Sheffield Hallam University

Applying sequence analysis to uncover 'real-world' clinical pathways from routinely collected data: a systematic review

MATHEW, Smitha, PEAT, George <http://orcid.org/0000-0002-9008-0184>, PARRY, Emma, SOKHAL, Balamrit Singh and YU, Dahai

Available from Sheffield Hallam University Research Archive (SHURA) at:

https://shura.shu.ac.uk/34420/

This document is the author deposited version.

Published version

MATHEW, Smitha, PEAT, George, PARRY, Emma, SOKHAL, Balamrit Singh and YU, Dahai (2024). Applying sequence analysis to uncover 'real-world' clinical pathways from routinely collected data: a systematic review. Journal of clinical epidemiology, 166: 111226. [Article]

Copyright and re-use policy

See http://shura.shu.ac.uk/information.html

Sheffield Hallam University Research Archive
http://shura.shu.ac.uk

Journal of Clinical Epidemiology 166 (2024) 111226

Journal of Clinical Epidemiology

ORIGINAL RESEARCH

Applying sequence analysis to uncover 'real-world' clinical pathways from routinely collected data: a systematic review

Smitha Mathew[a], George Peat[a,b], Emma Parry[a], Balamrit Singh Sokhal[a], Dahai Yu[a,*]

[a]*School of Medicine, Keele University, Staffordshire, UK*

[b]*Centre for Applied Health & Social Care Research, Sheffield Hallam University, Sheffield, UK*

Accepted 27 November 2023; Published online 28 November 2023

Abstract

Objectives: This systematic review aims to elucidate the methodological practices and reporting standards associated with sequence analysis (SA) for the identification of clinical pathways in real-world scenarios, using routinely collected data.

Study Design and Setting: We conducted a methodological systematic review, searching five medical and health databases: MEDLINE, PsycINFO, CINAHL, EMBASE and Web of Science. The search encompassed articles from the inception of these databases up to February 28, 2023. The search strategy comprised two distinctive sets of search terms, specifically focused on sequence analysis and clinical pathways.

Results: 19 studies met the eligibility criteria for this systematic review. Nearly 60% of the included studies were published in or after 2021, with a significant proportion originating from Canada ($n = 7$) and France ($n = 5$). 90% of the studies adhered to the fundamental SA steps. The optimal matching (OM) method emerged as the most frequently employed dissimilarity measure (63%), while agglomerative hierarchical clustering using Ward's linkage was the preferred clustering algorithm (53%). However, it is imperative to underline that a majority of the studies inadequately reported key methodological decisions pertaining to SA.

Conclusion: This review underscores the necessity for enhanced transparency in reporting both data management procedures and key methodological choices within SA processes. The development of reporting guidelines and a robust appraisal tool tailored to assess the quality of SA would be invaluable for researchers in this field.

Plain Language Summary: Clinical pathways (CPs) are like detailed plans for treating specific diseases or medical conditions. They help doctors and healthcare teams provide effective, evidence-based, and safe care to patients. However, clinical pathways are not consistently followed in clinical practice, especially for chronic conditions. Methods such as sequence analysis are used to identify 'real-world' patterns of care from patients' electronic health records. This study identified and reviewed 19 publications using sequence analysis to explore 'real world' clinical pathways. Despite the potential of sequence analysis, the methods used varied greatly. This review highlights the need for clearer reporting and some additional guidelines for researchers intending to use sequence analysis to explore clinical pathways using real-world patient data. *© 2023 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).*

Keywords: Clinical pathways; Care trajectories; Sequence analysis; Optimal matching; Cost setting; Cluster analysis

1. Introduction

A clinical pathway (CP), also known as a care pathway, serves as a multidisciplinary care plan for a specific disease or medical condition that outlines effective treatment for a patient [1]. CPs are widely used tools in evidence-based healthcare, facilitating the translation of clinical practice guideline recommendations into specific care processes,

Funding: EP is funded by a National Institute for Health and Care Research (NIHR) Academic Clinical Lectureship CL-2020-10-001. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

* Corresponding author. Primary Care Centre Versus Arthritis, School of Medicine, Keele University, Staffordshire, UK ST5 5BG. Tel.: +44-0-1782-734891; fax: +44-0-1782-734719.

E-mail address: d.yu@keele.ac.uk (D. Yu).

https://doi.org/10.1016/j.jclinepi.2023.111226

0895-4356/© 2023 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

What is new?

Key Findings

- Sequence analysis (SA) represents an innovative approach within healthcare research, increasingly applied in recent years to uncover trajectories across a spectrum of clinical outcomes.

What this adds to what is known?

- None of the studies in this review provided account of their data retrieval process, presumably due to the intricacies inherent in working with routinely collected data.
- This review underscores the pressing need for heightened precision and transparency in the reporting of data management procedures and pivotal methodological choices integral to the SA process.

What is the implication and what should change now?

