Public Health and Informatics
J. Mantas et al. (Eds.)

© 2021 European Federation for Medical Informatics (EFMI) and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/SHTI210200

Process Mining of Disease Trajectories: A Literature Review

Guntur P KUSUMA[a,b,1], Angelina P KURNIATI[c], Eric ROJAS[d], Ciarán D MCINERNEY[a], Chris P GALE[e], and Owen A JOHNSON[a]
[a] *School of Computing, University of Leeds, Leeds, UK*
[b] *School of Applied Science, Telkom University, Bandung, Indonesia*
[c] *School of Computing, Telkom University, Bandung, Indonesia*
[d] *Computer Science Department, School of Engineering, Pontificia Universidad Católica de Chile, Chile*
[e] *Leeds Institute of Cardiovascular and Metabolic Medicine, University of Leeds, Leeds, UK*

Abstract. Disease trajectories model patterns of disease over time and can be mined by extracting diagnosis codes from electronic health records (EHR). Process mining provides a mature set of methods and tools that has been used to mine care pathways using event data from EHRs and could be applied to disease trajectories. This paper presents a literature review on process mining related to mining disease trajectories using EHRs. Our review identified 156 papers of potential interest but only four papers which directly applied process mining to disease trajectory modelling. These four papers are presented in detail covering data source, size, selection criteria, selections of the process mining algorithms, trajectory definition strategies, model visualisations, and the methods of evaluation. The literature review lays the foundations for further research leveraging the established benefits of process mining for the emerging data mining of disease trajectories.

Keywords. Disease Trajectories, Process Mining, Electronic Health Records

1. Introduction

Electronic health records (EHR) are now well established as the basis for delivering modern healthcare [1]. The routine clinical use of EHRs supports predictive, preventive, personalised and participatory (known as "P4") systems medicine approaches [2, 3] and, more recently learning health systems [4]. The increasing volume of data within EHRs is of interest to medical informatics researchers for analysis and, in this paper, we focus on one area of interest - process mining of disease trajectories.

Disease trajectories model patterns of disease over time and can be mined by extracting diagnosis codes from EHR [5]. These models can help researchers and clinicians gain a better understanding of disease progression and how different diseases inter-relate with each other during deteriorating health. Process mining provides a collection of methods and tools that has been used extensively to extract clinical insights from EHRs by mining process models from event data [6]. Process mining

[1] Corresponding Author, Guntur P Kusuma, 7.26 E C Stoner Building, University of Leeds, LS2 9JT Leeds, United Kingdom; E-mail: scgpk@leeds.ac.uk.

provides a mature set of methods and tools that has been used to mine care pathways using event data from EHRs and, we believe, could be applied to disease trajectories. The objectives of this paper are: (1) to review the literature of process mining for identifying disease trajectories, (2) summarise data sources, data selections, methods and limitations, and (3) recognise the challenge of implementing process mining to identify disease trajectories.

2. Method

The first author firstly developed the search strategy defined by the research question "Can a process mining approach be used to identify disease trajectory(ies)?" The two facets of the search strategy were "process mining" and "disease trajectory(ies)". A concept map [7] is used to summarise the synonyms during the literature search.

The literature searching was conducted by the first author on 5th November 2020 following a similar approach as [6], using Google Scholar, PubMed, and dblp databases. Our search syntax was: *"("process mining" OR "workflow mining") AND ("disease trajectory" OR "disease network " OR "disease progression" OR "disease course" OR "illness trajectory" OR "diagnostic trajectory" OR "diagnosis trajectory")"*. We carefully applied the search criteria across databases, adjusting for differences in implementing the criteria from one database to another. For example, Google Scholar automatically searches the plural form of a singular term or vice versa, while we used both singular and plural terms in the two other databases. We also looked for relevant literature in the process mining website at http://www.processmining.org, but none of them met our search criteria seen in Figure 1.

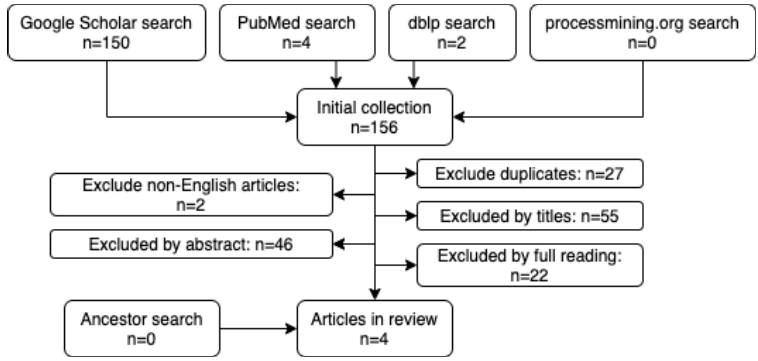

Figure 1. The article selection flow, following [8]

Figure 1 shows the flow diagram of articles included and excluded. From an initial set of 156 articles meeting our criteria, we excluded 27 duplicates and two non-English articles. The remaining articles were filtered for exclusion based on title (n=55), abstract (n=46), and by full reading (n=22). Any process mining articles which only mention the term "disease trajectory" or similar were excluded. The reference lists of the four remaining articles were reviewed against our criteria, but no new articles were included.

