Construction of Empirical Care Pathways Process Models From Multiple Real-World Datasets

Juan González-García, Carlos Tellería-Orriols, Francisco Estupiñán-Romero, and Enrique Bernal-Delgado

***Abstract*—Care pathways (CPWs) are "multidisciplinary care plans that detail essential care steps for patients with specific clinical problems." While CPWs impact on health or cost outcomes is vastly studied, an in-depth analysis of the real-world implementation of the CPWs is an area that still remains underexplored. The present work describes how to apply an existing process mining methodology to construct the empirical CPW process models. These process models are a unique piece of information for health services research: for example to evaluate their conformance against the theoretical CPW described on clinical guidelines or to evaluate the impact of the process in health outcomes. To this purpose, this work relies on the design and implementation of a solution that a) synthesizes the expert knowledge on how health care is delivered within and across providers as an activity log, and b) constructs the CPW process model from that activity log using process mining techniques. Unlike previous research based on ad hoc data captures, current approach is built on the linkage of various heterogeneous real-world data (RWD) sets that share a minimum semantic linkage. RWD, defined as secondary use of routinely collected data as opposite to ad hoc data extractions, is a unique source of information for the CPW analysis due to its coverage of the caregiving activities and its wide availability. The viability of the solution is demonstrated by constructing the CPW process model of Code Stroke (Acute Stroke CPW) in the Aragon region (Spain).**

Manuscript received June 5, 2019; revised October 14, 2019 and December 18, 2019; accepted January 25, 2020. Date of publication February 24, 2020; date of current version September 3, 2020. This work was supported in part by the ProCOMP project (LMP183_18) funded by the Aragon Government through the Operative Programme of European Regional Development Fund (ERDF) "Construyendo Europa desde Aragón," and in part by the ICTUSnet project (SOE2/P1/E0623) funded by the Interreg SUDOE Program that supports the development of south-east Europe regions through the ERDF. *(Corresponding author: Juan Gonzalez Garcia.)*

Juan González-García and Carlos Tellería-Orriols are with the Biocomputing Unit and the Data Science in Health Services and Policy Research Group and Health Services Research network on Chronic Patients (REDISSEC), Institute for Health Sciences in Aragon (IACS), 50009 Zaragoza, Spain (e-mail: jgonzalezgarc.iacs@aragon.es; ctelleria@aragon.es).

Francisco Estupiñán-Romero and Enrique Bernal-Delgado are with the Data Science in Health Services and Policy Research and Health Services Research network on Chronic Patients (REDISSEC), Institute for Health Sciences in Aragon (IACS), 50009 Zaragoza, Spain (e-mail: festupinnan.iacs@aragon.es; ebernal.iacs@aragon.es).

Digital Object Identifier 10.1109/JBHI.2020.2971146

***Index Terms*—Real-world data, process mining, data mining, health system research, code stroke, comparative effectiveness research.**

I. Introduction

CARE pathways (CPWs) (also called clinical pathways, integrated care pathways or care maps) are "multidisciplinary care plans that detail essential care steps for patients with specific clinical problems" [1]. As stated in the Cochrane review [2], which included a total of 27 studies of CPW implementation, CPWs "aim to link evidence to practice and optimize clinical outcomes whilst maximising clinical efficiency." This same review concluded that "Care pathways are associated with reduced in-hospital complications and improved documentation without negatively impacting on length of stay and hospital costs."

While CPWs benefits on health or cost outcomes have been studied and properly demonstrated in the Cochrane review [2], an in-depth evaluation of the real-world implementation of the CPWs is currently an active research area. The study of CPW implementation in the health care has evolved from basic before/after comparison of some variables of interest of works such as [3], to a more complex analysis to answer questions such as which is the adequacy of the real-world implementation of the CPW with respect to the theoretical or normative definition of the CPW, as in [4], or what is the effect of actual CPW implementation in terms of health outcomes for the different patients that traverse the pathway, as in [5]. From the computing and data science community process mining techniques have emerged as an alternative approach to solve those questions.

Process mining is a relatively new research field in computing science that combines unsupervised and supervised data mining methods with business process analysis techniques [6]. The goal of process mining is to discover and study the structural organization of productive processes undertaken in an organisation, namely the business process models.