- The development of comprehensive guideline for transparently documenting the technical intricacies associated with analytical process along with the creation of a critical appraisal tool tailored to assess the quality of studies, stands to greatly benefit researchers in this field.

and enhancing patient safety and efficiency in healthcare [2]. Since the 1990s, the adoption of CPs has increased in Europe, starting in the United Kingdom (UK), and gradually embraced by various European countries [2]. The European Pathways Association (EPA), established in 2004, supports the development, implementation, and evaluation of CPs throughout Europe. As per the EPA's mission, CPs aim to elevate the quality of care by improving patient outcomes, promoting patient safety, increasing patient satisfaction, and optimising resource utilisation [3].

Typically, CPs are developed collaboratively by healthcare teams and are grounded in evidence-based guidelines and practices [2]. Nevertheless, the universal application of this approach across all clinical contexts presents challenges due to healthcare professionals' time and resource constraints and the continuous evolution of healthcare processes [4]. To address these challenges, data-driven techniques can support empirically identifying effective care patterns by analysing routinely collected electronic health records (EHRs) [5]. Among these techniques, process mining, machine learning and latent class models have found their places in scrutinising CPs and care patterns from EHR data [6–8]. Nevertheless, they often fall short in the intricate temporal sequencing and event patterns in the patient's healthcare journey. Enter Sequence Analysis (SA), an emerging epidemiological approach derived from social sciences, designed to delve into ordered sequences of healthcare events and classify individuals into distinct groups exhibiting similar care patterns [5]. Process mining focuses on mapping and optimising patient flow within the healthcare system. Latent class analysis categorises patients into distinct groups based on shared healthcare characteristics. In contrast, machine learning is utilised for predicting patient outcomes, diagnosing diseases, and providing personalised treatment recommendations using historical patient data. SA, our focus, uniquely addresses the temporal dynamics and sequential patterns in patient care pathways. SA promises insights into treatment patterns and their effectiveness, which, in turn, can inform the redesign and optimisation of existing care pathways [9].

Abbott and Tsay (2000) outline the fundamental steps of SA, which encompass selecting a suitable alphabet to represent the states and time interval, determining dissimilarity or distance measures, and clustering sequences based on calculated dissimilarity [10]. In the initial step, the states are defined, guided by the research question. Once the states are established, decisions resolve around the time interval and observation period [11]. Subsequently, sequences are compared in pairs, and dissimilarity estimation comes into play. SA employs two techniques to measure sequence dissimilarity. The first method involves non-aligning methods, counting common attributes like longest common subsequences (LCS) or distinct common subsequences. The second method relies on optimal matching (OM), quantifying the degree of dissimilarity between pairs of sequences by considering elementary operations such as insertions, deletions, and substitutions, along with their associated minimum costs required to transform one sequence into the another [12]. Based on the computed dissimilarity measures, clustering techniques classify the sequences into groups with similar patterns. One such clustering method is agglomerative hierarchical clustering, which progressively groups cases into larger clusters by a linkage criterion, often utilising Ward's linkage criterion to ensure within-cluster homogeneity [12]. Another approach is partitioned around the medoid algorithm, which initially assigns cases to k clusters randomly. Through an iterative procedure, it allocates the cases to clusters based on their dissimilarity to the medoid, the reference case with minimal average dissimilarity within each cluster [12].

Liao and colleagues further enrich the SA methodology by discussing advancements in various facets, including sequence visualisation techniques, sequence complexity measures, dissimilarity indices, group and cluster analyses of dissimilarities, and extensions of SA to account for the complexity of sequences [13]. Over the past decade, methodological improvements and the availability of statistical software packages have significantly broadened SA's application, transcending various fields including healthcare research [13]. Consequently, the utilisation of SA in healthcare research has been on

a steady ascent, offering a robust tool for dissecting intricate longitudinal data.

There was a previous scoping review on the application of SA in health services research, which identified 13 studies published before 2019 [14]. However, most of the included studies used data from well-defined settings with limited sample size and may not represent the 'real world' scenario. Furthermore, none of the reviewed studies employed advanced multichannel SA, presumably because the utilization of SA methods in healthcare was in its infancy.

Given the developments, a comprehensive review of the medical literature concerning the application of SA in exploring care trajectories is warranted for several reasons. Firstly, an updated review is essential to focus on recent studies leveraging extensive and diverse datasets, such as EHRs and administrative databases, to effectively capture complex care trajectories within real-world healthcare settings. Secondly, the field of SA has witnessed notable advancements in recent years, and the revised review can shed light on these innovations and their potential implications for health services research, providing researchers with up-to-date insights into SA techniques. Lastly, the revised review addresses a crucial gap identified in the previous review: the need of more, transparent explanations of key methodological decisions within the context of SA. Thus, the primary objective of this review is to update the findings of the previous scoping review and systematically evaluate the SA methodological applicability in elucidating clinical pathways using real-world, routinely collected data.

