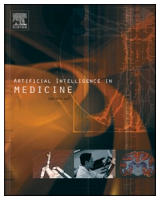
Artificial Intelligence In Medicine 144 (2023) 102645

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed



Process mining and data mining applications in the domain of chronic diseases: A systematic review

Kaile Chen [a,b,*], Farhad Abtahi [a,b,c], Juan-Jesus Carrero [d], Carlos Fernandez-Llatas [a,e], Fernando Seoane [a,c,f,g]

[a] Department of Clinical Science, Intervention and Technology, Karolinska Institutet, 17177 Stockholm, Sweden
[b] School of Engineering Sciences in Chemistry, Biotechnology and Health, Department of Biomedical Engineering and Health Systems, Division of Ergonomics, KTH Royal Institute of Technology, 14157 Stockholm, Sweden
[c] Department of Clinical Physiology, Karolinska University Hospital, 17176 Stockholm, Sweden
[d] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, 17177 Stockholm, Sweden
[e] SABIEN, ITACA, Universitat Politècnica de València, Spain
[f] Department of Medical Technology, Karolinska University Hospital, 17176 Stockholm, Sweden
[g] Department of Textile Technology, University of Borås, 50190 Borås, Sweden

ARTICLE INFO

Keywords:
Chronic disease
Data mining
Process mining
Systematic review

ABSTRACT

The widespread use of information technology in healthcare leads to extensive data collection, which can be utilised to enhance patient care and manage chronic illnesses. Our objective is to summarise previous studies that have used data mining or process mining methods in the context of chronic diseases in order to identify research trends and future opportunities. The review covers articles that pertain to the application of data mining or process mining methods on chronic diseases that were published between 2000 and 2022. Articles were sourced from PubMed, Web of Science, EMBASE, and Google Scholar based on predetermined inclusion and exclusion criteria. A total of 71 articles met the inclusion criteria and were included in the review. Based on the literature review results, we detected a growing trend in the application of data mining methods in diabetes research.

Additionally, a distinct increase in the use of process mining methods to model clinical pathways in cancer research was observed. Frequently, this takes the form of a collaborative integration of process mining, data mining, and traditional statistical methods. In light of this collaborative approach, the meticulous selection of statistical methods based on their underlying assumptions is essential when integrating these traditional methods with process mining and data mining methods. Another notable challenge is the lack of standardised guidelines for reporting process mining studies in the medical field. Furthermore, there is a pressing need to enhance the clinical interpretation of data mining and process mining results.

1. Introduction

Due to the increasing burden of chronic disease, the prevention and management of chronic disease is now a global concern. Currently, about one-third of adults worldwide live with more than one chronic disease [1]. The wide utilisation of information technology in healthcare practice leads to the collection of massive amounts of data, which contains valuable information waiting to be exploited. This “big data” has the potential to inform decisions regarding the health trajectory of patients and the management of chronic diseases. However, determining how to effectively analyse big data to obtain evidence that can be applied in healthcare is still a challenge.

Data mining plays a crucial role in chronic disease research, as it allows researchers to identify patterns and relationships within big data and apply predictive analytics, for example, in early intervention and diagnosis, and the development of models to forecast disease progression. Various data mining techniques have been developed to explore large datasets [2,3], and these techniques have the potential to be a powerful tool in the identification of behaviour models for long-term chronic disease and to guide clinical decisions [4,5]. The main techniques used in this area include classification, regression, and clustering [6]. The data mining methods used in different techniques depend on the characteristics of the data and the specific insights sought. Thus, it is important to review the literature about the application of data mining

* Corresponding author at: Department of Clinical Science, Intervention and Technology, Karolinska Institutet, 17177 Stockholm, Sweden.
E-mail address: kaile.chen@ki.se (K. Chen).

https://doi.org/10.1016/j.artmed.2023.102645
Received 2 March 2023; Received in revised form 24 August 2023; Accepted 28 August 2023
Available online 29 August 2023

Table 1
Search terms for the review.

No.	Database	Terms
Initial identification		
	Web of Science	TS=("process mining" OR "data mining") AND TS=("chronic disease*" OR "chronic condition"*" OR "long* condition*")
	PubMed	("process mining"[Title/Abstract] OR "data mining"[Title/Abstract]) AND ("chronic disease*"[Title/Abstract] OR "chronic condition*"[Title/Abstract] OR " long-term condition*"[Title/Abstract])
	Google Scholar	("process mining" OR "data mining") AND ("chronic disease*" OR "chronic condition*" OR " long-term condition*")
	Embase	('process mining' OR 'data mining'/exp OR 'data mining') AND ('chronic disease*' OR 'chronic condition*' OR 'long-term condition*')
Re-identification for process mining		
	Pubmed	(("process mining"[Title/Abstract]) AND ("chronic disease*"[Title/Abstract] OR "chronic condition*"[Title/Abstract] OR "long-term condition*"[Title/Abstract] OR "diabetes*"[Title/Abstract] OR "kidney*"[Title/Abstract] OR "cardiovascular*"[Title/Abstract] OR "dyslipidemia*"[Title/Abstract] OR "obesity*"[Title/Abstract] OR "cancer*"[Title/Abstract] OR "asthma*"[Title/Abstract] OR "chronic obstructive pulmonary disease*"[Title/Abstract] OR "migraine*"[Title/Abstract] OR "dementia*"[Title/Abstract] OR "multiple sclerosis*"[Title/Abstract] OR "hypertension"[Title/Abstract] OR "musculoskeletal*"[Title/Abstract] OR "periodontitis*"))

methods in the study of chronic diseases in order to identify trends.

Process mining, which builds on process model-driven approaches, has increased in popularity within the medical field in recent years [7]. The first publication that looked at process mining applications in healthcare was published in 2001 [8]. Process mining provides an alternative way to show the potential of constructing individualised behaviour models and has gained considerable traction in the domains of clinical pathways and disease trajectory analysis [9–12]. Moreover, the approach exhibits promising potential in exploring the associations between risk factors and chronic disease [13]. By leveraging real-world data, process mining endeavours to unravel the sequential steps followed by individuals and subsequently transform raw data into comprehensible process indicators.

Previous literature has included reviews of data mining in specific diseases [14] or within the broader healthcare context [6,15]. Similarly, there have been reviews focusing on process mining in healthcare [9,16,17]. However, few publications have combined these two disciplines to examine trends in their development and potential synergistic relationships. Therefore, the objective of our study is to offer a comprehensive summary of the available evidence regarding the application of both data mining and process mining in chronic diseases with a specific emphasis on scientific publications featuring case studies. To enhance clarity and guide our review, we have formulated the following research questions:

1) What are the characteristics of research papers that employ data mining and process mining in the study of chronic diseases?
2) What methods are frequently used in data mining and process mining? Our aim is to shed light on the most frequently used methods in data mining and process mining in the domain of chronic diseases. By identifying these prominent methods, we aim to provide a comprehensive understanding of the prevailing methodologies utilised in this research area.
3) What are the current gaps or research needs in data mining and process mining?

The subsequent sections of this paper are organised as follows. Section 2 presents a comprehensive description of the methodology employed in this systematic review. Section 3 presents a detailed account of the findings and outcomes of this systematic review. Section 4 discusses and highlights the gaps and limitations identified during the review process. Section 5 presents a conclusion by summarising the key findings of the present paper and offering concluding remarks.

Table 2
Criteria for inclusion and exclusion.

Inclusion criteria	Exclusion criteria
IC1: Articles written in English	EC1: Thesis, book section, conference papers, and review studies
IC2: Articles published from 1 January 2000 through 31 December 2022	EC2: Genome/gene studies, image data, wearable device, text mining
IC3: Articles must involve data mining or process mining related to chronic conditions/diseases	EC3: Mining social media platforms
IC4: At least one case study must be included in each article	

2. Methods

This section provides a comprehensive account of the methodology employed in this systematic review. Section 2.1 presents an account of the specific databases and search strategies that were used in this paper. Section 2.2 outlines the criteria employed for selecting and excluding relevant studies. Section 2.3 describes the data collection process, while Section 2.4 presents the methods used to present our summary and descriptive analysis.

2.1. Literature search

The search strategy aims to locate relevant studies about the application of process mining or data mining technologies in the domain of chronic diseases (search terms and strategies are shown in Table 1). Web of Science, PubMed, Embase and Google Scholar were searched. Google Scholar was used as an additional database for grey articles; only the first 200 results were included [18]. The publication date range was from 2000 to 2022.

Following the initial round of article identification, we encountered a limited number of papers related to process mining. To address this, we conducted a subsequent search in PubMed, focusing specifically on process mining in relation to all of the chronic diseases identified during the initial search phase.

2.2. Criteria for selection of studies

The inclusion and exclusion criteria are summarised in Table 2.

2.3. Data extraction

Based on the inclusion criteria, key information was extracted from the selected primary studies as follows: Title of the Study; First Author; Year of Publication; Disease Category; Study Duration (years); Country; Study design; Study Sample Size; Aims; Methods or Models; Results or Conclusions.

2.4. Summary and descriptive analysis methods

2.4.1. Definition of clinical category

The specific disease name in each study was recorded and retrieved using the MeSH database in order to identify the disease category.

2.4.2. Study design

All included studies were classified into longitudinal and non-longitudinal studies. Longitudinal study was defined as a study that uses continuous or recurring measurements to monitor specific

Table 3
Comparison of the current review with previously published reviews.