In the context of health services and policy research, process mining can be used to capture the CPWs' process model applying process discovery techniques from the logs available in their information systems. This empirical CPW process model, as it comes from RWD, can be later used to answer the questions suggested previously serving as example: it can be compared with the theoretical model, described in clinical guidelines, using conformance checking techniques; the paths or traces patients followed within the CPW, may be used to enrich a survival analysis to measure the effect of the exposition to the different paths.

The purpose of this paper is to demonstrate how to exploit the use of Real-World Data (RWD) sets for CPW analysis. RWD may be defined formally as "data used for decision-making that are not collected in conventional randomized controlled trials (RCTs)" [7] or, in other words, a secondary use of the administrative routine data collected in the health care systems as opposed to specific ad hoc data extractions. In this paper, RWD sets from a large health system are used within a process mining methodology to evaluate the actual implementation of a complex CPW. The use case selected is the Code Stroke implementation, the CPW for stroke management, in the Aragón Region health system, a public health system that comprises 9 acute hospitals covering a population of 1.3 million insurees.

This paper leverages the Process Mining Methodology (PM^2) described in [8] for the Code Stroke analysis use case. Three different RWD sets coming from three different information systems were processed to apply process discovery. Please, note that the approach presented in the paper can be applied to any other CPW whose activity is properly recorded in the routine data collected and maintained by the health care systems.

II. Related Work

A. Process Mining

Process mining is a relatively new research field in computing science that can be categorized under the umbrella of the data mining methods. Its main objective is to gain insight on how corporations or institutions organize their production, i.e. understand, analyse, and, at the end, improve productive processes. The extensive digitalisation in almost all industries made the case for process mining. In this field, the book "Process Mining: Data Science in Action" [6] by Wil M.P. van der Aalst may be considered the main reference.

In brief, the process mining cornerstone are the activity logs generated by different information systems in an organisation. Activity logs are files, usually in plain text, that contain a sequence of the ordered activities that took place in the productive process. Activity logs are processed to *discover* the process model by means of algorithms such as the α-algorithm [9], the Heuristics Miner [10], the Fuzzy mining [11] or the Inductive Miner [12]. Example of process models are depicted in results section.

As detailed in [6], the output of the process discovery algorithms is a map of the *control flow* perspective of the process. Depending on the extra information available in the activity logs, for example the specific timestamp when the activities took place or the person or resources involved in the activity, a broader insight can be achieved, for example a time-related perspective, a performance perspective or organisational perspective. Even it is not the main aim of the present work, the results section contains examples for control-flow, time-related and resource perspectives.

Pursuing the improvement of the production processes, conformance checking techniques and algorithms conform also an important area of process mining research. Extensively covered in the Munoz-Gama awarded PhD. thesis [13], conformance checking techniques aim to evaluate and quantify if the process models discovered follow the expected behavior and quantify. The differences between the actual and the expected behavior may suppose an opportunity for process improvement. Quantifications may be done from basic measurements of certain parameters of the process models, as the presented in the results section, to more complex ones based on topologies and graph algorithms, such as the ones collected in the Munoz-Gama PhD, or in the work by Rozinat *et al.* in [14].

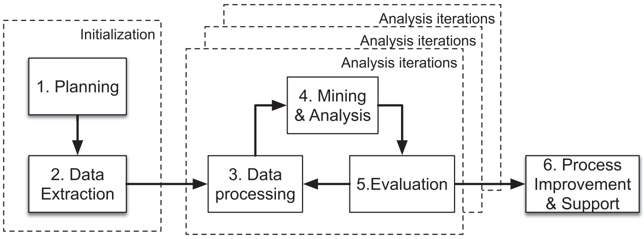

Fig. 1. PM^2 methodology schema adapted from [8].

Almost all algorithms in process mining field are implemented as part of the ProM tool [15], the reference process mining tool developed at the Technical University of Eindhoven. Commercial tools such as Disco [16] have done an effort to approach process mining to businesses by providing a neat interface. In the present paper the tool use for process mining tasks was `bupaR` [17], a R library for process mining. It was selected due to the interaction with all the available packages the R language offers and the possibility to create a reproducible process mining analysis workflows (note that ProM and Disco are principally interactive tools).