2. Methods

2.1. Study design

We conducted a methodological systematic review adhering to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guideline [15]. The protocol was registered in the PROSPERO international prospective register of systematic reviews (registration number: CRD42023420532).

2.2. Search strategy and information sources

The search strategy was employed in five medical and health databases: MEDLINE, PsycINFO, CINAHL, EMBASE and Web of Science. The databases were searched from their inception until February 28, 2023. The search strategy consisted of two specific groups of search terms focusing on sequence analysis and clinical pathways. The complete search strategy, along with the number of records retrieved, are presented in Appendix A. The citation and reference lists of the included papers were also reviewed, and potentially relevant publications were manually searched and included in the review.

2.3. Eligibility criteria

Our inclusion criteria encompassed real-world longitudinal observational studies employing SA to identify healthcare trajectories of individuals across all age groups and for any medical condition. We excluded studies utilising SA for purposes other than examining care pathways (e.g., health states, natural history of disease, biological, psychological, or social characteristics, events, processes, and outcomes). We also omitted publications that did not constitute original research articles, and non-English language articles that could not be translated within the timeframe of this review.

2.4. Study selection

All retrieved publications were imported into the Rayyan web application (https://www.rayyan.ai/) where they were de-duplicated and screened. The de-duplication process was executed manually. The main reviewer (SM) was responsible for screening all articles for eligibility, which involved scrutiny of title/abstracts followed by a full-text review. A second reviewer (BSS) independently reviewed a random 10% subset of the search results during both the title/abstract screening and full text screening phases. Any discrepancies arising from this process were subject to discussion and resolution among the remaining review team members (DY, GP and EP).

2.5. Data extraction

SM undertook data extraction using an Excel spreadsheet form. The form underwent piloting and refinement in consultation with team members. The information extracted from eligible research articles included the following details: author(s), year of publication, country of origin, the aim or objective of the study, data source used to obtain the relevant data, study population, sample size, disease/condition of interest, how healthcare events or states in sequence defined and measured, index date or starting point of the study period, duration of the observation, intervals and timeline of sequences, method employed for calculating the dissimilarity matrix and cost settings (if applicable), clustering algorithm utilised for grouping or categorising the data, decision made regarding the optimal number of clusters identified and methodologies used for identifying predictors of clusters.

2.6. Synthesis of the results

The review provided a descriptive summary of relevant features of eligible research studies, with a primary focus on the employed SA methodologies. Separate summaries were provided to describe the characteristics of study population, methodologies of SA used, statistical methods used to identify significant covariates among clusters, and

strategies employed for recognising and managing data quality issues.

2.7. Critical appraisal tool

As the studies in the review were longitudinal quantitative observational studies, the existing tools for conventional study designs, such as cohort study, case-control study, cross-sectional study or randomised controlled trial, could not be used to appraise the studies in the review critically. Therefore, we adapted items from the Real-World Observational Studies (ArRoWS) critical appraisal tool [16] and The REporting of studies Conducted using Observational Routinely collected health Data (RECORD) Statement [17] to suit the specific aspects of SA that were pertinent to the review (see Appendix B).

3. Results

3.1. Study selection and critical appraisal

The PRISMA flow diagram summarises the entire review process (Fig. 1). A total of 6,806 unique records underwent screening, ultimately resulting in the inclusion of 19 studies. Detailed information regarding the distribution of each item in critical appraisal tool, along with corresponding percentages, are presented in Table 1.

3.2. Characteristics of eligible studies

A noteworthy majority ($n = 14$) of the studies included in this review were published after the previous scoping review. These studies predominantly originated from Canada ($n = 7$) and France ($n = 5$) (Table 2). Various data sources were utilised across these studies, with sample size ranging from approximately 800−64,000 participants. Notably, none of the studies provided an explicit account of their data retrieval process, likely due to the inherent complexity of working with routinely collected data.

As detailed in Table 3, the SA method was applied to explore diverse clinical scenarios. Twelve (63%) studies focused on care trajectories post-diagnosis of diseases, while 2 (11%) studies delved into trajectories preceding a clinical condition or a medical procedure. Additionally, 3 (16%) studies examined hospitalisation patterns over a 3-year period, one study scrutinised care trajectories for patients admitted with infections caused by antibiotic-resistant bacteria and another study probed into care trajectories in prenatal care consumption.

3.3. Methods used in SA

A significant majority ($n = 16$) of the studies opted for unidimensional SA, including three studies that combined multiple dimensions of care into one dimension and one study that performed unidimensional SA on multiple dimensions independently (Table 3). Further, three studies chose multichannel SA to account for the multidimensional aspects of care trajectories.

The definition of time intervals exhibited considerable variability across studies, with daily intervals ($n = 6$) representing the shortest duration and yearly intervals ($n = 1$) signifying the widest timespan. Consequently, the sequence length of care trajectories ranged from 6 to 1095 depending on the time interval and study duration.