	General information				Focused review aspects							
	Review type	Disease scope	No. of papers included	Date intervals	Inclusion of other data analytics methods	Clinical disease categories of process mining applications	Dataset's Characteristics of included primary studies	Process mining methods/ algorithms	Research domains/ categories of process mining applications	Process mining tools	Challenges & gaps	Future work
Present study	Systematic review	Chronic disease	71	2000–2022	Yes (included data mining)	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Kurniati, Johnson et al. [21]	Literature review	Oncology	37	2008–2016	No	Oncology only	No	No	Yes	Yes	Yes	Yes
Ghasemi and Amyot [22]	Systematised review	General healthcare	36	No limit	No	Yes	No	Yes	No	Yes	Yes	No
Rojas, Munoz-Gama et al. [23]	Literature review	General healthcare	74	~Feb 2016	No	Yes	No	Yes	Yes	Yes	Yes	Yes
Kusuma, Hall et al. [24]	Literature review	Cardiology	32	Jan1998–Jun2017	No	Cardiology only	No	Yes	Yes	Yes	No	Yes
Batista and Solanas [25]	Systematic review	General healthcare	55	Not mentioned	No	Yes	No	Yes	Yes	Yes	Yes	Yes
Williams, Rojas et al. [26]	Literature review	Primary care	143	Feb2016–Oct2017	No	Yes	Yes	No	No	No	Yes	No
Sundari and Nayak [27]	Critical review	General healthcare	26	Not mentioned	No	No	No	Yes	No	No	No	Yes
Grüger, Bergmann et al. [28]	Systematic review	Oncology	55	Not mentioned	No	Oncology only	Yes	Yes	Yes	Yes	Yes	Yes
Guzzo, Rullo et al. [29]	Comprehensive review	General healthcare	172	2010–2021	No	No	No	Yes	Yes	Yes	Yes	Yes
De Roock and Martin [9]	Systematic review	General healthcare	263	~Jan 2021	No	No	No	Yes	Yes	No	Yes	Yes

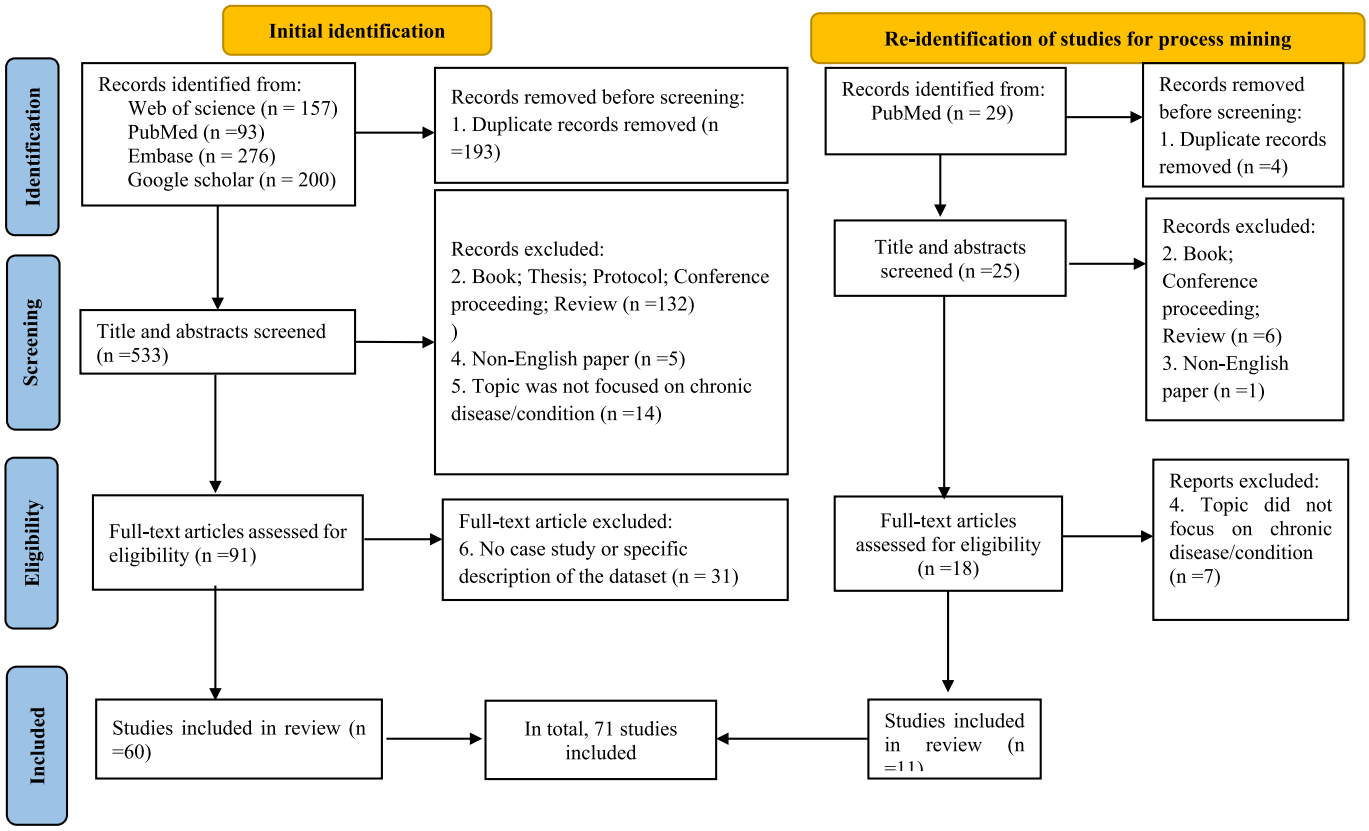


Fig. 1. Flow diagram of the review process.

individuals over a period of time [19]. Other study designs were grouped into the non-longitudinal study category.

2.4.3. Study category

In order to present a comprehensive overview of the research directions, we classified each primary study into a specific type based on the primary objective of each study. Specifically, a study was classified under the category of "prediction" if its principal aim involved making predictions, regardless of whether the study focused on early disease prediction or risk factor prediction. The term "association study" denoted the exploration of association rules, patterns, or the investigation of associated risk factors and subgroups. The "disease trajectory" category encompassed primary studies that investigated disease trajectory, comorbidity, or the network analysis of multi-morbidity to understand disease progression. Lastly, the "clinical pathway" category referred to primary studies where the main topic of investigation was the examination of clinical pathways.

2.4.4. Analysis methods and presentation of results

Overview of the search results and screening procedure for eligibility using the PRISMA flow diagram. The search results and screening process were presented in accordance with the PRISMA 2020 statement [20]. Descriptive methods and data visualisation techniques were employed to summarise and illustrate the findings from the included studies. Microsoft Excel was used to store the information that was extracted from papers. All of the collected information and data visualisation results were analysed using R 4.2.1 and Python 3.9.

3. Results

3.1. Prior literature reviews concerning process mining in healthcare

To emphasise both the contribution of our systematic review and the unique differentiators it possesses in comparison to others, we conducted a comparative analysis involving ten additional reviews. Table 3 strategically positions our review within this collection of literature reviews, examining two primary dimensions: the general information and the focal points within the review.

3.2. Eligibility screening of studies

A total of 726 studies were retrieved through a search of databases (Web of Science, n = 157; PubMed, n = 93; Embase, n = 276; Google Scholar, n = 200). After removing 193 duplicate records, 533 were screened by titles and abstracts, and 442 were excluded based on the inclusion and exclusion criteria. After reviewing all of the results, 31 additional records were excluded as they did not include a case study or a description of the dataset as per our search criteria. During the initial search round, 60 primary studies were identified as eligible for review to extract pertinent information. Among these, 58 primary studies were about data mining, while two were about process mining. Re-identification was employed to search for additional process mining studies in PubMed. After screening and assessment, 11 additional process mining studies were included. The specific reasons for the exclusion of records are displayed in Fig. 1.

3.3. Characteristics of eligible studies

A total of 71 primary studies were included, covering 25 countries/regions worldwide (see Fig. 2). A large proportion of included articles were published in the United States, with 16 articles (22.5 %). China and the United Kingdom were the next most prevalent, with 9 articles (12.7 %) and 7 articles (9.9 %), respectively. In the data mining studies, the United States stands out with the highest number of publications, accounting for 15 out of a total of 58 papers. Among the 13 process mining studies, the United Kingdom stands out with 5 publications.

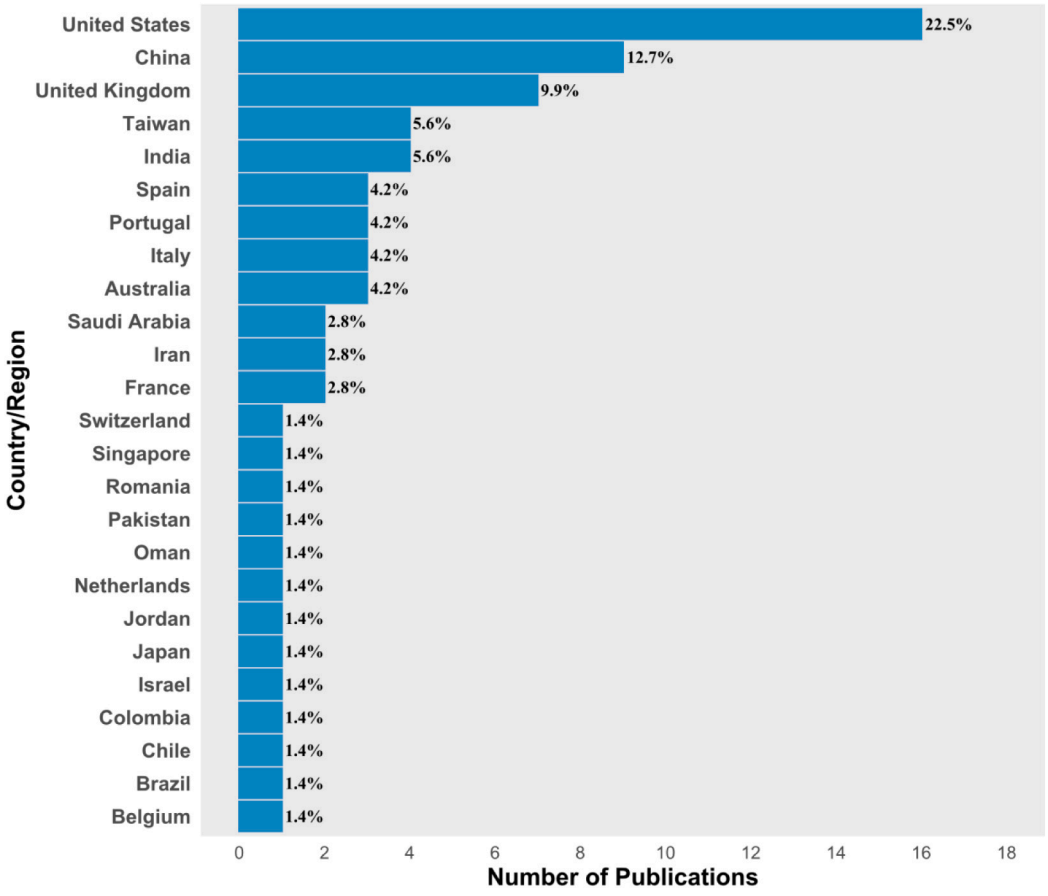


Fig. 2. The distribution of 71 included publications around the world from 2000 to 2022.