Last but not least, the process mining community has worked in giving a systematic approach to process mining projects by designing analysis methodologies. In this aspect, the `L*` life-cycle model [18] is one of the most accepted within process mining works. This paper follows a plus modern synthesis of the $L*$ life-cycle model, namely the PM^2 methodology, described in [8]. The PM^2 methodology schema is depicted in Fig. 1 and is described in Section III.

B. Process Mining in Healthcare and Health Services Research

The use of process mining has attracted the healthcare and, specifically, the health services research community attention since its initial steps. The 2005 work of Mărușter and Jorna in [19] is one of the very first examples of process oriented analysis. In this work authors proposed an initial approach to capture the interaction between "multidisciplinary" patients, i.e. pluri-pathogical patients that require multiple specialty care professionals, with the health system. In fact, this work appeared before the dawn of major Process Mining works and algorithms.

In recent years, up to 3 reviews covered the process mining works related to healthcare: in 2015, Yang and Su analyzed 37 works in [20]; in 2016, Rojas *et al.* produced the most extensive review up to date [21], including 74 works; and, finally, Batista and Solanas reviewed 55 works in 2018 [22].

Among these three reviews, it is interesting to highlight the one by Yang and Su [20], as it explicitly focuses on the application of process mining for CPW. The conclusions of Yang and Su after reviewing 37 works pointed that there is an important limitation in the usability of process mining for CPW analysis, mainly due to the lack of structure in the processes or in the clinical data itself. The review also pointed to limitations when dealing with integrated care as, in general, the data used was limited to reduced data sets.

It is in the data structure and availability context where the current work presents a notable improvement. In the vast majority of previous works (not cited here for space limitations) the data used wasn't previously modelled and usually corresponded to a single hospital information system (HIS) or, in some other cases, ad hoc data gatherings. Few works such as [23] presents a neat data model usable for a broad spectrum of analyses. Same applies for data sets, few works analysed data from multiple sources, for instance in [5] and [24] the authors gathered data from four different hospitals that share the same HIS to evaluate the pathway underwent by patients with chest pain symptoms, or in [25] where authors also gathered an ad hoc data registry in four Italian hospital to analyse stroke CPW; and very few of them linked multiple (RWD) sets, for example in [26], the authors linked the Austrian cancer registry with a treatments database.

The present work is an explicit effort in two facets: first, to define a data model able to capture the analysis requirements to exploit the use case data using process mining; second, to effectively link RWD sets routinely collected in the health system from multiple hospitals and emergency room departments, a challenging task to provide coherence to disjoint points of view. It is interesting to note that current health services research works share this fundamental vision of extensively link and use RWD sets to enable integrated care analysis, such as the study protocol for UK-wide primary care analysis detailed by Litchfield *et al.* in [27].

III. Methods

The contribution of this paper is to construct the empirical CPW process models using RWD sets by applying process mining techniques. To this purpose, the methodology used is an adaptation of the PM^2 methodology. PM^2 is a general methodology for process-mining-based projects proposed by Maikel L. van Eck in [8]. Fig. 1 contains a redraw of methodology schema from the original paper. As can be seen in the Figure, the methodology has 6 major tasks, divided en 3 stages:

- "Initialization" stage: composed by Task 1 "Planning" and Task 2 "Data Extraction."
- "Analysis" stage: composed by Task 3 "Data processing," Task 4 "Mining and Analysis" and Task 5 "Evaluation."
- "Process improvement and support," a single task stage (Task 6).

In order to ease the clarity of the current section it has been organised to align with the PM^2 methodology tasks.

Note that the source code of the implemented solution, which covers mainly the "Analysis" stage tasks, is available in a GitHub repository.[1] Sample datasets are also included to permit a full reproducibility of the approach.

A. Task 1: Planning. The Code Stroke Use Case

The use case selected to illustrate the present work is the construction of the Code Stroke process model in the Aragón region (Spain). Code Stroke is a complex organizational intervention that aims at delivering, in time, the best treatment available for acute stroke patients. Code Stroke provides a clear decision-making algorithm, i.e. a care pathway, privileging access to defined resources and treatments when the "Code Stroke" is *activated* [28]. Stroke represents the second cause of death in general population and first in female population, being also the first case of disability, generating high social costs. Evidence [29] showed that a proper management of stroke patients may lead to a mortality risk reduction as large as 45% (Odds ratio 0.61: CI95% 0.47-0.78).