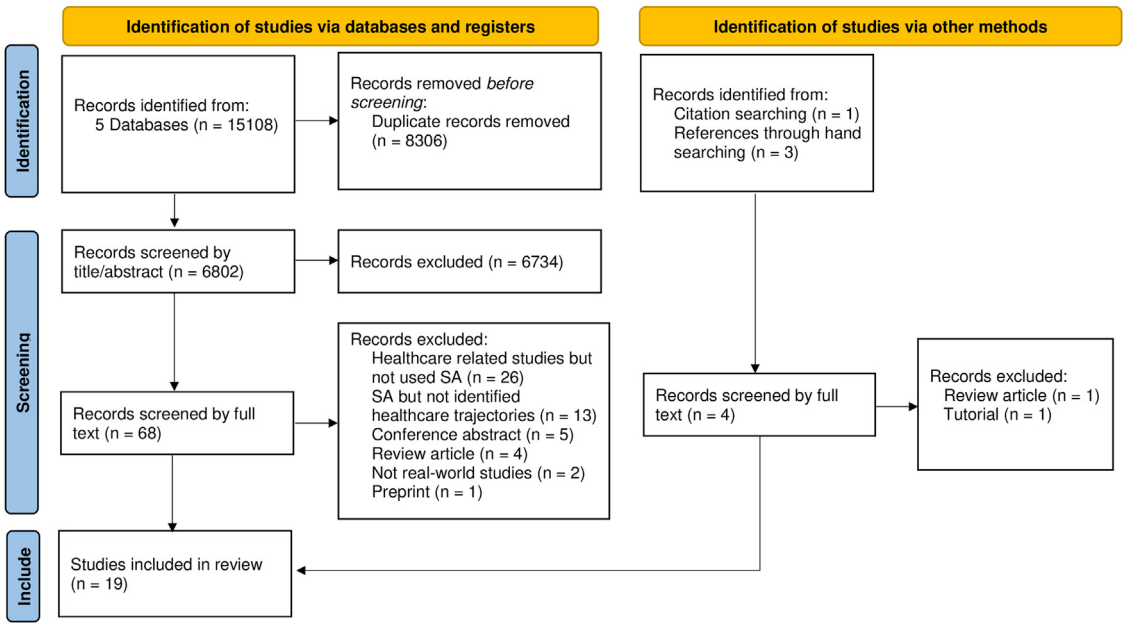

Fig. 1. PRISMA flow diagram. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 1. Summary of critical appraisal of included studies (n = 19)

Sl. No.	Items	Yes (%)	Not reported (%)	Not applicable (%)
1	Is the research question or objective(s) clear?	19 (100)	0	0
2	Is the study population clearly and fully described?	19 (100)	0	0
3	Are the three typical steps of SA described?			
	3.1) The selection of a suitable alphabet for the states and time unit	18 (95)	1 (5)	0
	3.2) The choice of a suitable:			
	Dissimilarity measure	17 (90)	0	2 (11)
	Cost setting	9 (47)	5 (26)	5 (26)
	3.3) Building of typology of trajectories	17 (90)	0	2 (11)
4	Are the statistical analyses clearly defined and appropriate?	19 (100)	0	0
5	Have the authors described handling of missing data in SA?	15 (79)	4 (21)	0
6	Have the authors described sensitivity analysis?	4 (21)	12 (63)	2 (11)
7	Are the limitations of the study described?	19 (100)	0	0
8	Have the authors drawn appropriate conclusions from their results?	19 (100)	0	0

Among the studies, two used SA solely as a descriptive tool for visualising the sequence of events. Within the remaining studies, Optimal Matching (OM) emerges as the predominant dissimilarity measure (n = 12). However, only seven of them specified the cost settings for the three basic operations (deletion, insertion, and substitution) within OM. These cost settings typically featured an indel cost of one and a substitution cost matrix predicated on estimated transition rates (Table 4). A majority of the studies (n = 10) performed agglomerative hierarchical clustering using Ward's linkage as the preferred criterion for ensuring within-cluster homogeneity. The selection of the optimum number of clusters varied across studies, highlighting the diversity in approaches. Importantly, most studies provided insufficient detail regarding the key methodological decisions associated with SA.

Beyond SA, most of the studies (n = 16) carried out parametric or non-parametric statistical tests to compare the patient characteristics among cluster memberships. Seven studies investigated the cluster membership as a predictor for various outcomes, including mortality, self-perceived general health, and pain inference.

This review found that 79% (n = 15) of the studies introduced an additional state to account for instances where no care events were recorded. Given the inherent subjectivity involved in the SA methodology, including choices related to time interval, dissimilarity metrics, cost setting, and clustering algorithm, only four studies conducted multiple alternative analyses with varying selections to evaluate the robustness of their findings.

