleading as *k*-means still performs well. Validation using replayability removes this concern as it shows which method produces high fidelity clusters.

Ensuring our new proposed replayability scores are improved from previous literature, evaluation is conducted of the proposed clustering results using metric shown in (Prodel et al. 2018). For all experiments, the previous scores constantly return *k*-means and G2V *k*-means significantly higher when averaging across scores. This is due to the clustering methods returning a sizable proportion of pathways in one group. Comparing with the queueing modelling results shown in Table 3 suggest the previous replayability scores do not represent clusters of clinical pathways suitable for simulations.

Our findings demonstrate that optimizing ML data clustering using surrogate optimization and replayability scores provides a more effective approach for clustering clinical pathways than *k*-means clustering. The replayability scores provided a more robust validation method, ensuring that the clustering model reflects real-world performance, reinforced with the queueing error validation metric.

While our study shows promising results, there are limitations to consider. The use of synthetic EHRs with limited clinical guidance may not fully reflect the complexity and heterogeneity of real-world patient populations. Additionally, the scalability of the approach may need further investigation to assess its effectiveness with larger datasets. Deeper analysis into node importance and weightings could improve clustering accuracy further as many common transitions cloud indicative pathways. Calculating replayability score can be computationally expensive on larger datasets. Further investigation into optimizing the scores can be made to scale the methods further. During our clustering methods, we arbitrarily select the upper cluster limit. Future work will investigate if *k*-means can be used to dictate the range for the number of clusters using replayability scores.

We investigated initiating both GA and Surrogate methods using k-means and G2V k-means results to assess how the differing starting points affect clustering performance. As they both failed to cluster accurately, there was no benefit from initiating using one or the other. Optimization starting points can have a significant effect on the speed to find optimal solutions, future work will produce a more efficient initiation search population to aid large searches. However, the introduction of the incremental surrogate search is less susceptible to poor starting points as the search is a fraction of the full search space.

Despite these limitations, our study provides valuable insights using the application of ML optimization and replayability scores for clinical pathway clustering. **These resulting clusters can direct simulation-based methods towards accurately and efficiently representing the system. The model will enable policy makers and clinicians to gain insights into the underlying structures, facilitating their understanding key issues such as health inequalities in care. This approach has the potential to be extended for investigating bottlenecks and conducting what-if planning within the healthcare pathway system.**

7 CONCLUSION

This study presents a novel method for clustering clinical pathways that improves upon existing techniques. By utilizing ML data clustering and introducing optimization methods such as surrogate optimization, we achieved improved accuracy and scalability. Additionally, our introduction of replayability scores for validation ensured that our method provided meaningful clusters from complex sub-populations. The results demonstrate that our method produces the least queueing network model errors and can be used for simulation-based methods in the future. This study provides a valuable contribution to the field of clinical pathway clustering and paves the way for further research to optimize patient care and resource allocation in healthcare settings.

ACKNOWLEDGMENTS

William Plumb is supported by UK Research and Innovation (UKRI) Centre for Doctoral Training in AI for Healthcare grant number EP/S023283/1.

REFERENCES

Aspland, E., D. Gartner, and P. Harper. 2021. "Clinical Pathway Modelling: A Literature Review". *Health Systems* 10(1):1–23.

Chang, C.-C., and C.-J. Lin. 2011. "LIBSVM: A Library For Support Vector Machines". *ACM Transactions on Intelligent Systems and Technology (TIST)* 2(3):1–27.

Choi, M. C., and V. Karumanchi. 2022. "Landscape Modification Meets Surrogate Optimization: Towards Developing an Improved Stochastic Response Surface Method". In *2022 Winter Simulation Conference (WSC)*, 3206–3216: IEEE.

Davies, D. L., and D. W. Bouldin. 1979. "A cluster separation measure". *IEEE transactions on pattern analysis and machine intelligence* (2):224–227.

De Oliveira, H., V. Augusto, B. Jouaneton, L. Lamarsalle, M. Prodel, and X. Xie. 2020. "Optimal Process Mining of Timed Event Logs". *Information Sciences* 528:58–78.

Elbattah, M., O. Molloy, and B. P. Zeigler. 2018. "Designing Care Pathways Using Simulation Modeling And Machine Learning". In *2018 Winter Simulation Conference (WSC)*, 1452–1463: IEEE.

Funkner, A., A. Yakovlev, and S. Kovalchuk. 2022. "Surrogate-Assisted Performance Prediction for Data-Driven Knowledge Discovery Algorithms: Application To Evolutionary Modeling of Clinical Pathways". *Journal of Computational Science* 59:101562.

Funkner, A. A., A. N. Yakovlev, and S. V. Kovalchuk. 2017. "Towards Evolutionary Discovery of Typical Clinical Pathways in Electronic Health Records". *Procedia Computer Science* 119:234–244.

Golberg, D. E. 1989. "Genetic Algorithms in Search, Optimization, and Machine Learning". *Addion Wesley* 1989(102):36.

Huang, D., T. T. Allen, W. I. Notz, and N. Zeng. 2006. "Global Optimization of Stochastic Black-Box Systems via Sequential Kriging Meta-Models". *Journal of Global Optimization* 34(3):441–466.

Huang, Z., W. Dong, F. Wang, and H. Duan. 2015. "Medical Inpatient Journey Modeling and Clustering: A Bayesian Hidden Markov Model Based Approach". In *AMIA Annual Symposium Proceedings*, 649: American Medical Informatics Association.

Kleijnen, J. P. 2009. "Kriging Metamodeling in Simulation: A Review". *European journal of operational research* 192(3):707–716.

