P. de Toledo, C. Joppien, M. P. Sesmero and P. Drews, "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), Berlin, Germany, 2019, pp. 354-357, doi: 10.1109/EMBC.2019.8857149.

DOI: 10.1109/EMBC.2019.8857149

Mining Disease Courses across Organizations: A Methodology Based on Process Mining of Diagnosis Events Datasets

Paula de Toledo, Carolin Joppien, Maria Paz Sesmero, Paul Drews

Abstract— **This work proposes the use of Process Mining methodologies on healthcare datasets containing diagnosis information as a means to identify the course of a disease across organizations. Datasets containing diagnosis information for administrative purposes are a good candidate due to its standardized format, widespread availability and coverage. We present a methodology to preprocess, cluster and mine diagnosis information and the results of a preliminary use case with diabetes type II. Some meaningful disease courses have been found but less useful patterns do also emerge. Future work involves lowering the level of granularity chosen (ICD three digit codes) and extending the time span of the data available (three years).**

I. Introduction

Process Mining is a rather young technique that lies between Data Science and Process Science [1]. The goal of Process Mining is to extract knowledge from event logs to discover, supervise and enhance real processes. Process Mining is applicable to any system that records real world events and helps organizations to observe, redefine and rationalize their processes. Furthermore, Process Mining can be used to analyze time-related patterns such as time between two events.

Electronic Healthcare Records (EHRs) are becoming more and more intertwined with the operational processes they assist [1]. The data stored contain information about patient diagnoses and treatments as well as associated care processes. Healthcare organizations are increasingly acknowledging the value of these data and exploiting them to improve their outcomes [2]. Process Mining can offer new perceptions about workflow processes, clinical pathways and compliance with medical guidelines. It can also be used to improve care processes and reduce costs [2].

We propose to use process mining methodologies not to discover processes but common disease patterns in clinical datasets. Specifically we put forward the use of healthcare data registered for administrative purposes as the basis for mining disease course. Although EHRs are widespread nowadays in developed countries, the way information is recorded is far from standard, as it strongly depends on the healthcare provider and the department within the organization. In parallel to the EHR many organizations keep what is called a Minimum Basic Data Set: for each encounter of a patient with the healthcare provider this dataset records basic information (date, demographics, diagnostics, and sometimes prescriptions) [10]. This minimum basic information is fairly standard along providers and care settings and we consider it an excellent start point for common disease patterns discovery due to its direct availability and coverage (it records all possible encounters with the healthcare provider).

II. Process Mining in Healthcare

In recent years, researchers have started to use Process Mining (PM) in healthcare, most commonly to visualize patient pathways in a realistic way. PM helps clinicians and managers [3] to identify typical processes of a patient, the real world counterpart of the clinical guidelines used to reduce variability of treatments and control costs. PM is mainly applied to discover three different types of processes: treatment, organizational and non-elective vs. elective care [4]. In a foundational paper [5] on the application of PM in healthcare published in 2008 the authors demonstrate the applicability to a gynecological oncology process in a Dutch hospital. They analyzed the process from control flow, organizational, and performance perspectives and identified the lack of structure of healthcare environments as the main difference with other application areas. They mention future areas of research should be oriented to methodologies that yield understandable, high-level information instead of too detailed models. A literature review was published in 2016 [4] in which authors examine the current status of PM in healthcare as well as trends and challenges. It shows how PM can identify regular behaviour, process variants, and spot exceptional medical cases. [2] provides a methodology for the application of process analytics on healthcare processes focused on the specific challenges in the healthcare environment. In other recent works authors applied inductive visual miner and heuristic process miner to achieve goals like reducing care costs, and improving processes and quality [6]. There are also attempts to manage and reduce the variability in clinical pathways [7]. To the best of our knowledge there is no published research that tries to identify processes involving different levels of healthcare or across organizations. [8] identifies four challenges for PM in healthcare: highly dynamic processes, highly complex processes, increasingly multi-disciplinary processes, and ad-hoc processes. These and additional issues like completeness, accuracy, complexity and bias are reported in [9], together with the fact that patient data in EHRs tends to be heterogeneous, complicating the analyses and increasing variance.

Paula De Toledo and Maria Paz Sesmero are with the Computer Science Department, Universidad Carlos III de Madrid, Spain (e-mail: mtoledo@inf.uc3m.es).