The aim of the study is to describe the overall pathway for patients with a suspicion of acute stroke and the specific pathway for patients with ischaemic stroke, deepening a bit further in the timely use of fibrinolytic treatment, as a major milestone in this care process. For this particular illustration, we retrieved the information for all suspected cases of an acute stroke episode in 2017. The episode linking algorithm of Tasks 3 determined the actual stroke cases.

This work used retrospective pseudonymised data and was conducted in accordance with the amended Helsinki Declaration, the International Guidelines for Ethical Review of Epidemiological Studies, and Spanish laws on data protection and patients' rights. The data went through a double dissociation process, i.e. in the original data source and once data is stored in the database, impeding patients re-identification.

B. Task 2: Data Extraction Task. Data Model and RWD Sets

1) The Data Model: a basic element in the data extraction task is the process model construction to identify and clarify the semantics of the data and its structure. The data model may limit the possibilities of further analyses or the comparability among different health systems once the process model has been established. To this respect, our data model was a straightforward adaptation of the data model present in the RWD sets used in this work. In different settings, the nonexistence of a clear data model or the necessity of dealing with datasets with different semantics will require an extensive pre-processing work to ensure not just the syntactic interoperability, but also the semantic interoperability.

To design the data model is important to understand an abstract structure of the CPW and its translation into the typical information systems available in the health system. In the present work, the real-world implementation of a CPW is abstracted as a set of episodes. An episode is a sequence of caregiving events related to the treatment of a given patient for an acute

[1] [Online]. Available: https://github.com/IACS-Biocomputing/process-mining-RWD

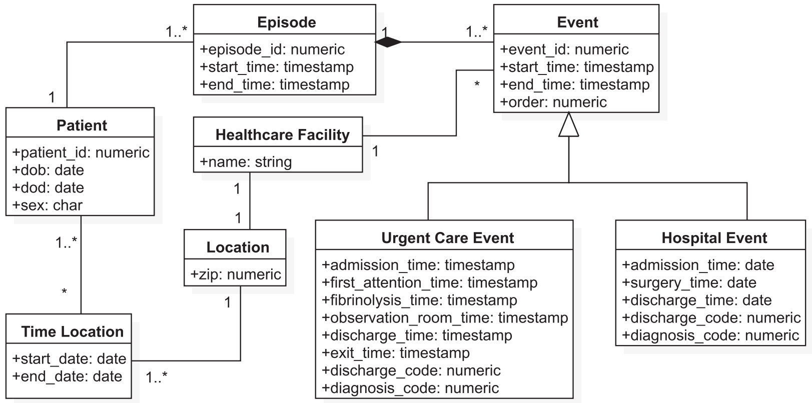

Fig. 2. UML diagram of the code stroke data model.

occurrence of a given illness, considered from an entry point (e.g., admission to an emergency room) to a final discharge (e.g. hospital discharge). The events are each individual record registered in the information system or the database of a given healthcare service (e.g., emergency room information system or a hospital information system). An event includes the series of activities that took place in the specific service for a specific episode. An activity is a timestamped clinical or administrative act that took place in a given service during the caregiving episode of an acute occurrence of a given illness (e.g., the emergency room admission or the hospital surgery).

The data model designed to capture the previous description is depicted in Fig. 2 using a Unified Modeling Language (UML) schema. Three main classes represent the data that will be captured from the RWD sets: "Patient," "Urgent Care Event" and "Hospital Event." Locations and facilities are, in fact, attributes of the "Patients" and "Events" but are promoted to classes for time-related expressivity. The superclass "Event" is used to provide a single interface independently of the type of the event. Note that the two timestamps attributes of the "Event" class correspond to the minimum time (start_time) and maximum time (end_time) observed in the subclasses. In the case of "Hospital Event" the granularity of the time attributes is casted from date to timestamp by adding midnight time.

Finally the "Episode" objects are created during the Task 3, detailed in III-C, by capturing the event continuity among the CPW that took place during a patient treatment.