Table 2. Characteristics of included studies (n = 19)

Characteristics	*n* (%)
Year of publication[a]	
Up to 2018	3 (16)
2019–2020	5 (26)
2021–2023	11 (58)
Country	
Canada	7 (37)
France	5 (26)
Germany	2 (11)
China	2 (11)
Switzerland	1 (5)
Finland	1 (5)
UK	1 (5)
Data source	
Health insurance data	7 (37)
National health data system/registries	4 (21)
Linked health administrative and community health survey database	3 (16)
Hospital medical records	3 (16)
Linked health administrative and clinical data	1 (5)
Electronic health records	1 (5)

[a] Three studies until 2018 and two studies from 2019 were included in a previous scoping review.

Table 3. Application and type of sequence analysis (n = 19)

Methods	*n* (%)
Application of SA in healthcare research	
Care pathways of individuals after the diagnosis of:	
Chronic obstructive pulmonary disease [18,19]	2 (11)
Multiple sclerosis [20,21]	2 (11)
Diabetes [22,23]	2 (11)
Heart failure [24,25]	2 (11)
Arthritis conditions [26]	1 (5)
End-stage renal disease [27]	1 (5)
Prostate cancer [28]	1 (5)
Schizophrenia [29]	1 (5)
Care pathways of individuals prior to:	
Diagnosis of schizophrenia [30]	1 (5)
Invasive coronary angiography [31]	1 (5)
Hospitalisation pattern following:	
Diagnosis of alcohol use disorders [32]	1 (5)
Diagnosis of major psychiatric disorders [33]	1 (5)
Deinstitutionalisation [34]	1 (5)
Care trajectories for patients admitted with infections caused by antibiotic-resistant bacteria [35]	1 (5)
Care trajectories in prenatal care consumption [36]	1 (5)
Dimensions in SA	
Multidimensional [18,21,29]	3 (16)
Combined dimensions [19,26,31]	3 (16)
Unidimensional [20,22–25,27–29,32–36]	13 (68)
Time interval: *n* (sequence length range)	
Days [23,29,32–35]	6 (360–1095)
Weekly [22,30,37]	3 (52–104)
Monthly [19,26,27,36]	4 (9–48)
Quarterly [24,31]	2 (6–8)
Half-yearly [21,28]	2 (7–26)
Yearly [20]	1 (7)
Not clear [25]	1

SA indicates sequence analysis.

4. Discussion

This systematic review marks the first comprehensive assessment of the methodological application of SA in identifying real-world CPs from routinely collected data. Several important findings emerge from this investigation. Firstly, there has been a notable surge in the utilisation of SA for comprehending care pathways in recent years. In contrast to a previous scoping review, which included articles up to 2019, that identified only five studies employing SA to explore care pathways utilising routinely collected data [20,24,25,27,36], there has been a substantial increase in the adoption of this methodology for understanding care trajectories associated with diverse medical conditions. Secondly, it is evident that studies included in the review continue to exhibit deficiencies in reporting key methodological decisions associated with SA, mirroring the findings of the previous scoping review. Notably, the subjective nature of choices related to SA parameters, including the definition of states, time intervals and observation period, and dissimilarity measures, exerts a significant influence on the resulting outcomes and insights derived from the analysis [22]. Aligning these parameter choices with the research objective is crucial [11,13]. A select few studies resorted to defining states of healthcare utilisation closely reflecting the "6W" model proposed by Vanasse and colleagues [18]. This model, which elucidates care trajectories by considering successive interactions in time ("when") between patients ("who") and healthcare services, specifically the healthcare settings ("where"), the reason for consultation ("why"), the care providers involved ("which"), and the sequence of services received ("what") over a specific period of time. This model allowed for the observation of multiple trajectories for each individual, and the goal of multichannel SA was to investigate how these trajectories unfold jointly [38]. Consequently, insights derived from cluster patterns generated through multichannel SA prove invaluable for informing patient care, management, and health services planning.

The granularity of the analysis is heavily influenced by the choice of time intervals [12]. Smaller intervals, such as daily, yield more data points, heightening sensitivity to changes and events. Conversely, broader time intervals, such as monthly, necessitate the establishment of hierarchies to account for the potential occurrence of multiple care events within the chosen time frame [22]. In such cases, events with higher priority are typically selected, prioritising hospital and emergency visits over ambulatory care visits, and specialists over primary care providers [37]. Thus, the researchers must judiciously select time intervals based on their interests in the presence of each state or the overall pattern of care rather than transitions between care states. Moreover, the selection of an observation period and time intervals crucially determines the length of the sequences. Dlouhy and Biemann's simulation study on career sequences, recommends a minimum sequence length of 25 ensuring high-quality outcomes [39]. The maximal distance between a pair of sequences depends on their length. Normalization becomes necessary to account for these differences when dealing with sequences of different lengths. Normalization seeks to set the maximal distance to 1, or a value independent of sequence lengths [40].