Carolin Joppien and Paul Drews are with the Institute of Information Systems, Leuphana University of Lüneburg, Germany.

This work was supported in part by grants TRA2015-63708-R and TRA2016-78886-C3-1-R (Spanish Government) and Topus (Madrid Regional Government).

III. MATERIALS

A. The minimum basic dataset

Spanish healthcare providers are legally required to keep a registry the so called minimum basic dataset (MBDS) for every patient encounter at primary, hospital or emergency care. Demographic data as well as information about the main diagnoses motivating the contact and additional diagnoses are stored [10]. Diagnoses are coded with ICD-9-CM (later updated to ICD-10). Although the MBDS data is gathered for administrative purposes and is less specific than a full EHR or a Departmental Information System, it has several features that make it useful towards constructing a methodology to mine the course of a disease: its span to the complete population, generalizability to all patients regardless of the disease, and direct access to temporal information, as Departmental Systems and general purpose EHRs sometimes require complex pre-processing to generate time-labeled diagnostic data.

This work is based on the data collected for the MBDS by the public healthcare provider of a suburban area to the south of Madrid in central Spain with a population of 225,000 people, encompassing different levels of healthcare: one hospital, nine primary healthcare centers and emergency care. Each encounter record holds patient demographic information, date, and care setting (primary care, hospital care, pharmacy or emergency) as well as codified diagnostic information (up to 15 diagnoses per encounter). The dataset covers a period of three years.

B. Process Mining tools

The ProM framework has become the de facto standard for Process Mining [12]. The framework consists of an extensible plug-in tool that was developed at TU Eindhoven [4] [11]. PM uses a variety of notations to model processes. The most widely used is the Petri net modeling language, but others are also in use such as Business Process Modeling Notation, Event-Driven Process Chains and Unified Modeling Language. Modeling languages depict events in terms of activities (a step in the process). Activities are related to a particular case (a process instance) and per case a trace of events can be recorded [1][3].

IV. METHODOLOGY FOR DISEASE COURSE MINING

The application of Process Mining to disease course brings along some challenges. Firstly, courses of diseases are assumed to be highly complex. This complexity leads to unstructured so-called spaghetti-process models that can’t be visualized in a comprehensible manner. Secondly, in comparison to mining patient pathways, the main difference is the lack of de jure reference models such as clinical guidelines. Hence, the methodology must be applicable on the sole basis of the data set available. As data sets such as the MBDS do not contain data directly related to processes, the underlying data-driven approach is a new field of use for Process Mining.

A. Case selection and Data transformation

The methodology proposed is based on the standard Knowledge Discovery in Databases model (KDD) and on previous studies [2][13][14][15]. Similarly to the KDD, it is essentially an iterative process and it shares the first two steps: Data Selection and Data Transformation, which consists of the transformation of event data from information systems into one event log [8].

B. Pre processing

The third step in our methodology is pre-processing of event logs and the generation of event traces ready for mining. It is based on data filtering and abstraction to reduce dimensionality of data, remove unrelated details and focus on the most significant aspects. Pre-processing is an essential step for unstructured processes such as disease course. The outcome of this step are events traces, which, in this specific application, are lists of diagnoses codes assigned to a patient, in a temporal order. The proposed steps are:

3.1. Case selection: healthcare data sets may contain data for a large spectrum of patients and sometimes a case is chosen such as patients suffering from the same or related diseases.

3.2. Dimensionality reduction: The dimensionality is reduced by grouping related activities or by lowering granularity. Diagnosis and prescription coding systems are usually hierarchical and dimensionality can be reduced by selecting a higher hierarchy level. For ICD-9-CM, code 250 corresponds to diagnosis “Diabetes mellitus”, whereas the four-digit code 250.0 corresponds to “Diabetes mellitus without mention of complication” and the five-digit code specifies the diabetes type. The level of granularity strongly depends on the scope of the research and conditions the outcome.

3.3. Event support: This step involves removing rare events that are irrelevant for discovering typical behaviour. This may also be the case with too common events. This involves computing and analysing the support of events and removing them according to the goals [13].

3.4. Repetitions: a common approach in PM is to identify repeating events and plot them as a loop. A better approach for our purpose is keeping only one instance in a series of repeating diagnoses.

3.5. Noise: all events that are not related to the case selected, outliers, errors, and those not relevant for other reasons are considered as noise and filtered out.