3. Results

Table 1 summarizes the four articles identified at the conclusion of our literature search [9-12]. The four papers use four different data sources, four different model visualizations, three different trajectory approaches and three different conformance checking metrics.

Table 1. Reviewed articles of process mining in disease trajectories.

#	Authors	Country/ Region	Data source	N data	Standard coding	Disease	PM Methodology	Model visualisation	Discovery algorithm	Trajectory approach	Conformance checking
1	Kusuma et al.[9]	Boston, USA	BIDMC Hospital	46520	ICD-9	General	PM^2	Directly-followed graph	iDHM	correlation measurement, binomial test	Replay fitness, precision, generalisation, k-folds cross validation
2	de Toledo et al.[10]	Spain	MBDS by the public healthcare provider	225,000	ICD-9	Type 2 Diabetes	KDD	Heuristics net	Heuristics miner and Fuzzy miner	n-grams	N/A
3	De Oliveira et al.[11]	England	NHS Hospital Episode Statistic	76,523	ICD-10	Sepsis	N/A	Private company app	Metaheuristics optimization algorithm	Metaheuristics optimisation algorithm	Replay fitness
4	Kusuma et al.[12]	N/A	(synthetic)	50	ICD-10	General	PM^2	Disco	iDHM	N/A	Replay fitness, precision, generalisation

Four data sources were identified from the reviewed articles: (1) The MIMIC-III dataset of the Beth-Israel Deaconess Medical Center (BIDMC) hospital in Boston, USA (n=46,520) which is available as open access [9, 13]; (2) a healthcare dataset from a suburban area in Spain (n=225,000) [10]; (3) England's national Hospital Episode Statistic (HES) dataset (n=76,523) [11]; (4) a synthetic EHR dataset used to demonstrate the concept of process mining to identify disease trajectories (n=50) [12]. All articles used the International Classification of Disease (ICD) codes published by the World Health Organisation (WHO) [14], either the 9th revision (ICD-9) [9, 10] or the 10th revision (ICD-10) which was used by [11] and [12].

Patient selection based on a specific diagnosis was performed by de Toledo et al. [10] with Type-2 Diabetes and by De Oliveira et al. [11] with a focus on patients with sepsis. In contrast to these, Kusuma et al. [9] did not pre-select diseases of interest but did excluded those ICD codes that do not relate directly to a diagnosis.

Three approaches were identified to define the order of the diagnostic codes to form a trajectory: (1) a set of statistical analysis of correlation measurement and a binomial test [9]; (2) n-grams [10]; (3) metaheuristics optimisation algorithm [11]. The implementation of process mining in the study of disease trajectories was found in three ways: process discovery to build the disease trajectory model, conformance checking, and for model visualisation. The *Fuzzy Miner* and *Heuristics Miner algorithms* [15] were used by de Toledo et al. [10] for mining the event traces and for visualise the model respectively. A proprietary *Metaheuristics optimisation algorithm* [16] was used by De Oliveira et al. [11] while a private-company-developed app was used to visualise the model. Kusuma et al. [9] used the *interactive Data-aware Heuristics Miner* [15] to mine the model, and the directly-followed graph for model visualisation.

There are two articles which reported the use of conformance checking. Kusuma et al. [9] applied replay fitness, precision, generalisation, and k-fold cross-validation for further evaluation, while the metaheuristics optimisation algorithm in De Oliveira et al. [11] is embedded with the replay fitness. Data clustering was used in two studies: de

Toledo et al. [10] applied k-means, agglomerative hierarchical (AHC), and model-based clustering (MBC); and De Oliveira et al. [11] applied the clinical classification software (CCS) for ICD-10-PCS which grouped the diagnostic codes into 220 coded-events [17] to reduce complexity. One study followed a Knowledge Discovery in Database (KDD) method [10, 18] and only one study followed PM^2, the most widely used process mining methodology [9, 19].

4. Discussion

Our review reveals there is currently very little research in use of process mining for identifying disease trajectories. We found that there has been no study which identifies general disease trajectories using national level EHR data. One potential use of process mining is to check the conformance of a model against the original or new data sources. Conformance checking is essential to confirm the representativeness of the disease trajectory model to the data but this was not undertaken in [10]. We recommend that models of how diseases progress over time should be evaluated using conformance checking [20] and clinical review against established medical knowledge.

The study of mining disease trajectories using process mining should build on medical domain knowledge, medical informatics and process mining or data science in general. It is relevant to involve experts in these domains from the beginning of the study because research insights are dependent on the entire research process, and are likely to be more robust when embedded within a multidisciplinary approach. The length of time covered by a data extract may give insights on how diseases are progressing across seasons, years, and the relationship with age and aging. A nationwide disease trajectory model may help policymakers decide where to focus medical informatics research. The opportunity to comprehend patterns of disease progression over time may help clinicians design better interventions for patients targeted at preventing otherwise likely disease progression.

