Systematic literature review on disease trajectory modelling and prediction and data science techniques

Hamza Qureshi[1,2*], Aqeel Malik[2,3] and Raiz [1,2†]

[†]These authors contributed equally to this work.

Abstract

The impact of researchers' strategies for publishing or producing scientific content can be profound and enduring. Identifying the most influential strategies for future impact has garnered increasing attention in academic literature. This study conducts a systematic review of recommendations concerning long-term strategies in research analytics and their implementation methodologies. Our objective is to provide a comprehensive overview from 2002 to 2018 of the evolution of this field, encompassing trends and contextual considerations. A primary focus is on identifying data-driven approaches for understanding long-term research strategies, particularly through process mining. Our approach adheres to a structured and rigorous protocol for systematic reviews in engineering research. The findings underscore the necessity for studies that offer specific recommendations grounded in data mining. This underscores the need for further research, particularly in two key areas: the application of process mining methodologies in research analytics and the exploration of data science techniques for assessing the feasibility of long-term strategies.

Keywords: modelling; data science; long-term strategies: trajectory

1 Introduction

Sánchez-Torres and Miles discussed tools for systematically assessing challenges and opportunities using future-oriented technology analysis in the context of e-government policy development. However, existing reviews primarily focus on contexts such as medicine, lacking an emphasis on data science as a strategic tool but rather integrating strategies into empirical conclusions. This study aims to address this gap by integrating scientometrics or research analytics to identify and contextualize case studies

combining methodologies and tools comprehensively. As far as we know, no systematic review has yet consolidated studies focusing on long-term strategies within this context.

The objective of this systematic review is to identify articles presenting long-term strategies utilizing data science, specifically process mining, within research analytics. We aim to explore the methodologies employed in the development and evaluation of these strategies. Our interest lies in understanding how these strategies are conceptualized in literature and their long-term effects.

Gómez et al. argue that systematic reviews provide a robust tool for identifying, evaluating, and interpreting studies within defined contexts. They emphasize the synthesis of information with rigor and impartiality to ensure scientific value. This approach is pivotal for uncovering insights that might otherwise be overlooked.

2 Methodolgy

In selecting sources for this study, several criteria were defined. Publications were sourced from websites hosting search engines utilizing specific keywords and recommendations from experts. All selected studies were required to be in English. The search strategy employed various combinations of keywords, including process mining, research analytics, long-term strategies, scientometrics, long-term learning, model, scientific career, and data mining. Detailed combinations are listed in Table 1. The primary source utilized was Scopus. Following an evaluation based on predefined quality criteria, sources meeting these standards were selected. Additionally, each selected reference underwent thorough scrutiny and approval by two researchers from the Instituto Tecnológico y de Estudios Superiores de Monterrey, ensuring consensus and adherence to selection criteria.

Table 1 Inclusion and exclusion criteria.

Criteria	Description
CI1	Includes publications whose titles contain disease trajectory modelling and prediction.
CI2	Include publications that contain keywords that match the selected keywords.
CI3	Includes publications where the abstract contains selected keywords.
CI4	Includes publications that are available in full-text.
CE1	Excludes publications that do not meet the inclusion criteria described above.
CE2	Excludes any duplicate publication.

Each article underwent classification using the 2012 ACM Computing Classification System [32]. The process involved several steps: firstly, keywords were employed to identify specific categories associated with each article. Secondly, titles were scrutinized for relevant keywords to facilitate category assignment. Lastly, abstracts were carefully reviewed to confirm and finalize category placements. Figure 1 illustrates the flow diagram detailing the inclusion criteria process.

For taxonomy specificity, classifications extended up to four levels: Level 1 denoted the category, Level 2 the domain, and Level 3 the subdomain. In this study, focus was

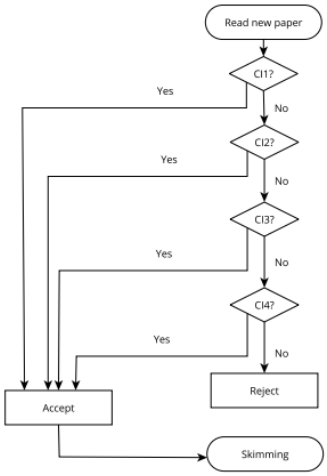

Fig. 1 Selection Criteria Explained in Block Diagram

Table 2 Results

Process Mining Technique	Papers
Conformance	1
Discovery	15
Discovery, Conformance checking	7
Event log aggregation	1
No	5

primarily on the first three levels of the taxonomy. Articles classified within any of these levels were considered not applicable for subsequent levels. Table 4 summarizes the results categorized across three levels: Level 1 (Category), Level 2 (Domain), and Level 3 (Subdomain), with categories in bold, domains underlined, and subdomains in italics.