2) RWD Sets Extraction: the RWD sets used for this study were three: emergency room data warehouse (SUH BI), the hospital discharge database (CMBD), the insurance database (BDU). These three datasets are distributed across the Aragon Health System facilities with different security policies, data models, but with a common identifier available in all of them: the patient's personal identification code (indicated as patient_id).

In a first extraction step the data was captured into a single repository leveraged to a health data platform. This platform is a hardware/software solution to integrate the health system data sources into a single RWD Data Lake and a computing platform to exploit the the RWD Data Lake. The platform integrates the Extract, Transform and Load (ETL) processes that captured the data from the aforementioned and store them in tabular formats.

A second extraction step was required to gather the specific events and patients related to the Code Stroke and create the objects in the data model, creating the objects for all classes except the "Episode." This data gathering is a complex task divided in four data selections: 1) SUH BI emergency room records where diagnosis_code corresponds to one of the ICD-9/ICD-10 stroke codes provided by neurologists and the minimum timestamp among its activities takes place during 2017; 2) CMBD records where variable diagnosis_code corresponds to one of the ICD-9/ICD-10 stroke codes, where the date reflected in the variable admission_time occurs in 2017; 3) records from SUH BI where variable patient id appears into the union of patient id variables from records in two previous selections; 4) records from CMBD where variable patient_id appears into the union of patient_id variables from records in selections 1 and 2; 5) records from BDU where variable patient_id appears into the union of patient_id variables from records in data selection 1 and 2.

The rationale of these selections is the following: first two data selections correspond to the direct detection of suspicious stroke events in ER or hospital facilities; selections 3 and 4 are the events that may reflect side conditions of the pathway but did not contain an explicit stroke diagnosis (e.g., a patient who came to ER with problems with speech and it is confirmed as a stroke in the hospital). Data selections 1 and 3 are merged and codified as "Emergency Room Event" objects, and equivalently done with data captures 2 and 4 to generate "Hospital Event" objects. Finally, data selection 5 is transformed into "Patient" objects.

C. Task 3: Data Processing Task. Episode Linking Algorithm and Activity Log Generation

The episode linking algorithm is the core algorithm of the present work. As stated previously, patient_id is the only common identifier across the datasets and no other information links the continuity of the Code Stroke episodes. The episode linking algorithm is in charge of determine the correctness of the episodes and construct the episode continuity not explicitly recorded in the RWD sets.

The episode linking algorithm is a Python application that runs as follows:

Single event timestamp correctness and correction procedure: Those events with incoherent timestamps are marked as incorrect, e.g. hospital surgery out of bounds of the hospitalization. Events marked as incorrect take part of the whole linking procedure to capture and *consume* the rest of the episode events that will be also considered incorrect avoiding the use of them in other episodes.

A correction procedure processes manually-inserted activity timestamps (e.g., ER Fibrinolysis activity). The correction consists on checking the coherence between year, month and day of the manually-inserted activities and the rest of activity timestamps available in the same event. If there is a discrepancy between the majority of the timestamps and the manually-inserted ones, the manually-inserted are corrected by using the most observed value. This correction may be to the lowest level of detail for ER events (i.e., days) and only for years in Hospital events.

Event bound detection and censoring: Detect the minimum and maximum times in all events and store them in dummy variables for each event (`start_time` and `end_time`). In case the event bounds are out of the study period (initially defined) are marked as left or right censored appropriately.

Global sorting: When two events from same patient occur the same **date**, the Emergency Room precedes the Hospital Event. This definition avoids the timestamp granularity mismatch between types of events. This step generates the `order` variable in the events.

Linking rules execution: Apply chronological and logical linking rules comparing pairs of ordered events of each patient. The chronological linking rules consider the types of events to compare and focus in the time sequence adding a time tolerance between the end of an event and the beginning of the following one. The tolerance depends on the type of events compared. Linking rules complement chronological linking rules ensuring that the `discharge_code` of the first event is compatible with the second event type.