The choice of dissimilarity measure significantly impacts the results of SA. While OM remains the most prevalent dissimilarity measure, various alternative methods are available [30]. Studer and Ritschard conducted a comprehensive and critical review of these dissimilarity measures,

Table 4. Clustering techniques and decision on optimum number of clusters (n = 17)

Methods	*n* (%)
Dissimilarity measures and cost-setting	
Optimal matching with indel cost of one and substitution cost matrix of estimated transition rates	7 (41)
Optimal matching (not specified cost setting)	5 (29)
Localised optimal matching with substitution cost matrix based on Gower distance method	1 (6)
Simple Hamming distance	1 (6)
Dynamic Hamming distance with substitution cost matrix based on time varying state transitions	1 (6)
Longest common subsequence	2 (12)
Clustering techniques	
Agglomerative hierarchical clustering using Ward's linkage	10 (59)
Hierarchical clustering (not specified linkage criterion)	1 (6)
Cluster analysis (not specified which method)	2 (12)
Partitioned around medoid	3 (18)
Regression tree analysis	1 (6)
Criteria on optimum no. of clusters	
Dendrogram	6 (35)
Inertia curve	6 (35)
Interpretability and clinical meaning	8 (47)
Average Silhouette width	6 (35)
Hubert's C	2 (12)
Minimum sample size in cluster	2 (12)
Not mentioned	1 (6)

providing researchers with guidance for selecting an appropriate measure aligned with their research objective [41]. The application of agglomerative hierarchical clustering based on Ward's linkage, which minimises within-cluster variance [12]; was predominant among the reviewed studies. However, three studies opted for the partitioned around the medoid algorithm [24,28,31], which necessitates the initial specification of the number of clusters [12]. Different initial partitions can yield distinct final clusters, highlighting the importance of selecting the optimum number of clusters based on quality indices, statistical criteria and, moreover, clinical meaning and interpretation of the clusters [42].

Finally, defining medical events in SA using real-world data present inherent challenges due to the vast information and complexity in sources like EHRs and administrative databases. Researchers must clearly define the data requirements and efficiently extract and organise the data for SA. Additionally, real-world data may contain missing or incomplete information, posing challenges in accurately identifying sequences of medical events. Common approaches to handling missing data include either excluding incomplete sequences or treating the missing state as another state in the alphabet [13]. While most reviewed studies where no care events were recorded, they did not provide explicit information on how they are addressing missing data issues.

The distinctive feature of this systematic review lies in its focus on the methodological issues related to SA and evaluation of various SA methods for exploring real-world CPs. However, a limitation arises from the challenge of adapting appraisal tools designed for longitudinal observational studies to the context of SA. To mitigate this limitation, we developed a customised tool for assessing real-world observational studies incorporating pertinent aspects of SA.

This systematic review shed light on the methodological practical issues of SA within the context of clinical pathway analysis. Key initial steps include defining medical events or states within the sequences, such as diagnoses, treatments, procedures, or healthcare interactions, pertinent to the research question. Decisions on the time unit for measuring events (day, weeks, months, or years) and the observation period are crucial. The choice between methods like LCS or OM for sequence comparison is central. LCS, a non-aligning method, identifies the most common attributes in the same order across sequences. In contrast, OM, including Hamming distance and localized OM, calculates dissimilarity by assessing the cost to transform one sequence into another, factoring in insertion, deletion, and substitutions. Clustering techniques, such as agglomerative hierarchical clustering and the partitioned-around medoid algorithm, are employed to classify sequences based on dissimilarity measures. The quality of clusters is evaluated using indices like Average Silhouette Width (ASW) and Hubert's C (HC) index, with further consideration for cluster size, clinical relevance, and visualisation tools, and predefined rationale for choices.

This review offers insights for researchers embarking on studies involving specific conditions or diseases. It provides guidance on various dimensions of care to consider, defining states in sequence, selecting appropriate time intervals and observation period, and making informed choices regarding statistical methodologies for SA. Given the depth and implication of the current study, we will consider the development of a dedicated review paper to outline the main procedures, different choices of statistical methods, and practical considerations as a future research endeavor. Such a review paper has the potential to offer essential guidelines and insights for subsequent sequence analyses within the context of CPs. Care trajectories based on SA can provide personalised treatment plans for patients, leading to more effective and patient-centered care. This, in turn, facilitates efficient healthcare resource allocation, directing resources toward patients likely to require intensive or long-term care. Policymakers can utilize this

information to design and implement policies that enhance disease prevention, early detection, and management.

5. Conclusion

This review underscores the imperative need for more precise and transparent reporting concerning data management and the key methodological decisions underpinning the SA process. The development of guidelines for the clear reporting of technical details related to the analytical process and the creation of a critical appraisal tool tailored to assess the quality of studies using SA would greatly benefit the research community.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2023.111226.