The characteristics of the 71 included articles are summarised in Table 4. Almost half, 31 (43.7 %) of included articles, have a sample size of >10,000 individuals. As for the type of study design, about 67.6 % of articles were longitudinal study designs. For the data source, most of the studies (81.7 %) used private datasets, while 19 % of studies analysed a public database, with the UCI machine learning repository being the most popular public dataset platform. Most of the articles presented in this study used R, WEKA, and Python. In the process mining studies, R emerged as the most commonly employed programming language, with pMineR [30] and bupaR [31] being the primary tools used for process mining. ProM, PMApp, and DISCO are specialised toolkits for process mining (see Table 4). According to the specific disease name in the study, we used the MeSH disease category to categorize the chronic diseases in the 71 papers into nine categories. These nine categories are shown in Table 4 and Fig. 3, which includes the category *multiple chronic diseases*, which indicates that the objective disease in the study was a combination of more than one chronic disease. Based on these nine categories, including the specific disease name, study title, study aims, method, and conclusion, an overview of all 71 articles is presented in Supplementary T1. Within data mining studies, the research area that emerged as the most prevalent was Nutritional and Metabolic Diseases (34.5 %). Notably, diabetes stood out as the most prevalent among these diseases. Conversely, in the realm of process mining studies, the research area that was the most prevalent was Neoplasms (53.8 %) (see Fig. 3).

3.4. The synergy of data mining, process mining, and traditional statistical methods

During our review, primary studies that incorporated statistical analysis, involved statistical hypothesis testing, or employed statistical inference were classified as utilising traditional statistical methods. In the 58 data mining studies reviewed, about 81 % utilised data mining methods as their primary approach, although they often incorporated and compared multiple data mining methods. Furthermore, 19 % of the data mining studies combined data mining methods with traditional statistical methods. Among the 13 process mining studies examined, most (69 %) integrated process mining methods with either data mining or traditional statistical methods. The other 31 % of the process mining studies exclusively employed process mining methods. Regarding the specific process mining methods (also referred to as miners in the literature), 12 out of the 13 studies utilised a single process mining method/miner, while one study compared and contrasted four different methods/miners (see Table 4).

3.5. Time trend

As shown in Fig. 4, the number of studies published each year fluctuated from 1 January 2000 to 31 December 2022. A notable absence of eligible papers was observed between the years 2000 and 2006. However, a sharp increase in the number of studies was observed starting in 2017 in both data mining studies and process mining studies, reaching a peak in 2021. In terms of data mining studies, the most prevalent category among the publications was prediction. Conversely, the most frequently published category in process mining studies focused on clinical pathways.

3.6. Overview of the methods employed in the included primary studies

In total, 62 methods were used in the 71 included articles. The distribution of methods differed between longitudinal and non-longitudinal studies (Fig. 5). Here, we summarised all process mining methods/miners as process mining. The most prevalent methods in the

Table 4
The characteristics of the 71 included studies published worldwide between 2000 and 2022.

Characteristics		Number of articles (%)	References
Number of study population	<100	2 (2.8)	[32,33]
	100–999	21 (29.6)	[31,34–53]
	1000–9999	17 (23.9)	[30,54–69]
	10,000~	31 (43.7)	[70–100]
Study design	Longitudinal	48 (67.6)	[30,31,35,39,50,51–57,60–63,65–76,78–84,86–91,93–95,97–100]
	Non-longitudinal	23 (32.4)	[32–34,36–38,40–49,58,59,64,77,85,92,96]
Follow-up for study length (years)	<5	22 (31.0)	[30,32,33,35,38,39,41,45,52,56,62,63,79,81,82,85,88,91,93,95,98,100]
	5–<10	15 (21.1)	[50,53–55,69–72,76,80,89,94,96,97,99]
	10–<20	14 (19.7)	[31,57,60,61,65,66,68,74,83,84,86,87,90,92]
	20~	2 (2.8)	[73,78]
	NA	18 (25.4)	[34,36,37,40,42–44,46–49,51,58,59,64,67,75,77]
Data source	Private	58 (81.7)	[30–33,35,36,38–41,47,50–63,68–100]
	UCI Machine Learning Repository	9 (12.7)	[34,37,42–46,48,49]
	SEER database	1 (1.4)	[65]
	NIDDK (the National Institute of Diabetes and Digestive and Kidney Diseases, U.S.)	1 (1.4)	[66]
	StatLib & NHANES	1 (1.4)	[64]
	Medical Information Mart for IntensiveCare III (MIMICIII)	1 (1.4)	[67]
Study category	Prediction	36 (50.7)	[32–34,36,38,40,42–49,55,58,59,62,64,66,67,70,71,73,76–78,80,82–84,91,94–97]
	Association study	13 (18.3)	[35,37,39,41,56,57,60,61,65,79,85,92,93]
	Disease trajectory	11 (15.5)	[54,63,72,74,75,81,86–90]
	Clinical pathway	11 (15.5)	[30,31,50–53,68,69,98–100]
Analysis software	R	23 (29.1)	[30,31,33,36,42,51,55,61,63,65,68,69,74,76,81,82,85–89,91,97]
	WEKA	8 (10.1)	[38,44,46,48,57,59,77,84]
	Python	4 (5.1)	[34,43,67,70]
	ProM	4 (5.1)	[31,51,52,99]
	SAS	3 (3.8)	[73,79,95]
	SPSS	3 (3.8)	[47,56,85]
	MATLAB	3 (3.8)	[64,83,92]
	Gephi	2 (2.5)	[54,83]
	PMApp	2 (2.5)	[69,72]
	CLUS	1 (1.3)	[40]
	Data Curation Tool	1 (1.3)	[75]
	JAVA	1 (1.3)	[41]
	MEKA	1 (1.3)	[40]
	RapidMiner	1 (1.3)	[84]
	SPSS Clementine	1 (1.3)	[35]
	STATA	1 (1.3)	[95]
	Teradata Warehouse Miner	1 (1.3)	[39]
	VESA Visualization Tool	1 (1.3)	[62]
	DISCO	1 (1.3)	[98]
	Not mentioned	17 (21.5)	[32,37,45,49,50,53,58,60,66,71,78,80,90,93,94,96,100]
Clinical category of disease	Nutritional and metabolic diseases:	22 (31.0)	[34,36,41,42,44,48,49,54–56,58,59,64,67,69,74,77,82–84,88,93]
	Diabetes		
	Dyslipidemia		
	Obesity		
	Multiple chronic diseases	16 (22.5)	[35,39,70,72,75,79,80,85,87,89–92,94–96]
	Neoplasms:	14 (19.7)	[30,31,37,45,46,51–53,60,65,68,73,81,99]
	Breast cancer		
	Colorectal cancers		
	Cervical cancer		
	Multiple cancers		
	Thyroid carcinoma		
	Urogenital diseases:	8 (11.3)	[38,43,50,63,66,76,97,98]
	Chronic kidney disease		
	Diabetic nephropathy		
	Respiratory tract diseases:	3 (4.2)	[32,33,71]
	Asthma		
	Chronic obstructive pulmonary disease		
	Nervous system diseases:	3 (4.2)	[40,47,78]
	Chronic migraine		
	Dementia		
	Multiple sclerosis		
	Cardiovascular diseases:	3 (4.2)	[57,62,100]
	Cardiovascular disease		
	Hypertension		

(*continued on next page*)

Table 4 (*continued*)

Characteristics		Number of articles (%)	References
	Musculoskeletal diseases: Spondyloarthritis	1 (1.4)	[61]
	Stomatognathic diseases: Periodontitis	1 (1.4)	[86]
Data mining (58 articles)	Data mining alone	47 (81.0)	[32–44,46–50,54,55,57–64,66,70,71,73,75–82,84,87,90,92–95]
	Data mining + traditional statistical methods	11 (19.0)	[45,56,65,74,83,85,88,89,91,96,97]
Process mining (13 articles)	Process mining alone	4 (30.8)	[31,51,98,99]
	Process mining + traditional statistical methods	3 (23.1)	[30,72,86]
	Process mining + data mining	4 (30.8)	[52,53,67,100]
	Process mining + data mining + traditional statistical methods	2 (15.4)	[68,69]
Prevalent data mining methods (n = 116)	Random Forest	10 (8.6)	[34,44,49,73,82–84,91,94,96]
	Logistic Regression	8 (6.9)	[42,44,73,82,83,85,91,94,96]
	Naive Bayes	6 (5.2)	[34,42,44,82–84]
	Neural Network	6 (5.2)	[34,49,56,57,73,82]
	Association Rule Mining	6 (5.2)	[35,37,57,60,85,89]
	Decision Tree	6 (5.2)	[41,47,73,77,83,96]
	SVM	6 (5.2)	[34,58,59,83,94,96]
	Clustering	4 (3.4)	[57,60,80,89]
	kNN	4 (3.4)	[34,82,83,96]
	AdaBoost	2 (1.7)	[34,77]
	Association Analysis	2 (1.7)	[39,80]
	Comorbidity Network	2 (1.7)	[54,55]
	Ensemble Deep Neural Networks	2 (1.7)	[33,78]
	Gradient Boosting	2 (1.7)	[42,96]
	J48	2 (1.7)	[44,84]
	K-means	2 (1.7)	[36,49]
	Sequential Pattern Mining	2 (1.7)	[71,93]
	Other*	44 (37.9)	
Prevalent process mining methods/miners (n = 16)	PALIA	2 (12.5)	[69,72]
	Inductive Miner-Infrequent (IMf)	2 (12.5)	[31,51]
	Careflow Miner (CFM)	2 (12.5)	[30,68]
	Inductive Visual Miner (IvM) & Petri Net	1 (6.3)	[99]
	Integer Linear Programming (ILP)	1 (6.3)	[31]
	interactive Data-Aware Heuristics Miner (iDHM)	1 (6.3)	[31]
	Inductive Miner (IM)	1 (6.3)	[31]
	Process Map	1 (6.3)	[86]
	Sequence Clustering Algorithm	1 (6.3)	[52]
	Decay Replay Mining	1 (6.3)	[67]
	No reported specific process mining methods	3 (18.8)	[53,98,100]

* Other indicates methods only shown once; one primary study can contain more than one method.

longitudinal studies were Process mining (n = 13, 31 %), Random Forest (n = 6, 14.3 %), Logistic Regression (n = 5, 11.9 %), Association Rule Mining (n = 4, 9.5 %), Neural Network (n = 4, 9.5 %), and Clustering (n = 4, 9.5 %). For the non-longitudinal studies, the most prevalent methods were Random Forest (n = 4, 13.3 %), Naive Bayes (n = 4, 13.3 %), Decision Tree (n = 4, 13.3 %), and SVM (n = 4, 13.3 %).