5. Conclusions

The use of process mining for identifying disease trajectories has significant potential as an emerging area of study. This paper provides a summary of current work process mining disease trajectories with respect to data sources, size and the coverage of the data, methods of identifying the trajectory between two diagnostic codes, discovery algorithms, and the visualisations. The small number of studies using formal process mining methodologies highlights a lack of awareness of these methods which we hope this paper addresses.

Using process mining, disease trajectories can be created for both a specific cohort of patients with a specific disease or to discover population level patterns of disease using larger EHR datasets from healthcare providers, regional and national sources. Further comparative studies between trajectories from different countries would provide a better insight on the progression of diseases globally. To date there are only four papers identified with work in this area of study and our conclusion is that there is a great opportunity to develop both the method and case studies using process mining for disease trajectories. There is considerable scope to explore the potential value of these approaches to advancing medical informatics.

Acknowledgements

This study was supported by the UK National Institute for Health Research (NIHR) Yorkshire and Humber Patient Safety Translational Research Centre (NIHR YH PSTRC) and the Indonesia Endowment Fund for Education (LPDP).

References

[1] Campanella P, Lovato E, Marone C, Fallacara L, Mancuso A, Ricciardi W, et al. The impact of electronic health records on healthcare quality: a systematic review and meta-analysis. Eur J Public Health. 2015;26(1):60-4.
[2] Flores M, Glusman G, Brogaard K, Price ND, Hood L. P4 medicine: how systems medicine will transform the healthcare sector and society. Personalized Medicine. 2013;10(6):565-76.
[3] Hood L, Friend SH. Predictive, personalized, preventive, participatory (P4) cancer medicine. Nature Reviews Clinical Oncology. 2011;8(3):184-7.
[4] McLachlan S, Potts HWW, Dube K, Buchanan D, Lean S, Gallagher T, et al. The Heimdall Framework for Supporting Characterisation of Learning Health Systems. J Innov Health Inform. 2018;2058-4563.
[5] Jensen AB, Moseley PL, Oprea TI, Ellesøe SG, Eriksson R, Schmock H, et al. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. Nature Comm. May 2014;5:1-10.
[6] Rojas E, Munoz-Gama J, Sepulveda M, Capurro D. Process mining in healthcare: A literature review. Journal of Biomedical Informatics. April 2016;61:224-36.
[7] Novak J, Cañas A. The Theory Underlying Concept Maps and How to Construct Them. Florida Institute for Human and Machine Cognition. 2006;1.
[8] Williams R, Rojas E, Peek N, Johnson OA, editors. Process mining in primary care: A literature review. Studies in health technology and informatics; 2018.
[9] Kusuma G, Kurniati A, McInerney CD, Hall M, Gale CP, Johnson O, editors. Process Mining of Disease Trajectories in MIMIC-III: A Case Study. LNBIP 2nd International Conference on Process Mining (ICPM 2020). Virtual conference managed by the University of Padua: Springer, Verlag; 2020.
[10] de Toledo P, Joppien C, Sesmero MP, Drews P, editors. Mining Disease Courses across Organizations: A Methodology Based on Process Mining of Diagnosis Events Datasets. 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC): IEEE; 2019.
[11] De Oliveira H, Prodel M, Lamarsalle L, Inada-Kim M, Ajayi K, Wilkins J, et al. "Bow-tie" optimal pathway discovery analysis of sepsis hospital admissions using the Hospital Episode Statistics database in England. JAMIA Open. 2020;3(3):439-48.
[12] Kusuma G, Sykes S, McInerney C, Johnson O. Process Mining of Disease Trajectories: A Feasibility Study. 13th International Conference on Health Informatics.2020. p. 705-12.
[13] Johnson AEW, Pollard TJ, Shen L, Lehman LWH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Scientific Data. 2016;3:160035.
[14] World Health Organization. International statistical classification of diseases and related health problems. - 10th revision. Fifth ed: WHO Press; 2016.
[15] Mannhardt F, De Leoni M, Reijers HA, editors. Heuristic mining revamped: An interactive, data-Aware, and conformance-Aware miner. BPM 2017: CEUR-WS.org.
[16] Prodel M, Augusto V, Jouaneton B, Lamarsalle L, Xie X. Optimal Process Mining for Large and Complex Event Logs. IEEE Transactions on Automation Science and Engineering. 2018;15(3):1309-25.
[17] The Healthcare Cost and Utilization Project. Clinical Classifications Software (CCS) for ICD-10-PCS: Agency for Healthcare Research and Quality, Rockville, MD, Available at: https://www.hcup-us.ahrq.gov/toolssoftware/ccs10/ccs10.jsp.
[18] Han J, Kamber M, Pei J. Data mining concepts and techniques third edition. The Morgan Kaufmann Series in Data Management Systems (Selected Titles) ed: Morgan Kaufmann; 2011. p. 83-124.
[19] van Eck ML, Lu X, Leemans SJJ, van Der Aalst WMP. PM2: A process mining project methodology. In: Zdravkovic J, Kirikova M, Johannesson P, editors. 9097: Springer, Cham; 2015. p. 297-313.
[20] van der Aalst WMP. Process Mining: Data Science in Action. 2 ed: Springer-Verlag Berlin Heidelberg; 2016. p. 467.