3.6. Sequence length: After the pre-processing steps, some event traces may contain only a few events or even a single one and therefore are of little relevance for the goals. Similarly, some traces could be extremely long and bring undesirable effects in the analysis results. Trace length can be limited to a certain range according to the objective [13].

C. Clustering

We propose the use of clustering techniques to divide the event log into sub-logs containing groups of patients that have similar characteristics. PM techniques perform poorly when used on datasets such as ours, due to the heterogeneity of the cases. Our goal is to divide the logs into subsets that are more homogeneous prior to the mining task itself [14]. An iterative clustering approach is proposed as the traces in healthcare records appear to be highly unique and a single clustering iteration does not always lead to satisfactory results. To select the optimal number of clusters we used both direct methods that aim to optimize a criterion such as average silhouette, and statistical testing methods that compare evidence against null hypothesis [16].

1) Event traces and distance metrics

The result of the pre-processing step is an event trace that has to be converted into a representation supporting the calculation of distances among traces for clustering. The distance metric used is a key feature defining the outcome. Two main representations are used: vector based and syntactic. In a vector based approach an event trace is represented as a vector where each dimension corresponds to a diagnosis in the event log. Binary vector retains only the absence or presence of the diagnosis whereas numeric representation holds the frequency count of occurrences of the diagnosis in the event trace. These representations can be seen as bag-of-activities feature set [17] as information about the order of occurrence of events is disregarded. To incorporate order of occurrence into the vector based representation, n-grams are used. An n-gram is a sub-sequence of n events, i.e. trace A→B→C represented as two-grams would be {AB, BC}. The size of the n-gram model increases drastically with the size of n and number of events, generating a vast computational overhead. The value of n must be a trade-off between space complexity and accuracy of representation. In all vector based approaches common distance metrics such as Euclidean distance are used.

In the syntactic approach, cases are viewed as traces: a trace in an event log corresponds to a sequence of events executed in a process instance and retain the order of execution. The traces are considered in totality for clustering and the distances between sequences are defined in terms of error transformations, such as the Hamming (next to edit) and Levenshtein (edit) distances.

2) Clustering methods

Given an event log containing a set of traces generated with vector based or syntactic methods and an appropriate distance/similarity function, the next step is clustering the event log into sub logs. We propose the use of commonly used algorithms such as K-means clustering, hierarchical clustering [18] and model-based clustering (MBC) [19].

D. Data Mining and Evaluation

The actual Process Mining is performed on each of the clusters derived. The mining techniques used are heuristics miner and fuzzy miner, in their ProM implementation described before. These techniques have proven their applicability b [4]. The heuristics miner is an improvement of the alpha miner (α-algorithm). The alpha miner process models results are often unstructured and not sound. The heuristics miner takes frequencies into account, filtering out noisy or infrequent behaviours. Additionally, it is able to detect short loops and skip single activities. On the other hand, the fuzzy miner presents process models as process graphs and is used by commercial tools because of its high practical value as is able to cluster events and provides a more comprehensive process model on a higher abstraction level. It reveals patterns such as relations between activities and information about preceding and following activities [20] but process graphs do not make evident if a sequence indicates choice or parallelism [21]. This technique is used in this work because of its visual and comprehensible outcome.

Evaluation is the final part of the process: knowledge elicited during the analysis is presented to medical experts for evaluation. This feedback usually results in iterations of previous phases to refine the analysis.

V. RESULTS

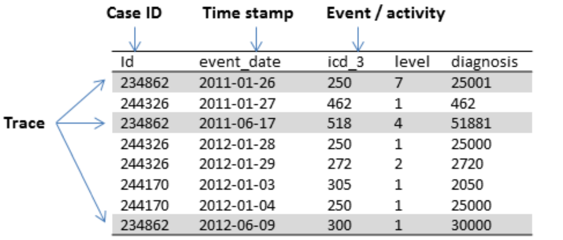
As a case study to test the methodology, we focused on Diabetes type II patients, due to the high prevalence of the disease, high number of co-morbidities and economic impact. All patients with any diagnostic of Diabetes Type II at any point in the three years analysed where included in the study. Table I describes the dataset for both the complete and the diabetes population. Diagnoses that are documented most frequently besides diabetes are “unspecified essential hypertension” (ICD9-4019), “unspecified effects of heat and light” (9929), “late effect of traumatic amputation” (9059), and unspecified cataract (3669). All of them are well known side effects or conditions associated to diabetes.