We conducted a succinct overview of the commonly employed models in the included studies: Random Forest, Logistic Regression, Naive Bayes, Neural Networks, Clustering, Association Rule Mining, Decision Tree, SVM, and Process Mining (Table 5).

4. Discussion

A total of 71 primary studies in chronic diseases were included (58 relating to data mining and 13 relating to process mining). This systematic review presents a comprehensive overview of the characteristics and methodologies employed in data mining and process mining for chronic conditions. Additionally, it highlights research gaps and challenges, and outlines future directions in the field.

4.1. Characteristics of included primary studies

In order to detect potential trends and ensure comprehensive coverage, this paper focuses on the period from 2000 to 2022, capturing advancements and developments both in data mining and process mining during this time frame. The time frame for this systematic review was selected based on the notable upsurge in the prevalence of data mining during the 1990s [108], with process mining gaining recognition starting in the 2000s [8]. Data availability plays a critical role in accurately assessing the utilisation of advanced analytics supported by data mining or process mining. The exclusion of research articles that do not report actual case studies aligns with the primary focus of this literature study, which aims to explore the application of information mining methods in real-world healthcare data scenarios. The number of publications in both data mining and process mining has significantly increased, indicating a growing interest in these disciplines. This suggests that methods from these domains are likely to be increasingly applied in the field of chronic disease research in the future [14].

Our study revealed that diabetes has emerged as a prominent research area within the domain of chronic disease studies. Out of the 71 articles examined, 19 specifically focused on the application of data mining methods in the context of diabetes. This chronic and complex

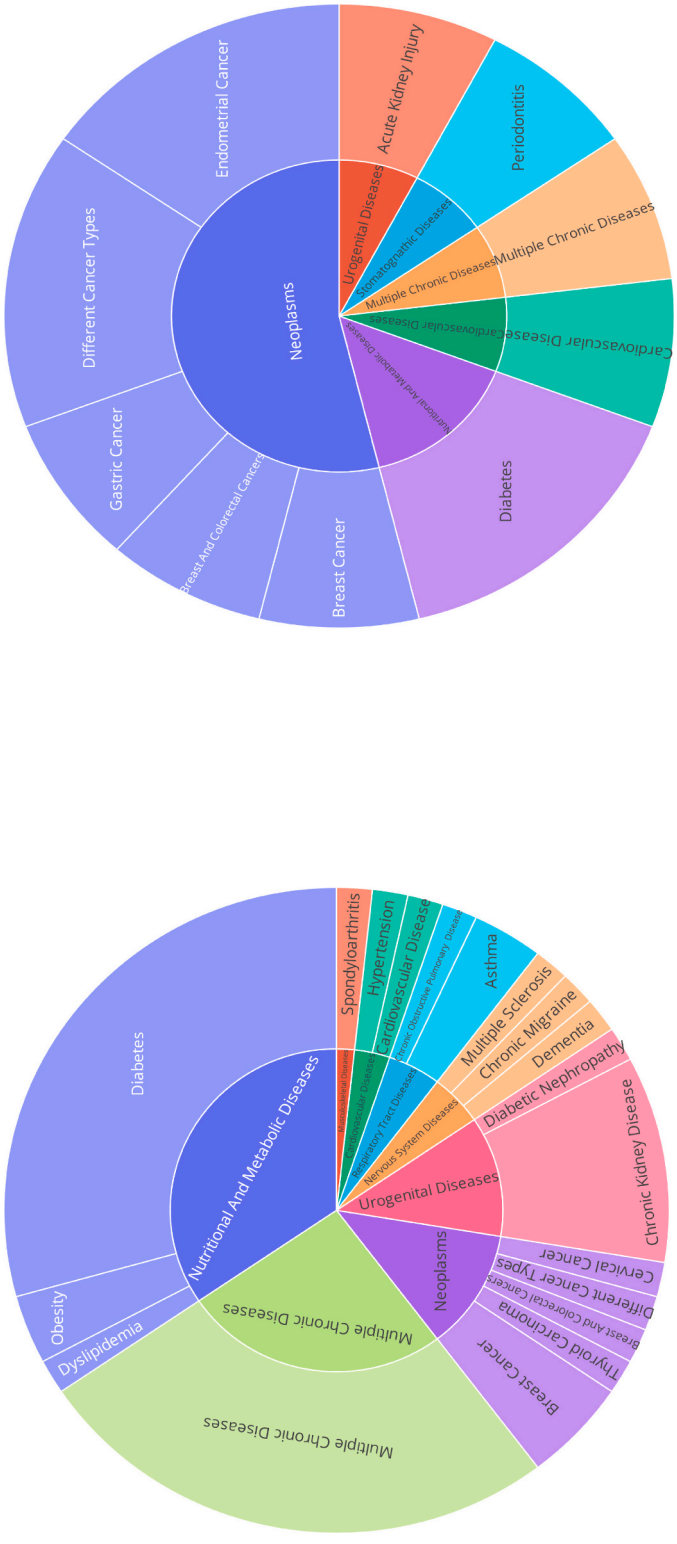


Fig. 3. Clinical category of disease and specific disease from the 71 primary studies in this review.

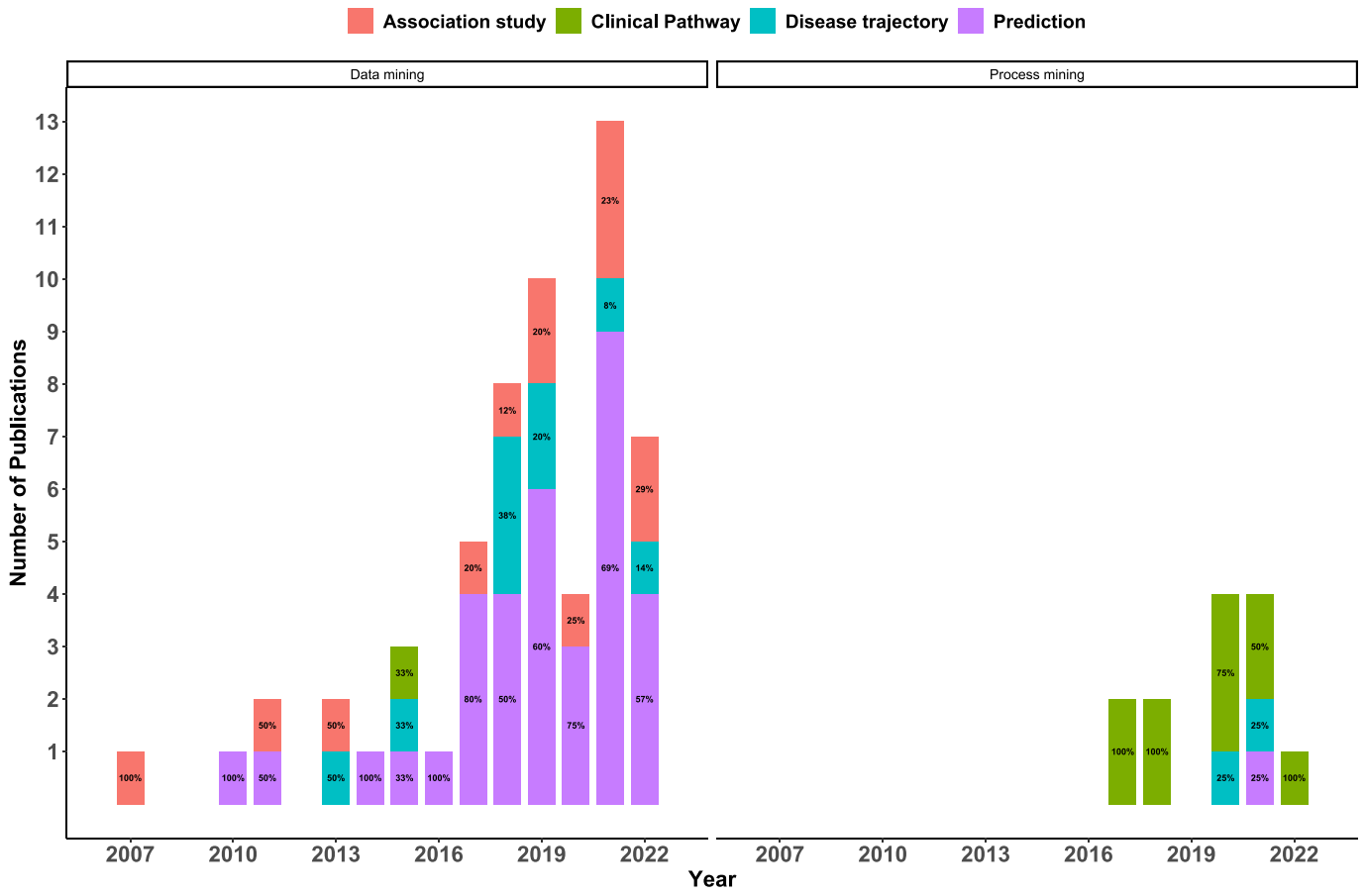


Fig. 4. Time trend for the number of publications and distribution of study categories between data mining studies and process mining studies from 2000 to 2022.

disease is associated with severe multi-morbidity, encompassing conditions such as cardiovascular disease, nephropathy, and retinopathy. The prevalence of diabetes-related illnesses is increasing among older and younger populations [109,110]. Considering the complexities associated with diabetes-related high prevalence multimorbidity, it is reasonable that most of the included primary studies directly address the application of advanced data mining methods in diabetes.

Furthermore, we have observed that cancer is the primary research area within the field of process mining. Cancer is a complex and multifaceted disease that poses substantial challenges in prevention, early detection, treatment, and management [111]. Understanding cancer aetiology and identifying effective strategies for prevention and management is crucial for the optimisation of cancer treatment [21,112]. Our analysis shows that process mining is widely used in cancer clinical pathway research because it allows researchers to analyse and detect complex healthcare processes and enhance the cancer treatment process.

Chronicity and complexity in the care of patients suffering from chronic diseases. Predictive analytics can substantially reduce healthcare costs by identifying people who are more prone to chronic disease and by facilitating early intervention before problems affect the population as a whole [109]. More than 50 % of the studies evaluated in this literature review focused specifically on prediction, including early diagnosis and incidence risk detection. The identified focus on better care through prevention in chronic care is not unexpected, as the link between healthcare sustainability, prevention, and chronic care was presented by Morgan, Zamora and Hindmarsh [113] >15 years ago.