CRediT authorship contribution statement

Smitha Mathew: Conceptualization, Data curation, Formal analysis, Writing − original draft, Writing − review & editing. **George Peat:** Conceptualization, Writing − review & editing, Supervision. **Emma Parry:** Conceptualization, Writing − review & editing, Supervision. **Balamrit Singh Sokhal:** Data curation, Writing − review & editing. **Dahai Yu:** Conceptualization, Writing − review & editing, Supervision.

Data availability

No data was used for the research described in the article.

Declaration of competing interest

The authors declare no conflict of interest.

Acknowledgments

The authors wish to extend their gratitude to Nadia Corp and the other members of the systematic review team at Keele University for their invaluable support, especially during the protocol development and data retrieval from the EMBASE database. SM was supported by PhD studentship from the Faculty of Medicine and Health Sciences at Keele University.

References

[1] Kinsman L, Rotter T, James E, Snow P, Willis J. What is a clinical pathway? Development of a definition to inform the debate. BMC Med 2010;8:31.
[2] Rotter T, de Jong RB, Lacko SE. Clinical pathways as a quality strategy. In: Busse R, Klazinga N, Panteli D, Quentin W, editors. Improving healthcare quality in Europe. Denmark: European Observatory on Health Systems and Policies; 2019. vol. Health Policy Series. No. 53.
[3] European Pathways Association. Available at https://e-p-a.org/. Accessed August 14, 2023.
[4] Cho M, Kim K, Lim J, Baek H, Kim S, Hwang H, et al. Developing data-driven clinical pathways using electronic health records: the cases of total laparoscopic hysterectomy and rotator cuff tears. Int J Med Inform 2020;133:104015.
[5] Savaré L, Ieva F, Corrao G, Lora A. Mining and evaluation of patients' diagnostic therapeutic paths through state sequences analysis. arXiv 2022. https://doi.org/10.48550/arXiv.2209.04384.
[6] Duma D, Aringhieri R. An ad hoc process mining approach to discover patient paths of an emergency department. Flex Serv Manuf J 2020;32:6−34.
[7] Kempa-Liehr AW, Lin CY-C, Britten R, Armstrong D, Wallace J, Mordaunt D, et al. Healthcare pathway discovery and probabilistic machine learning. Int J Med Inform 2020;137:104087.
[8] Hastings SN, Whitson HE, Sloane R, Landerman LR, Horney C, Johnson KS. Using the past to predict the future: latent class analysis of patterns of health service use of older adults in the emergency department. J Am Geriatr Soc 2014;62:711−5.
[9] Huang Z, Lu X, Duan H. On mining clinical pathway patterns from medical behaviors. Artif Intell Med 2012;56:35−50.
[10] Abbott A, Tsay A. Sequence analysis and optimal matching methods in sociology. Sociol Methods Res 2000;29:3−33.
[11] Blanchard P. Sequence analysis. In: Atkinson PA, Williams RA, Cernat A, editors. Encyclopedia of research methods. London: Sage; 2019:1−22.
[12] Raab M, Struffolino EIn: Sequence analysis, vol. 190. Thousand Oaks, CA: SAGE Publications, Inc; 2022.
[13] Liao TF, Bolano D, Brzinsky-Fay C, Cornwell B, Fasang AE, Helske S, et al. Sequence analysis: its past, present, and future. Soc Sci Res 2022;107:102772.
[14] Liao W. The use of sequence analysis to study primary care pathways: an exploratory study of people at high risk of lung cancer in England. Southampton: University of Southampton; 2022: Doctoral Thesis.
[15] Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009;6:e1000097.
[16] Coles B, Tyrer F, Hussein H, Dhalwani N, Khunti K. Development, content validation, and reliability of the assessment of real-world observational Studies (ArRoWS) critical appraisal tool. Ann Epidemiol 2021;55:57−63.e15.
[17] Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, et al. The REporting of studies conducted using observational routinely-collected health data (RECORD) statement. PLoS Med 2015;12:e1001885.
[18] Vanasse A, Courteau M, Ethier JF. The "6W" multidimensional model of care trajectories for patients with chronic ambulatory care sensitive conditions and hospital readmissions. Public Health 2018;157:53−61.
[19] Henri S, Herrera R, Vanasse A, Forget A, Blais L. Trajectories of care in patients with chronic obstructive pulmonary disease: a sequence analysis. Can J Respir Crit 2022;6:237−47.
[20] Roux J, Grimaud O, Leray E. Use of state sequence analysis for care pathway analysis: the example of multiple sclerosis. Stat Methods Med Res 2019;28:1651−63.
[21] Roux J, Kingwell E, Zhu F, Tremlett H, Leray E. Care consumption of people with multiple sclerosis: a multichannel sequence analysis in