TABLE I. DESCRIPTION OF THE DATASET

	MBDS	Diabetes TII
Number of patients	235,460	7,023
Average Patient age	41	67
Number of records	4,500,000	440,197
Number of different diagnoses	9,330	2,654
Average records per patient [max-min]	13 [1-488]	32 [1-488]

A. Pre-processing

The main design decisions during pre-processing were to select the level of granularity of the disease codes at three digits, based on the hierarchical structure of ICD-9. We removed diagnoses that were considered less relevant for our analysis (“740-759: Congenital Anomalies” and “800-999: Injury and Poisoning”). We also removed all diagnoses that affected less than ten per cent of patients.


Case ID
Time stamp
Event / activity
Trace

Id	event_date	icd_3	level	diagnosis
234862	2011-01-26	250	7	25001
244326	2011-01-27	462	1	462
234862	2011-06-17	518	4	51881
244326	2012-01-28	250	1	25000
244326	2012-01-29	272	2	2720
244170	2012-01-03	305	1	2050
244170	2012-01-04	250	1	25000
234862	2012-06-09	300	1	30000

Figure 1 – Extract from event log after abstraction and selection steps showing how a trace is made up of events from a patient

B. Clustering

In the next step we applied clustering techniques to divide the event log into sub logs. Vector based and syntactic representations were used, with the later generating very heterogeneous traces not adequate for the subsequent mining steps. We only report here results of the vector based approach. Different clustering approaches - k-means, agglomerative hierarchical (AHC) and model based (MBC) - are applied and results compared to select the most suitable method. Three different representations were used: simple events (ICD codes), 2 grams and 3 grams. We found that 2 and 3 grams (Table II) yielded more relevant results. As an example of the order preservation in n-grams, we see that sequence 250 (Diabetes) → 401 (Hypertension) occurs with a much higher frequency than 401 → 250. After building the n-grams and transferring the traces into a feature space, clustering is performed. Best results are achieved using AHC algorithm with Ward's clustering method both for 2 and 3-grams. 9 clusters are selected in the 2-gram model, with three clusters having an average silhouette (ASW) of 1. For 3-grams clustering further iteration is not required as one step provides good results. 5 clusters are identified, four of them with ASW =1.

TABLE II. MOST FREQUENT 2-GRAMS AND 3 GRAMS

2-gram	Frequency	3-gram	Frequency
250_272	1646	250_272_401	377
250_401	1406	250_272_278	256
401_250	968	250_272_305	167
719_250	507	250_272_366	96
272_401	479	250_278_401	93
250_278	429	250_366_401	93
724_250	398	401_250_272	81

C. *Data Mining and evaluation*

Fuzzy miner is used to mine the event traces within each cluster. Heuristics miner is then used to visualize the event logs as a process model. The resulting heuristics net for cluster 2 is shown in Figure 2.

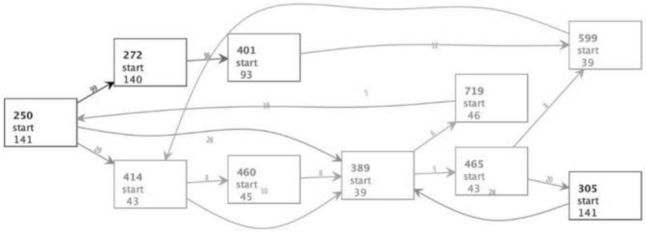


Figure 2 – Heuristics net of cluster 2 containing 141 traces

The heuristics net highlights frequent events with a darker colour and provides information about the frequencies of the flows between events. In this cluster the predominant sequence is: Diabetes mellitus → Disorders of lipoid metabolism → Essential hypertension, which is meaningful from the clinical point of view. In other cluster the most prevalent sequence is apparently meaningless (Cataract → Diabetes mellitus → Essential hypertension) showing the need for a final evaluation step to interpret disease courses. Iterations to remove frequent but meaningless diagnosis could be useful as well if carefully considered.

We found that the representation option to select depends on the objective: when the goal is to find patient groups with similar events, simple single ICD codes are preferred as this clusters patients based on their diagnoses an thereby reduces the variety. If the goal is to find similar pathways concerning the order of diagnoses, n-grams are to be used as they are a proficient trade-off to cluster traces as it keeps order information but less infrequent subsequence are removed.