López, C. P. 2014. "Optimization Techniques via the Optimization Toolbox". *MATLAB Optimization Techniques*:85–108.

Lu, X., S. A. Tabatabaei, M. Hoogendoorn, and H. A. Reijers. 2019. "Trace Clustering on Very Large Event Data in Healthcare Using Frequent Sequence Patterns". In *Business Process Management: 17th International Conference*, 198–215: Springer.

Narayanan, A., M. Chandramohan, R. Venkatesan, L. Chen, Y. Liu, and S. Jaiswal. 2017. "graph2vec: Learning Distributed Representations of Graphs". *arXiv preprint arXiv:1707.05005*.

Prodel, M., V. Augusto, B. Jouaneton, L. Lamarsalle, and X. Xie. 2018. "Optimal Process Mining for Large and Complex Event Logs". *IEEE Transactions on Automation Science and Engineering* 15(3):1309–1325.

Prodel, M., V. Augusto, X. Xie, B. Jouaneton, and L. Lamarsalle. 2015. "Discovery of Patient Pathways From a National Hospital Database Using Process Mining and Integer Linear Programming". In *2015 IEEE International Conference on Automation Science and Engineering (CASE)*, 1409–1414: IEEE.

Queipo, N. V., R. T. Haftka, W. Shyy, T. Goel, R. Vaidyanathan, and P. K. Tucker. 2005. "Surrogate-Based Analysis and Optimization". *Progress in Aerospace Sciences* 41(1):1–28.

Regis, R. G., and C. A. Shoemaker. 2007. "A Stochastic Radial Basis Function Method for the Global Optimization of Expensive Functions". *INFORMS Journal on Computing* 19(4):497–509.

Reiser, M., and S. S. Lavenberg. 1980. "Mean-value Analysis of Closed Multichain Queuing Networks". *Journal of the ACM (JACM)* 27(2):313–322.

Rousseeuw, P. J. 1987. "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis". *Journal of Computational and Applied Mathematics* 20:53–65.

Shen, Y., and C. A. Shoemaker. 2020. "Global Optimization for Noisy Expensive Black-Box Multi-Modal Functions via Radial Basis Function Surrogate". In *2020 Winter Simulation Conference (WSC)*, 3020–3031: IEEE.

Singh, P., and A. Hellander. 2017. "Surrogate Assisted Model Reduction for Stochastic Biochemical Reaction Networks". In *2017 Winter Simulation Conference (WSC)*, 1773–1783: IEEE.

Walonoski, J., M. Kramer, J. Nichols, A. Quina, C. Moesel, D. Hall, C. Duffett, K. Dube, T. Gallagher, and S. McLachlan. 2018. "Synthea: An Approach, Method, and Software Mechanism for Generating Synthetic Patients and the Synthetic Electronic Health Care Record". *Journal of the American Medical Informatics Association* 25(3):230–238.

Yi, W., J. Zhong, S. Tan, W. Cai, and N. Hu. 2017. "Surrogate Assisted Calibration Framework for Crowd Model Calibration". In *2017 Winter Simulation Conference (WSC)*, 1216–1227: IEEE.

Zhang, G., B. E. Patuwo, and M. Y. Hu. 1998. "Forecasting with Artificial Neural Networks: The State of the Art". *International Journal of Forecasting* 14(1):35–62.

Zhang, Y., R. Padman, and N. Patel. 2015. "Paving the COWpath: Learning and Visualizing Clinical Pathways from Electronic Health Record Data". *Journal of Biomedical Informatics* 58:186–197.

AUTHOR BIOGRAPHIES

WILLIAM PLUMB is a PhD student within the UKRI Centre for Doctoral Training in AI for Healthcare at Imperial College London, his research focuses on the use of artificial intelligence (AI) methods to understand and optimize clinical pathways design. Using AI-based methods such as simulations, his research seeks to improve the design of clinical pathways to make them more efficient, effective, and clear. His email address is w.plumb20@imperial.ac.uk.

ALEX BOTTLE is a Professor of Medical Statistics in the School of Public Health at Imperial College London. His research focus is on using routinely collected data to measure and explain variations in health service quality and safety with the aim of improving patient care. He devised a national hospital mortality monitoring system that detected the problems at Mid Staffordshire Hospital and elsewhere. Having published over 400 peer reviewed papers, he serves as an Associate Editor of BMJ Quality & Safety and Deputy Statistics Editor at Thorax. His email address is robert.bottle@imperial.ac.uk.

GIULIANO CASALE joined the Department of Computing at Imperial College London in 2010, where he is currently a Reader. Previously, he worked as a scientist at SAP Research UK. He teaches and does research in performance engineering and cloud computing, topics on which he has published more than 150 refereed papers. He has served on the technical program committee of over 100 conferences and workshops. He serves on the editorial boards of IEEE TNSM and ACM TOMPECS and as current chair of ACM SIGMETRICS. His email address is g.casale@imperial.ac.uk.

ALEX LIDDLE is Clinical Senior Lecturer in Orthopaedic Surgery at the MSK Lab at Imperial College London. He also holds a role as a Consultant Orthopaedic Surgeon specializing in knee and hip surgery at St Mary's and Charing Cross Hospitals. His principal research interest is the use of technology and big data to optimize outcomes and minimize complications after joint replacement. Having published in all major orthopaedic journals, he provides expert review for several orthopaedic journals such as the BMJ and the National Institute for Health Research, He is on the Editorial Board of the Bone and Joint Journal. His email address is a.liddle@imperial.ac.uk.