a population-based setting in British Columbia, Canada. Multiple Sclerosis. Journal 2022;28:309−22.
[22] Kurkela O, Nevalainen J, Arffman M, Lahtela J, Forma L. Foot-related diabetes complications: care pathways, patient profiles and costs. BMC Health Serv Res 2022;22:559.
[23] McKay R, Letarte L, Lebel A, Quesnel-Vallée A, Vanasse A, Bartlett G, et al. Exploring social inequalities in healthcare trajectories following diagnosis of diabetes: a state sequence analysis of linked survey and administrative data. BMC Health Serv Res 2022;22:131.
[24] Vogt V, Scholz SM, Sundmacher L. Applying sequence clustering techniques to explore practice-based ambulatory care pathways in insurance claims data. Eur J Public Health 2018;28:214−9.
[25] Rao A, Kim D, Darzi A, Majeed A, Aylin P, Bottle A. Long-term trends of use of health service among heart failure patients. Eur Heart J Qual Care Clin Outcomes 2018;4:220−31.
[26] Nguena Nguefack HL, Pagé MG, Choinière M, Vanasse A, Deslauriers S, Angarita-Fonseca A, et al. Distinct care trajectories among persons living with arthritic conditions: a two-year state sequence analysis. Front Pain Res 2022;3:1014793.
[27] Le Meur N, Vigneau C, Lefort M, Lebbah S, Jais J-P, Daugas E, et al. Categorical state sequence analysis and regression tree to identify determinants of care trajectory in chronic disease: example of end-stage renal disease. Stat Methods Med Res 2019;28:1731−40.
[28] Baulain R, Jové J, Sakr D, Gross-Goupil M, Rouyer M, Puel M, et al. Clustering of prostate cancer healthcare pathways in the French National Healthcare database. Cancer Innovat 2023;2:52−64.
[29] Brodeur S, Vanasse A, Courteau J, Courteau M, Stip E, Fleury M, et al. Antipsychotic utilization trajectories three years after initiating or reinitiating treatment of schizophrenia: a state sequence analysis approach. Acta Psychiatr Scand 2022;145:469−80.
[30] Vanasse A, Courteau J, Courteau M, Roy M-A, Stip E, Fleury M-J, et al. Multidimensional analysis of adult patients' care trajectories before a first diagnosis of schizophrenia. Schizophrenia 2022;8:52.
[31] Novelli A, Frank-Tewaag J, Bleek J, Günster C, Schneider U, Marschall U, et al. Identifying and investigating ambulatory care sequences before invasive coronary angiography. Med Care 2022;60:602−9.
[32] Han X, Jiang F, Zhou H, Needleman J, Guo M, Chen Y, et al. Hospitalization pattern, inpatient service utilization and quality of care in patients with alcohol use disorder: a sequence analysis of discharge medical records. Alcohol Alcohol 2020;55:179−86.
[33] Han X, Jiang F, Needleman J, Guo M, Chen Y, Zhou H, et al. A sequence analysis of hospitalization patterns and service utilization in patients with major psychiatric disorders in China. BMC Psychiatry 2021;21:245.
[34] Golay P, Morandi S, Conus P, Bonsack C. Identifying patterns in psychiatric hospital stays with statistical methods: towards a typology of post-deinstitutionalization hospitalization trajectories. Soc Psychiatry Psychiatr Epidemiol 2019;54:1411−7.
[35] Touat M, Brun-Buisson C, Opatowski M, Salomon J, Guillemot D, Tuppin P, et al. Costs and Outcomes of 1-year post-discharge care trajectories of patients admitted with infection due to antibiotic-resistant bacteria. J Infect 2021;82:339−45.
[36] Le Meur N, Gao F, Bayat S. Mining care trajectories using health administrative information systems: the use of state sequence analysis to assess disparities in prenatal care consumption. BMC Health Serv Res 2015;15:200.
[37] Vanasse A, Courteau J, Courteau M, Benigeri M, Chiu YM, Dufour I, et al. Healthcare utilization after a first hospitalization for COPD: a new approach of State Sequence Analysis based on the ''6W'' multidimensional model of care trajectories. BMC Health Serv Res 2020;20:177.
[38] Gauthier J-A, Widmer ED, Bucher P, Notredame C. Multichannel sequence analysis applied to social science data. Sociol Methodol 2010;40:1−38.
[39] Dlouhy K, Biemann T. Optimal matching analysis in career research: a review and some best-practice recommendations. J Vocat Behav 2015;90:163−73.
[40] Gabadinho A, Ritschard G, Müller NS, Studer M. Analyzing and visualizing state sequences in R with TraMineR. J Stat Softw 2011;40:1−37. https://doi.org/10.18637/jss.v040.i04.
[41] Studer M, Ritschard G. What matters in differences between life trajectories: a comparative review of sequence dissimilarity measures. J Roy Stat Soc 2016;179:481−511.
[42] Han Y, Liefbroer A, Elzinga C. Comparing methods of classifying life courses: sequence analysis and latent class analysis. Longit Life Course Stud 2017;8:319−41. https://doi.org/10.14301/llcs.v8i4.409.