VI. CONCLUSION

This work is a preliminary step towards using process mining to identify common courses of disease in healthcare datasets containing diagnoses information. The results show that meaningful patterns are identified but also less useful and contradictory information emerges from the data. Further research will be directed towards better defining the level of granularity for a diagnosis that yields useful results, fine tune the size of the n-grams and find ways to use syntactic distance among traces that keep variability low. We identified that the time span of data (three years) is too short to properly mine courses of diseases. To the best of our knowledge this work is the first attempt to mine healthcare information gathered from different levels of care, which is one of the main values of the dataset used. While the dataset is cross-organizational, widely available and general purpose, it might be too unspecific for mining disease courses, as it does not hold enough expressive capacity to represent aspects such as confirmed vs. suspected diagnosis and it would be useful to extend the methodology to more complex EHR systems. On practical grounds the value of the methodology would is to be the basis of a tool to support clinicians and managers, in interpreting the results without direct support of data scientists.

REFERENCES

[1] W Van der Aalst 2016. Process Mining: Data Science in Action. Springer

[2] J Lismont, AS Janssens et al. 2016. A guide for the application of analytics on healthcare processes: A dynamic view on patient pathways. Computers in Biology and Medicine.

[3] RS Mans, W Van der Aalst, RJB Vanwersch. 2015. Process Mining in Healthcare: Evaluating and Exploiting Operational Healthcare Processes. Springer.

[4] E Rojas, J Munoz-Gama et al. 2016. Process Mining in Healthcare: A Literature Review. ScienceDirect.2016.

[5] RS Mans, MH. Schonenberg, et al. 2008. Application of Process Mining in Healthcare: A case study in a Dutch hospital Communications in Computer and Information Science. Springer.

[6] KM Ganesha, K Soundarya V, Supriya. 2017. Analyzing the waiting time of patients in hospital by applying heuristics process miner. Int Conf on Inventive Communication and Computational Technologies.

[7] Detro S, Pereira E et al. 2017. Managing Business Process variability through process mining and semantic reasoning: An Application in healthcare. In Collaboration in a Data-Rich World. Springer.

[8] A Rebuge DR Ferreira. 2012. Business process analysis in healthcare environments: a methodology based on process mining. IS Information Systems.

[9] G Hripcsak, DJ Albers, A Perotte. 2011. Exploiting time in electronic health record correlations. Journal of the American Medical Informatics Association.

[10] A Sarria-Santamera,. 2013. The Spanish Minimum Basic Data Set. European Journal of Public Health

[11] W Van der Aalst 2014b. Process Mining discovery, conformance and enhancement of business processes. Springer.

[12] RS Mans, W van der Aalst et al. 2013. Process Mining in healthcare: data challenges when answering frequently posed questions. LNCS:

[13] Veiga GM, DR. Ferreira. 2009. Understanding spaghetti models with sequence clustering for ProM. In Business Process Management Workshops. Lecture Notes in Business Information Processing.

[14] Bose RP. J Chandra. 2012. Process Mining in the large: preprocessing, discovery, and diagnostics.

[15] D Klimov, A Shknevsky. 2015. Exploration of patterns predicting renal damage in patients with diabetes type II. J Am Med Info 22 (2).

[16] Kassambara, Alboukadel. 2017. Practical guide to cluster analysis in R. Unsupervised machine learning. CreateSpace.

[17] G Greco, A Guzzo et al. 2006. Discovering expressive process models by clustering log traces. IEEE Trans. Knowl. Data Eng. 18 (8).

[18] J Leskovec, A Rajaraman, JD Ullman. 2015. Mining of Massive Datasets.

[19] AP Dempster, NM Laird, DB Rubin. 1977. Maximum likelihood from incomplete data via the EM Algorithm. Journal of the Royal Statistical Society. Series B (Methodological) 39 (1).

[20] CW Günther, W Van der Aalst. 2007. Fuzzy mining – adaptive process simplification based on multi-perspective metrics. Business Process Management. Lecture Notes in Computer Science. Springer.

[21] GT Lakshmanan, S Rozsnyai, F Wang. 2013. Investigating clinical care pathways correlated with outcomes. In Business Process Management, 323–38. Lecture Notes in Computer Science. Springer.