4.2. *The prevailing methods used in the domain of chronic diseases*

This literature review indicates that the most prevalent data mining method is the random forest, which has shown better performance in risk factor assessment and the prediction of hospital readmission, survivability, and mortality in the present study [73,83,84,91]. Other popular applications of the random forest include gene expression microarrays, genome-wide studies, and protein-interaction prediction [114]. The popularity of RF may be due to its flexibility and applicability to different data features, including both longitudinal and non-longitudinal study designs. Although RF has “black box” characteristics and is more complex than decision trees, it contains effective techniques to provide explainability, such as counterfactual sets [115]. Furthermore, RF has been proven effective in survival analysis within the medical field. As an example, Scheffner et al. [116] conducted a study where they developed a survival RF model to predict the survival of renal transplant patients. The researchers opted for RF as the primary analytical technique due to its ability to handle a substantial number of variables while minimising overfitting, in contrast to the Cox Proportional Hazard Regression Model. Furthermore, RF provides an advantage in addressing intricate data interactions, which is a familiar challenge in conventional regression models [117]. However, it is important to note that RF has certain limitations in handling rare events and large datasets [101,117].

Process mining was a key focus of this review. It is a relatively new data-driven method in the healthcare domain that employs multidisciplinary techniques and visualisation methods to generate interpretable information [118]. Process mining encompasses three main techniques: process discovery, process conformance, and process enhancement [119]. Among these techniques, process discovery stands out as the most common approach used to detect underlying processes. In this review, we have observed substantial variations in the specific methods employed across different applications for process mining. Notably, within the studies included in our analysis, the most prevalent process

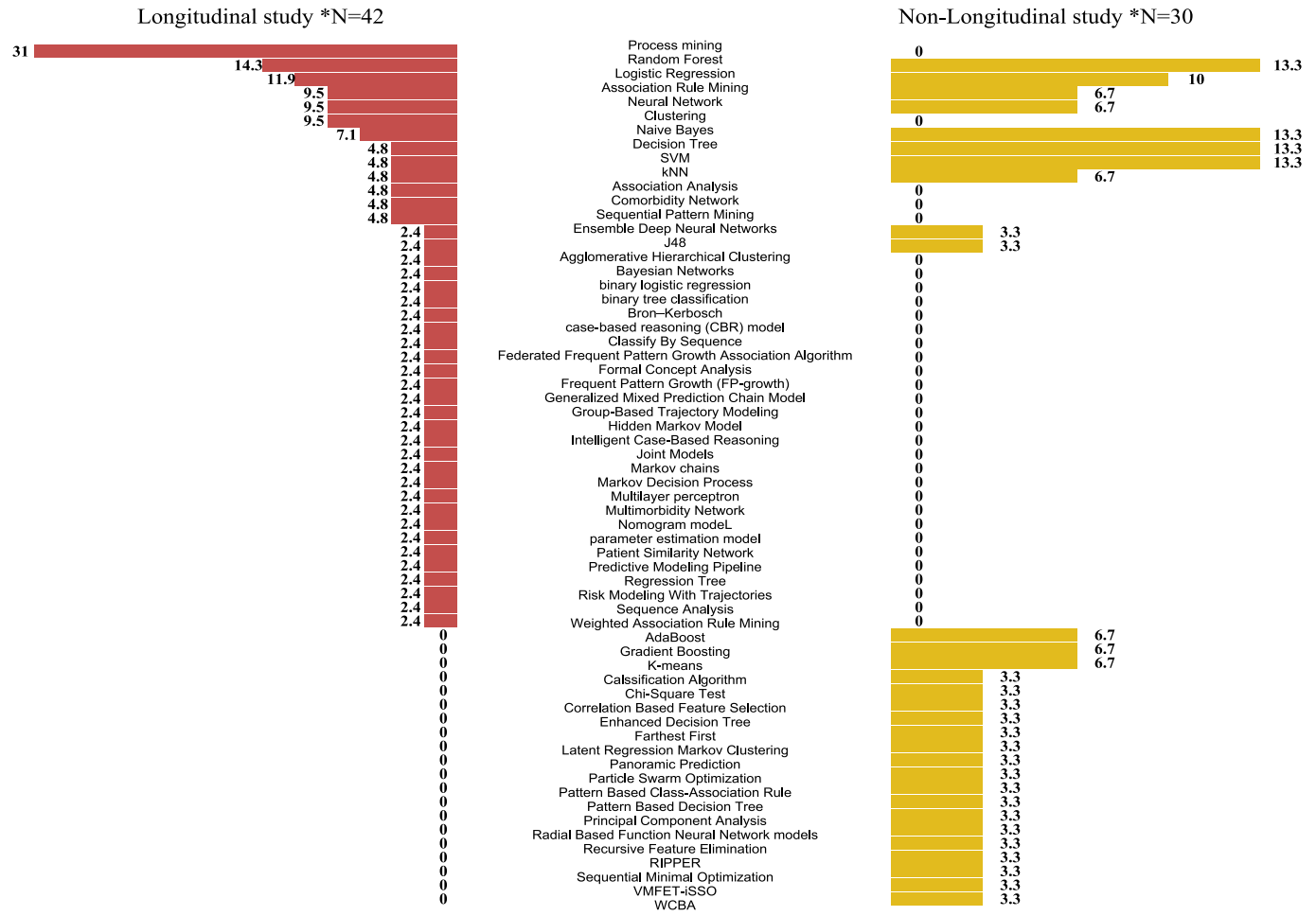


Fig. 5. The proportion (%) of methods utilised in longitudinal or non-longitudinal studies from the 71 included articles. *The frequency of methods appearing in longitudinal/non-longitudinal studies; a single study may employ multiple methods.

mining methods were identified within the domain of process discovery. Specifically, PALLA [69,72], Inductive Miner-Infrequent (IMf) [31,51], and Careflow Miner (CFM) [30,68] were prominent methods used for process discovery (each reported in 2 primary studies). However, it is important to note that these methods are tailored to distinct situations. Furthermore, three studies did not explicitly specify the specific process mining methods used but instead referred to the broader terms “process mining” or “process mining method”. Moreover, it is important to note that the use of distinct methods requires the utilisation of specific toolkits or applications. For instance, PALLA is applicable within the PMApp, the Inductive Miner-Infrequent method can be employed within ProM, and Careflow Miner can be used within pMineR.

As process mining was developed with a focus on specific study designs, it is particularly well-suited for longitudinal studies that target process-oriented analyses. Although the application of process mining to the study of chronic diseases can be challenging due to the inherent complexity and intricacy in healthcare processes [9], the approach has gained significant prominence in modelling clinical pathways. In fact, most process mining studies in our review focused on modelling clinical pathways. This finding aligns with another systematic review on process mining, which reported that 69 % of healthcare studies focused on clinical processes [9]. This implies that process mining offers certain advantages in terms of managing complex, unstructured healthcare processes. The application of process mining in healthcare has the potential to enhance care process efficiency and reduce wasteful and non-value-added activities. The World Health Organization (WHO) has reported a significant level of wasteful and non-added value activities (30 %), further emphasising the need for improvement in this area [120].

Among the 13 process mining studies reviewed, only four publications exclusively utilised process mining, while the remaining nine publications employed a combination of traditional statistical methods or data mining methods. Process mining offers the advantage of directly visualising traces, thus facilitating the comprehension of complex healthcare processes. However, when the research objective involves comparison, prediction, or the identification of associations, relying solely on process mining may present challenges in achieving satisfactory outcomes. Traditional statistical methods are widely accepted as the primary means of establishing associations in the medical field. Therefore, incorporating both statistical methods and data mining methods alongside process mining can be valuable complementary approaches. Notably, we found that Fisher’s exact test [30,69] and Cox regression [30,68,69,86] were the traditional statistical methods that were most frequently combined with process mining. While process mining alone may not yield significant results, the integration of process mining with statistical methods presents a promising approach to obtain meaningful outcomes. We also found that clustering [52,69] and the Markov model [53,100] are often combined with process mining. Clustering is an important preprocessing method that is frequently used in process mining, which can assist in identifying process variants or different behavioural patterns within a process. By clustering similar process traces or event sequences, researchers can uncover distinct process models or patterns. Markov models provide a mathematical framework for modelling dynamic systems and processes [53,100]. Researchers can capture and represent the probabilistic behaviour and

Table 5
Overview of common methods from the 71 included primary studies.