This systematic classification underscores the study's exploration of relationships between scientometrics, data mining, long-term strategies, and the pivotal role of process mining within data mining practices.

Recent studies in disease trajectory modelling (2021-2024) have extensively utilized AI and Data Science/Data Mining methodologies to predict disease progression and outcomes. Machine learning techniques, including supervised learning (such as logistic regression and random forests), unsupervised learning (like clustering algorithms), and reinforcement learning, have been prominent. Deep learning methods, particularly neural networks and convolutional neural networks, have also gained traction for their ability to handle complex data structures and extract intricate patterns inherent in disease trajectories. These methodologies enable researchers to leverage large-scale datasets effectively, providing insights into disease progression dynamics and individualized treatment responses.

The strengths of AI and Data Science/Data Mining methodologies lie in their capacity to analyze vast amounts of heterogeneous data, thereby capturing subtle relationships and temporal patterns in disease trajectories. Machine learning techniques offer high predictive accuracy and robust performance in handling diverse datasets. Deep learning excels in feature extraction from raw data, particularly images and unstructured text, enhancing the understanding of disease progression markers. However, these methodologies face challenges such as model interpretability, especially in complex neural networks, and the need for large annotated datasets for effective training. Computational complexity and the potential for overfitting are additional limitations that researchers must address to ensure the reliability and generalizability of their models.

Predominantly employed evaluation metrics in disease trajectory modelling include predictive accuracy, sensitivity, specificity, and clinical utility. AI and Data Science/-Data Mining models demonstrate strong performance in predicting disease progression and outcomes, often achieving high accuracy and sensitivity in detecting early disease markers or treatment responses. However, challenges persist in achieving high specificity, particularly in distinguishing between disease progression stages or identifying rare events. Clinical utility, which measures the practical impact of model predictions on patient care and treatment decisions, remains a critical but evolving area of evaluation. Addressing these metrics effectively is crucial for translating research findings into clinical practice and improving patient outcomes.

Commonly used open datasets in disease trajectory modelling research between 2021 and 2024 exhibit diverse characteristics tailored to study disease progression and treatment outcomes comprehensively. These datasets typically include longitudinal health records spanning diverse patient demographics, clinical variables, genetic profiles, and lifestyle factors. They are curated to facilitate the training and validation of AI and Data Science/Data Mining models, providing researchers with ample opportunities to explore disease dynamics over time. The availability of such datasets fosters collaboration and benchmarking efforts within the scientific community, promoting transparency and reproducibility in disease trajectory modelling studies.

The suitability of open datasets for training and evaluating AI and Data Science/-Data Mining models in disease trajectory modelling is influenced by their richness in data diversity, volume, and annotation quality. These datasets enable researchers to develop robust models capable of capturing complex disease trajectories and predicting patient outcomes accurately. However, challenges such as data heterogeneity, missing values, and data privacy concerns necessitate careful preprocessing and validation strategies. Establishing standardized protocols for dataset curation and model evaluation enhances the reliability and comparability of research findings across different studies. Collaborative efforts to enhance dataset accessibility and usability are essential for advancing the field of disease trajectory modelling and driving innovations in personalized medicine.

Common challenges in disease trajectory modelling research using AI and Data Science/Data Mining techniques include the interpretability of complex models, scalability to handle large-scale datasets, and the integration of diverse data sources

for comprehensive analysis. Model overfitting, particularly in deep learning architectures, remains a concern, along with the ethical implications of data privacy and security in handling sensitive patient information. Limited access to high-quality annotated datasets and the need for domain-specific expertise in healthcare informatics further constrain research advancements. Addressing these challenges requires interdisciplinary collaborations, robust validation frameworks, and transparent reporting standards to foster trust and credibility in disease trajectory modelling applications.

Emerging research directions in disease trajectory modelling focus on integrating multimodal data sources, including genomic, imaging, and real-time sensor data, to enhance predictive modeling capabilities. Advanced machine learning algorithms, such as transfer learning and federated learning, offer promising avenues to mitigate data privacy concerns while leveraging distributed datasets across healthcare systems. Incorporating explainable AI techniques facilitates model interpretability, enabling clinicians to understand and trust AI-driven predictions in clinical decision-making. Future research efforts should prioritize developing standardized benchmarks and validation protocols, fostering open science initiatives to promote data sharing and collaboration. These initiatives are crucial for advancing the state-of-the-art in disease trajectory modelling and translating research findings into actionable insights for personalized healthcare delivery.