Methods	Characters	Highlights from the included primary studies
Random Forest (RF)	RF is a flexible method that offers a multitude of advantages [101], thereby establishing its position as the most widely adopted methodology. Nonetheless, it is important to acknowledge that RF possesses a "black box" nature and may encounter limitations when confronted with large datasets.	A study [83] showed that random forest is the best algorithm to use if the purpose is to find type 2 diabetes patients who are truly at risk of CVD. In another study [84], random forest also showed the best performance in predicting hospital readmission of diabetic patients, with an accuracy of 0.898. In a third study [73], random forest outperforms the other five models in lowering type I errors in the prediction of overall survivability in cancer patients with comorbidity. In a fourth study [91], random forest outperformed logistic regression for predicting the mortality of patients with multi-morbidity.
Logistic regression (LR)	LR is commonly referred to as a "white box" model, where the size of coefficients serves as an indicator of their relative significance in influencing the classification outcome. However, logistic regression exhibits comparatively lower discriminative performance when compared to other "black box" methods [102].	In one study [44], the logistic regression model showed the best performance for predicting diabetic patients. In other studies [42,73,82,91], logistic regression was used to predict diabetes risk, which showed good performance but not the best.
Naive Bayes (NB)	Naive Bayes is a classification technique that uses Bayes' theory [93], exhibiting efficiency and effectiveness even with small datasets. However, it is important to note that Naive Bayes relies on the assumption of conditional independence among features, which may limit its performance in scenarios where this assumption does not hold [103].	In a number of studies [42,44,82–84,96], Naive Bayes was used to develop a predictive model that showed good performance but was not the best compared to other algorithms.
Neural networks	Neural networks are also known as Artificial neural networks (ANNs). The main advantage of neural networks is that they offer the ability to identify intricate non-linear relationships between dependent and independent variables implicitly. However, ANNs do have a "black box" character [104].	In a number of studies [34,49,82], neural networks showed the best performance in the prediction of a diabetes diagnosis. ANNs are better than other statistical techniques at capturing non-linear relationships [49]. In another study [56], ANNs were used to identify the incidence of dyslipidemia risk factors.
Clustering	Clustering is a statistical data analysis method widely used in various disciplines, such as pattern identification, image analysis, machine learning, data mining, and bioinformatics [105].	A number of studies [57,60,80,89] used clustering to group patients based on various similarities and heterogeneity criteria to better detect behaviour within a cluster. These studies also used association rule mining after clustering. While clustering approaches construct clusters based on similarities, association rule mining uncovers relationships based on co-occurrences [106]. In practice, clustering deals with samples (patients) and association rule mining deals with predictors (variables).
Association rule mining (ARM)	ARM is frequently used to explore relationships among all selected variables. Link analysis and sequence mining are two frequently used association rule learning techniques. Link analysis is used to find the connections among many possible related factors. Sequence mining finds relationships that occur over time [35].	One study [35] used ARM to extract useful association rules from big data to detect information that clinicians could use to manage chronic diseases. One study [37] used ARM to extract useful knowledge that could assist medical professionals in improving cervical cancer screening programmes to provide patients with prompt treatment at an earlier stage. ARM is also useful in finding associated hidden symptom patterns in patients with cancer [60]. Two studies [85,89] used association rule mining analysis to explore multi-morbidity patterns and relationships among chronic diseases.
Decision tree	A decision tree is a classification technique [73]. It is capable of handling data that contains errors or missing values during the training process, although it requires large datasets to yield accurate outcomes [103].	One study [41] used a decision tree with a particle swarm optimisation algorithm to identify risk factors for type II diabetes. Another study [47] proposed the use of a decision tree to develop diagnostic criteria and assessment tools for myologic encephalomyelitis/ multiple sclerosis. In other studies [73,77,83,96], a decision tree was used to develop the predictive model, which showed good performance but not the best compared to other algorithms.
Support vector machine (SVM)	SVM is a supervised machine learning technique that uses classification algorithms for bi-category classification tasks. It is built on statistical learning theory [107]. However, SVM also possesses "black box" characteristics.	One study [58] proposed SVM combined with an additional rule-based explanation component to diagnose diabetes and provide a comprehensive explanation simultaneously. The results of another study [59] showed that the SVM-based prediction model offered many advantages, such as high accuracy, robustness, rapid learning speed, and superior classification performance. Several studies [22,73,86] used SVM to develop a predictive model that showed good performance but not the best performance.
Process mining (PM)	Process transparency is the advantage of PM. However, domain knowledge is essential for drawing meaningful insights from process mining analyses [9].	Most of the process mining studies included in the analysis revolve around the domain of clinical pathways. One study [40] compared various process mining methods for process discovery, wherein the iDHM miner demonstrated the best performance. Another study [67] pioneered a process mining-based methodology for the prediction of in-hospital mortality. Other studies [53,90] employed process mining in conjunction with Markov models to effectively model treatment pathways in patients

(*continued on next page*)

Table 5 (*continued*)

Methods	Characters	Highlights from the included primary studies
		diagnosed with cardiovascular diseases and cancer. Before conducting process mining discovery, two other studies [52,69] applied clustering techniques to their respective datasets. Another study [72] showed that process mining has the capacity to develop dynamic risk models for chronic conditions; this study obtained three process mining maps corresponding to three chronic conditions, which are easier for researchers to understand and measure, and help manage the disease process. Another study [86] used process mining to discover disease trajectories in periodontitis patients, through which researchers could check related multi-morbidity progression.

transitions between different states or activities by incorporating process mining (e.g. of event logs or process traces) into a Markov model. This combined approach allows researchers to gain significant insights and draw robust conclusions within the context of chronic disease research.

4.3. Research gaps, challenges, and future research directions

In the present review, we observed a lack of standardisation in the structure of reports, leading to potential confusion and difficulties in interpretation. Inconsistencies in reporting methods, such as variations in report structure and terminology, further exacerbate this issue. One widely recognised guideline is the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) [121]. Although TRIPOD primarily focuses on prediction models, it provides valuable guidance for the reporting of data mining studies in the medical domain. Furthermore, several of the included studies adhered to the STROBE (STrengthening the Reporting of OBservational studies in Epidemiology) guideline [122], which provides support for reporting in the context of research involving observational study designs. However, as far as we know, there is currently no standardised guideline for reporting process mining studies in the medical field. Moreover, the diverse landscape of process mining methods, as discussed in Section 4.2, reveals variations across different applications. Consequently, there is a need to refine the reporting structure and terminology pertaining to process mining in order to establish a standard framework. Such standardisation is essential for enhancing the consistency and clarity of process mining research in both academic and practical domains.

Furthermore, caution should be exercised when combining traditional statistical methods with data mining or process mining. For instance, multiplicity can arise in situations where statistical comparisons are made between outcomes across multiple trajectories or nodes identified through process mining. In order to avoid misleading results, it is important to understand the underlying assumptions of the statistical methods and their appropriateness in the given context. Therefore, it is important for researchers to have an adequate understanding of statistical assumption when applying information mining methods in clinical research [123].

Challenges related to model explainability and interpretation in clinical settings persist. Despite ongoing efforts to address these challenges, such as the use of counterfactual sets [115] in random forest, there is a continued need for more interpretable methods in advanced models. In the area of medical or clinical interpretation, our review primarily focused on primary studies that included case studies. However, we observed a tendency among many studies to prioritise the development of methods over the final medical outcome, resulting in limited clinical interpretation. In the context of process validation in process mining, the involvement of domain experts, such as clinical physicians, is crucial. Process mining techniques can benefit from the application of interactive methodologies that facilitate the bi-directional interaction between artificial intelligence and human knowledge. This enables a better understanding of decisions and allows for corrections and improvements guided by human expertise [124], which is in line with the findings of another study [9]. Our review found that only one out of the thirteen primary studies [69] involved clinical physicians in the evaluation of process models; other studies were performed by clinical researchers or a multidisciplinary team. This emphasises the need to include clinical expertise in order to ensure the relevance and usability of derived process models. Neglecting the proper interpretation of the significance of the application in the clinical field can result in missed opportunities to improve patient care and outcomes.

4.4. Limitations

There are some limitations in our systematic review. Firstly, the search strategy used in this review followed the Prisma standard in order to create better universality and standardisation, which may have resulted in the omission of relevant primary studies. Additionally, due to the notable disparity in conference paper formats between the fields of technology and medicine, we excluded these papers. There is the potential that these papers may have contained significant applications in the area of data mining or process mining in chronic diseases that we failed to include. Additionally, only one author performed the initial screening and retrieval of all the included articles, which may entail a certain level of bias. To avoid potential bias, the remaining authors screened and supervised the selection methodology.

5. Conclusion

In conclusion, there has been an increasing use of data mining and process mining methods in chronic disease research, often through collaborations that involve process mining, data mining, and traditional statistical methods. It is essential to focus on the most relevant questions in chronic disease research in order to improve sustainability and public health. Our systematic review has provided a comprehensive overview of the current trends in these two disciplines. Additionally, we have identified several research gaps and challenges that warrant further attention. Future research endeavours should focus on standardising reporting practices, enhancing the interpretation of results in the medical field, and exploring the potential to conduct interactive process mining that involves domain experts. These efforts hold the potential to advance the application of process mining in healthcare.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.artmed.2023.102645.

Declaration of competing interest

Farhad Abtahi is the founder and CEO of Wergonic AB. Wergonic AB is a start-up company developing digital solutions for precision ergonomics.

Acknowledgements

This activity has received funding from EIT Health (www.eithealth.eu), ID 220649, the innovation community on Health of the European Institute of Innovation and Technology (EIT), a body of the European

Union, under Horizon 2020, the EU Framework Programme for Research and Innovation.

References

[1] Hajat C, Stein E. The global burden of multiple chronic conditions: a narrative review. Prev Med Rep 2018;12:284–93.
[2] Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019;110:12–22.
[3] Mannila H. Data mining: machine learning, statistics, and databases. IEEE; 1996. p. 2–9.
[4] Battineni G, Sagaro GG, Chinatalapudi N, Amenta F. Applications of machine learning predictive models in the chronic disease diagnosis. J Personalized Med 2020;10:21.
[5] Campbell H, Hotchkiss R, Bradshaw N, Porteous M. Integrated care pathways. BMJ. 1998;316:133–7.
[6] Jothi N, Husain W. Data mining in healthcare–a review. Procedia Comput Sci 2015;72:306–13.
[7] van der Aalst W. Data science in action. In: van der Aalst W, editor. Process mining: data science in action. Berlin, Heidelberg: Springer Berlin Heidelberg; 2016. p. 3–23.
[8] Maruster L, van der Aalst W, Weijters T, van den Bosch A, Daelemans W. Automated discovery of workflow models from hospital data. B Kr€ oose, M de Rijke 2001:18.
[9] De Roock E, Martin N. Process mining in healthcare–an updated perspective on the state of the art. J Biomed Inform 2022;103995.
[10] Fernandez-Llatas C, Martinez-Millana A, Martinez-Romero A, Benedi JM, Traver V. Diabetes care related process modelling using Process Mining techniques. Lessons learned in the application of Interactive Pattern Recognition: coping with the Spaghetti Effect. Annu Int Conf IEEE Eng Med Biol Soc 2015;2015:2127–30.
[11] Kusuma G, Sykes S, Mcinerney C, Johnson O. Process mining of disease trajectories: a feasibility study. In: Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies. SCITEPRESS - Science and Technology Publications; 2020.
[12] de Toledo P, Joppien C, Sesmero MP, Drews P. Mining disease courses across organizations: a methodology based on process mining of diagnosis events datasets. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE; 2019. p. 354–7.
[13] Pebesma J, Martinez-Millana A, Sacchi L, Fernandez-Llatas C, De Cata P, Chiovato L, et al. Clustering cardiovascular risk trajectories of patients with type 2 diabetes using process mining. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE; 2019. p. 341–4.
[14] Marinov M, Mosa ASM, Yoo I, Boren SA. Data-mining technologies for diabetes: a systematic review. J Diabetes Sci Technol 2011;5:1549–56.
[15] Ahmad P, Qamar S, Rizvi SQA. Techniques of data mining in healthcare: a review. Int J Comput Appl 2015:120.
[16] Guzzo A, Rullo A, Vocaturo E. Process mining applications in the healthcare domain: a comprehensive review. Wiley Interdiscipl Rev Data Min Knowl Discov 2022;12:e1442.
[17] Dallagassa MR, dos Santos Garcia C, Scalabrin EE, Ioshii SO, Carvalho DR. Opportunities and challenges for applying process mining in healthcare: a systematic mapping study. J Ambient Intell Humaniz Comput 2022:1–18.
[18] Haddaway NR, Collins AM, Coughlin D, Kirk S. The role of Google Scholar in evidence reviews and its applicability to grey literature searching. PloS One 2015;10:e0138237.
[19] Caruana EJ, Roman M, Hernández-Sánchez J, Solli P. Longitudinal studies. J Thorac Dis 2015;7:E537–40.
[20] Page MJ, Mckenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Syst Rev 2021;10.
[21] Kurniati AP, Johnson O, Hogg D, Hall G. Process mining in oncology: a literature review. In: 2016 6th International Conference on Information Communication and Management (ICICM); 2016. p. 291–7.
[22] Ghasemi M, Amyot D. Process mining in healthcare: a systematised literature review. Int J Electron Healthc 2016;9.
[23] Rojas E, Munoz-Gama J, Sepúlveda M, Capurro D. Process mining in healthcare: a literature review. J Biomed Inform 2016;61:224–36.
[24] Kusuma GP, Hall M, Gale CP, Johnson OA. Process mining in cardiology: a literature review. Int J Biosci Biochem Bioinforma 2018;8:226–36.
[25] Batista E, Solanas A. Process mining in healthcare: a systematic review. In: 2018 9th International Conference on Information, Intelligence, Systems and Applications (IISA). IEEE; 2018. p. 1–6.
[26] Williams R, Rojas E, Peek N, Johnson OA. Process mining in primary care: a literature review. Stud Health Technol Inform 2018;247:376–80.
[27] Sundari MS, Nayak RK. Process mining in healthcare systems: a critical review and its future. Int J Emerg Trends Eng Res 2020:8.
[28] Grüger J, Bergmann R, Kazik Y, Kuhn M. Process mining for case acquisition in oncology: a systematic literature review. LWDA. 2020:162–73.
[29] Guzzo A, Rullo A, Vocaturo E. Process mining applications in the healthcare domain: a comprehensive review. WIREs Data Min Knowl Discov 2021:12.
[30] Cuendet MA, Gatta R, Wicky A, Gerard CL, Dalla-Vale M, Tavazzi E, et al. A differential process mining analysis of COVID-19 management for cancer patients. Front Oncol 2022;12:1043675.
[31] Kurniati AP, McInerney C, Zucker K, Hall G, Hogg D, Johnson O. Using a multi-level process comparison for process change analysis in cancer pathways. Int J Environ Res Public Health 2020;17.
[32] Lee CH, Chen JCY, Tseng VS. A novel data mining mechanism considering bio-signal and environmental data with applications on asthma monitoring. Comput Methods Programs Biomed 2011;101:44–61.
[33] Khasha R, Sepehri MM, Taherkhani N. Detecting asthma control level using feature-based time series classification. Appl Soft Comput 2021;111:16.
[34] Chaves L, Marques G. Data mining techniques for early diagnosis of diabetes: a comparative study. Appl Sci Basel 2021;11:12.
[35] Huang MJ, Sung HS, Hsieh TJ, Wu MC, Chung SH. Applying data-mining techniques for discovering association rules. Soft Comput 2020;24:8069–75.
[36] Muhammad MU, Ren JD, Muhammad NS, Hussain M, Muhammad I. Principal component analysis of categorized polytomous variable-based classification of diabetes and other chronic diseases. Int J Environ Res Public Health 2019;16:15.
[37] Lee CKH, Tse YK, Ho GTS, Chung SH. Uncovering insights from healthcare archives to improve operations: an association analysis for cervical cancer screening. Technol Forecast Soc Chang 2021;162:11.
[38] Turiac AS, Zdrodowska M. Data mining approach in diagnosis and treatment of chronic kidney disease. Acta Mech Automatica 2022;16:180–8.
[39] Imamura T, Matsumoto S, Kanagawa Y, Tajima B, Matsuya S, Furue M, et al. A technique for identifying three diagnostic findings using association analysis. Med Biol Eng Comput 2007;45:51–9.
[40] Bravo FP, Garcia AAD, Veiga ABG, de la Sacristana MMG, Pinero MR, Peral AG, et al. SMURF: systematic methodology for unveiling relevant factors in retrospective data on chronic disease treatments. IEEE Access 2019;7:92598–614.
[41] Abdullah AS, Selvakumar S. Assessment of the risk factors for type II diabetes using an improved combination of particle swarm optimization and decision trees by evaluation with Fisher's linear discriminant analysis. Soft Comput 2019;23:9995–10017.
[42] Birjais R, Mourya AK, Chauhan R, Kaur H. Prediction and diagnosis of future diabetes risk: a machine learning approach. SN Appl Sci 2019;1:8.
[43] Chaudhuri AK, Sinha D, Banerjee DK, Das A. A novel enhanced decision tree model for detecting chronic kidney disease. Netw Model Anal Health 2021;10:22.
[44] Battineni G, Sagaro GG, Nalini C, Amenta F, Tayebati SK. Comparative machine-learning approach: a follow-up study on type 2 diabetes predictions by cross-validation methods. Machines. 2019;7:11.
[45] Osman AH, Aljahdali HMA. An effective of ensemble boosting learning method for breast cancer virtual screening using neural network model. IEEE Access 2020;8:39165–74.
[46] Alwidian J, Hammo BH, Obeid N. WCBA: weighted classification based on association rules algorithm for breast cancer disease. Appl Soft Comput 2018;62:536–49.
[47] Ohanian D, Brown A, Sunnquist M, Furst J, Nicholson L, Klebek L, et al. Identifying key symptoms differentiating myalgic encephalomyelitis and chronic fatigue syndrome from multiple sclerosis. Neurology (ECronicon) 2016;4:41–5.
[48] Howsalya Devi RD, Bai A, Nagarajan N. A novel hybrid approach for diagnosing diabetes mellitus using farthest first and support vector machine algorithms. Obes Med 2020:17.
[49] Mahboob Alam T, Iqbal MA, Ali Y, Wahab A, Ijaz S, Imtiaz Baig T, et al. A model for early prediction of diabetes. Inform Med Unlocked 2019;16.
[50] Zhang Y, Padman R. Innovations in chronic care delivery using data-driven clinical pathways. Am J Manag Care 2015;21:e661–8.
[51] Kurniati AP, Rojas E, Zucker K, Hall G, Hogg D, Johnson O. Process mining to explore variations in endometrial cancer pathways from GP referral to first treatment. Stud Health Technol Inform 2021;281:769–73.
[52] Villamil MDP, Barrera D, Velasco N, Bernal O, Fajardo E, Urango C, et al. Strategies for the quality assessment of the health care service providers in the treatment of gastric cancer in Colombia. BMC Health Serv Res 2017;17:654.
[53] Baker K, Dunwoodie E, Jones RG, Newsham A, Johnson O, Price CP, et al. Process mining routinely collected electronic health records to define real-life clinical pathways during chemotherapy. Int J Med Inform 2017;103:32–41.
[54] Khan A, Uddin S, Srinivasan U. Comorbidity network for chronic disease: a novel approach to understand type 2 diabetes progression. Int J Med Inform 2018;115:1–9.
[55] Khan A, Uddin S, Srinivasan U. Chronic disease prediction using administrative data and graph theory: the case of type 2 diabetes. Expert Syst Appl 2019;136:230–41.
[56] Rezaei M, Fakhri N, Pasdar Y, Moradinazar M, Najafi F. Modeling the risk factors for dyslipidemia and blood lipid indices: Ravansar cohort study. Lipids Health Dis 2020;19:8.
[57] Pasanisi S, Paiano R. A hybrid information mining approach for knowledge discovery in cardiovascular disease (CVD). Information. 2018;9:14.
[58] Barakat NH, Bradley AP, Barakat MNH. Intelligible support vector machines for diagnosis of diabetes mellitus. IEEE T Inf Technol Biomed 2010;14:1114–20.
[59] Guo H, Fan ZC, Zeng Y. Novel data mining analysis method on risk prediction of type 2 diabetes. J Sign Process Syst 2022;94:1183–98.
[60] Luo X, Gandhi P, Storey S, Zhang ZY, Han Z, Huang K. A computational framework to analyze the associations between symptoms and cancer patient attributes post chemotherapy using EHR data. IEEE J Biomed Health Inform 2021;25:4098–109.

[61] Barata C, Rodrigues AM, Canhao H, Vinga S, Carvalho AM. Predicting biologic therapy outcome of patients with spondyloarthritis: joint models for longitudinal and survival analysis. JMIR Med Inf 2021;9:17.
[62] Sun J, McNaughton CD, Zhang P, Perer A, Gkoulalas-Divanis A, Denny JC, et al. Predicting changes in hypertension control using electronic health records from a chronic disease management program. J Am Med Inform Assoc 2014;21:337–44.
[63] 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.
[64] Lai CM, Chiu CC, Shih YC, Huang HP. A hybrid feature selection algorithm using simplified swarm optimization for body fat prediction. Comput Methods Programs Biomed 2022:226.
[65] Jin S, Liu H, Yang J, Zhou J, Peng D, Liu X, et al. Development and validation of a nomogram model for cancer-specific survival of patients with poorly differentiated thyroid carcinoma: a SEER database analysis. Front Endocrinol 2022;13.
[66] You YX, Hua ZS, Dong FQ. Generalized mixed prediction chain model and its application in forecasting chronic complications. J Oper Res Soc 2023;74(7): 1815–35.
[67] Theis J, Galanter WL, Boyd AD, Darabi H. Improving the in-hospital mortality prediction of diabetes ICU patients using a process mining/deep learning architecture. IEEE J Biomed Health Inform 2022;26:388–99.
[68] Chiudinelli L, Dagliati A, Tibollo V, Albasini S, Geifman N, Peek N, et al. Mining post-surgical care processes in breast cancer patients. Artif Intell Med 2020;105: 101855.
[69] Conca T, Saint-Pierre C, Herskovic V, Sepúlveda M, Capurro D, Prieto F, et al. Multidisciplinary collaboration in the treatment of patients with type 2 diabetes in primary care: analysis using process mining. J Med Internet Res 2018;20:e127.
[70] Hu ZX, Qiu H, Wang LY, Shen MH. Network analytics and machine learning for predicting length of stay in elderly patients with chronic diseases at point of admission. BMC Med Inform Decis Mak 2022;22:15.
[71] Cheng YT, Lin YF, Chiang KH, Tseng VS. Mining sequential risk patterns from large-scale clinical databases for early assessment of chronic diseases: a case study on chronic obstructive pulmonary disease. IEEE J Biomed Health Inform 2017;21: 303–11.
[72] Valero-Ramon Z, Fernandez-Llatas C, Valdivieso B, Traver V. Dynamic models supporting personalised chronic disease management through healthcare sensors with interactive process mining. Sensors. 2020;20:25.
[73] Zolbanin HM, Delen D, Zadeh AH. Predicting overall survivability in comorbidity of cancers: a data mining approach. Decis Support Syst 2015;74:150–61.
[74] Oh W, Kim E, Castro MR, Caraballo PJ, Kumar V, Steinbach MS, et al. Type 2 diabetes mellitus trajectories and associated risks. Big Data 2016;4:25–30.
[75] Carmona-Pirez J, Poblador-Plou B, Poncel-Falco A, Rochat J, Alvarez-Romero C, Martinez-Garcia A, et al. Applying the FAIR4Health solution to identify multimorbidity patterns and their association with mortality through a frequent pattern growth association algorithm. Int J Environ Res Public Health 2022;19: 10.
[76] Nenova Z, Shang J. Chronic disease progression prediction: leveraging case-based reasoning and big data analytics. Prod Oper Manag 2022;31:259–80.
[77] Alshammari R, Almutairi N. Building diabetes early warning system using data mining techniques. J Med Imaging Health Inform 2017;7:655–9.
[78] Tsang G, Zhou SM, Xie X. Modeling large sparse data for feature selection: hospital admission predictions of the dementia patients using primary care electronic health records. IEEE J Transl Eng Health Med 2021;9:1–13.
[79] Newcomer SR, Steiner JF, Bayliss EA. Identifying subgroups of complex patients with cluster analysis. Am J Manag Care 2011;17:E324–32.
[80] Ding RK, Jiang F, Xie JG, Yu YG. Algorithmic prediction of individual diseases. Int J Prod Res 2017;55:750–68.
[81] Jay N, Nuemi G, Gadreau M, Quantin C. A data mining approach for grouping and analyzing trajectories of care using claim data: the example of breast cancer. BMC Med Inform Decis Mak 2013;13:9.
[82] Olivera AR, Roesler V, Iochpe C, Schmidt MI, Vigo A, Barreto SM, et al. Comparison of machine-learning algorithms to build a predictive model for detecting undiagnosed diabetes ELSA-Brasil: accuracy study. Sao Paulo Med J 2017;135:234–46.
[83] Hossain E, Uddin S, Khan A. Network analytics and machine learning for predictive risk modelling of cardiovascular disease in patients with type 2 diabetes. Expert Syst Appl 2021;164:13.
[84] Neto C, Senra F, Leite J, Rei N, Rodrigues R, Ferreira D, et al. Different scenarios for the prediction of hospital readmission of diabetic patients. J Med Syst 2021; 45:11.
[85] Lin WQ, Yuan LX, Sun MY, Wang C, Liang EM, Li YH, et al. Prevalence and patterns of multimorbidity in chronic diseases in Guangzhou, China: a data mining study in the residents' health records system among 31 708 community-dwelling elderly people. BMJ Open 2022;12:e056135.
[86] Larvin H, Kang J, Aggarwal VR, Pavitt S, Wu J. Multimorbid disease trajectories for people with periodontitis. J Clin Periodontol 2021;48:1587–96.
[87] Shi X, Nikolic G, Van Pottelbergh G, van den Akker M, Vos R, De Moor B. Development of multimorbidity over time: an analysis of Belgium primary care data using Markov chains and weighted association rule mining. J Gerontol A Biol Sci Med Sci 2021;76:1234–41.
[88] Madlock-Brown C, Reynolds RB. Identifying obesity-related multimorbidity combinations in the United States. Clin Obes 2019:9.
[89] Zemedikun DT, Gray LJ, Khunti K, Davies MJ, Dhalwani NN. Patterns of multimorbidity in middle-aged and older adults: an analysis of the UK Biobank data. Mayo Clin Proc 2018;93:857–66.
[90] Faruqui SHA, Alaeddini A, Jaramillo CA, Potter JS, Pugh MJ. Mining patterns of comorbidity evolution in patients with multiple chronic conditions using unsupervised multi-level temporal Bayesian network. PloS One 2018;13.
[91] Sahni N, Simon G, Arora R. Development and validation of machine learning models for prediction of 1-year mortality utilizing electronic medical record data available at the end of hospitalization in multicondition patients: a proof-of-concept study. J Gen Intern Med 2018;33:921–8.
[92] Alaeddini A, Jaramillo CA, Faruqui SHA, Pugh MJ. Mining major transitions of chronic conditions in patients with multiple chronic conditions. Methods Inf Med 2017;56:391–400.
[93] Sun W, Shen W, Li X, Cao F, Ni Y, Liu H, et al. Mining information dependency in outpatient encounters for chronic disease care. Stud Health Technol Inform 2013; 192:278–82.
[94] Brisimi TS, Xu T, Wang T, Dai W, Adams WG, Paschalidis IC. Predicting chronic disease hospitalizations from electronic health records: an interpretable classification approach. Proc IEEE 2018;106:690–707.
[95] Ben-Assuli O, Padman R. Trajectories of repeated readmissions of chronic disease patients: risk stratification, profiling, and prediction. MIS Q 2020;44.
[96] Yao L, Xue-xue D, Xiao-ru X, Rong-hua F. Research on the establishment of a risk prediction model for multiple chronic diseases in the elderly based on big data. Chin J Gen Pract 2021;19:1979–82.
[97] Nenova Z, Shang J. Personalized chronic disease follow-up appointments: risk-stratified care through big data. Prod Oper Manag 2022;31(2):583–606.
[98] Sawhney S, Tan Z, Black C, Marks A, McLernon DJ, Ronksley P, et al. Validation of risk prediction models to inform clinical decisions after acute kidney injury. Am J Kidney Dis 2021;78:28–37.
[99] Marazza F, Bukhsh FA, Geerdink J, Vijlbrief O, Pathak S, Keulen MV, et al. Automatic process comparison for subpopulations: application in cancer care. Int J Environ Res Public Health 2020;17.
[100] Huang Z, Ge Z, Dong W, He K, Duan H. Probabilistic modeling personalized treatment pathways using electronic health records. J Biomed Inform 2018;86: 33–48.
[101] Liu Y. Random forest algorithm in big data environment. Comput Model New Technol 2014;18:147–51.
[102] Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. J Biomed Inform 2002;35:352–9.
[103] Kalcheva N, Todorova M, Marinova G. Naive Bayes Classifier, Decision Tree and AdaBoost Ensemble Algorithm–advantages and disadvantages. Knowl Based Sustain Dev 2020;2020:153.
[104] Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. J Clin Epidemiol 1996;49: 1225–31.
[105] Madhulatha TS. An overview on clustering methods. arXiv preprint arXiv: 12051117. 2012.
[106] Padua, R.D., Carmo, L.P., Rezende, S.O., & Carvalho, V.O. (2018). An Analysis on Community Detection and Clustering Algorithms on the Post-Processing of Association Rules. 2018 International Joint Conference on Neural Networks (IJCNN), 1-7.
[107] Noble WS. What is a support vector machine? Nat Biotechnol 2006;24:1565–7.
[108] Coenen F. Data mining: past, present and future. Knowl Eng Rev 2011;26:25–9.
[109] Corrao S, Natoli G, Nobili A, Mannucci P, Perticone F, Arcoraci V, et al. The "diabetes comorbidome": a different way for health professionals to approach the comorbidity burden of diabetes. Healthcare. 2022;10:1459.
[110] Cousin E, Duncan BB, Stein C, Ong KL, Vos T, Abbafati C, et al. Diabetes mortality and trends before 25 years of age: an analysis of the Global Burden of Disease Study 2019. Lancet Diabetes Endocrinol 2022;10:177–92.
[111] Meyskens Jr FL, Mukhtar H, Rock CL, Cuzick J, Kensler TW, Yang CS, et al. Cancer prevention: obstacles, challenges and the road ahead. J Natl Cancer Inst 2016; 108.
[112] Scully, Crispian M, Stefano Petti. Overview of cancer for the healthcare team: aetiopathogenesis and early diagnosis. Oral oncology 2010;46(6):402–6.
[113] Matthew W, Morgan NEZ, Michael FH. An inconvenient truth: a sustainable healthcare system requires chronic disease prevention and management transformation. HealthcarePapers. 2007;7:6–23.
[114] Ziegler A, König IR. Mining data with random forests: current options for real-world applications. Wiley Interdiscip Rev Data Min Knowl Discov 2014;4:55–63.
[115] Fernández RR, Martín de Diego I, Aceña V, Fernández-Isabel A, Moguerza JM. Random forest explainability using counterfactual sets. Inf Fusion 2020;63: 196–207.
[116] Scheffner I, Gietzelt M, Abeling T, Marschollek M, Gwinner W. Patient survival after kidney transplantation: important role of graft-sustaining factors as determined by predictive modeling using random survival Forest analysis. Transplantation. 2020;104:1095–107.
[117] Sapir-Pichhadze R, Kaplan B. Seeing the forest for the trees: random forest models for predicting survival in kidney transplant recipients. Transplantation. 2020;104: 905–6.
[118] Fernandez-Llatas C. Interactive process mining in healthcare. Springer; 2021.
[119] Van Der Aalst W. Process mining: discovery, conformance and enhancement of business processes. Springer; 2011.
[120] Evans DB, Etienne C. Health systems financing and the path to universal coverage. Bull World Health Organ 2010;88:402.

[121] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. BMC Med 2015;13:1.
[122] von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet. 2007;370:1453–7.
[123] Weissgerber TL, Garovic VD, Milin-Lazovic JS, Winham SJ, Obradovic Z, Trzeciakowski JP, et al. Reinventing biostatistics education for basic scientists. PLoS Biol 2016;14:e1002430.
[124] Fernández-Llatas C, Meneu T, Traver V, Benedi J-M. Applying evidence-based medicine in telehealth: an interactive pattern recognition approximation. Int J Environ Res Public Health 2013;10:5671–82.