Timestamp synchronization procedure: This procedure adjusts the timestamps across the different events in an episode and transforms the granularity of all timestamps to date plus time (hours, minutes and seconds). Those episodes where there is an overlap between Emergency Room events, the latter activity timestamp of the first event is substituted by the first activity timestamp of the following event. For those episodes that include an Emergency Room event followed by a hospital event, the admission activity timestamp of the hospital event (`admission_time`) is copied from the timestamp of the last activity of the previous Emergency Room event (`end_time`) to include the time granularity. This time part is also added to the rest of activities in the hospital event. When only hospital events take part in an episode, midday time are added to the hospital activities.

Using the use case data for 2017, the episode linking algorithm processed a total of 5802 suspect stroke episodes. Just 3265 were determined, the rest were discarded due to some of the previously commented reasons, for example 268 corresponded to right censored episodes or 319 correspond to any failed linking rules. Once finished the episode linking, the correct episodes are given a unique identifier (`episode_id`) and stored in MongoDB [30]. The storage is done in a patient-centric collection (the term used in MongoDB to refer to group of documents with similar structure and contents). Each document of the collection corresponds to a patient present and includes the stroke episodes described plus the patient socio-economic data captured from the BDU dataset. MongoDB was selected to handle stroke episodes as it is able to manage a variable number of events within an episode and a variable number of episodes for each patient easily within the structure of the patient-centric collection.

The log generation is a transformation process of the patient collection into a new MongoDB collection, in which each document contains the triplet `<episode_id, activity_id, timestamp>`.

The transformation process, currently included in the episode linking script, traverses all the patients in the patient collection. For each patient, the process takes all the correct episodes and then traverses the events that conform the episode. Each event is then decomposed on the activities it contains, and this information is stored in the activity log collection. In the specific case where an episode contains two consecutive hospital related events and the first activity discharge type indicates a discharge to a long stay hospital, the activities of the second event are specifically distinguished as long stay hospital activities. The collection is then sorted by the timestamp, resulting the orthodox organization of a classical log.

D. Task 4: Mining and Analysis. Care Pathways Process Model Generation

The care pathways process model generation leverages process discovery techniques. The tool selected for process model is `bupaR` [17], a R language package that offers an exhaustive variety of process mining analysis techniques.

In a R script, the activity log collection is transformed to an activity log data frame (the term used in R for its data table structure). The activity log data frame is used by `bupaR` to construct the care pathway process model. `bupaR` offers then multiple ways to present and explore the resulting process model. As presented in the following section, process traces, process maps and process timelines are the most useful process model outputs for the purposes of the current work.

E. Task 5: Evaluation

This section contains the outputs of the process mining analysis that were provided to the Aragón Code Stroke Implementers Group, as part of a face-validation approach of the results with domain experts. The Code Stroke Implementers Group is conformed by neurologists, emergency room managers, hospital managers and the Code Stroke strategy directors. The face-validation obtained positive results, confirming the high plausibility of the results with experts' vision of the CPW process. It remains for further iterations of the process mining analysis the inclusion of minor suggestions regarding the merge of some the CPW traces shown below, when the empirical

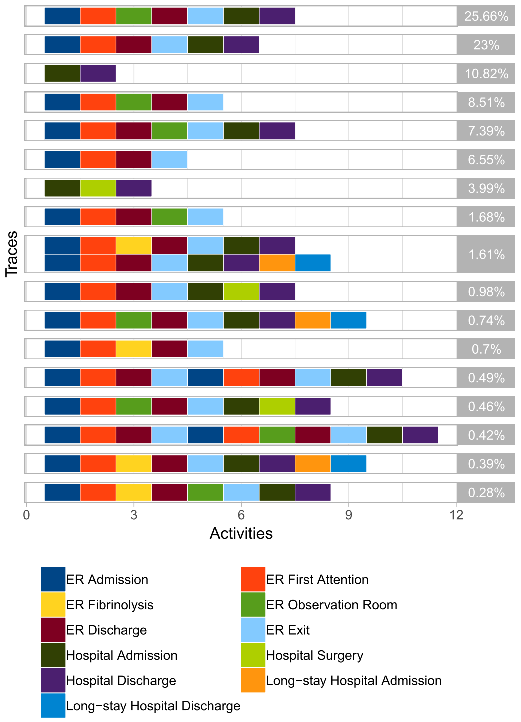

Fig. 3. Process traces covering more than 95% of total stroke episodes.

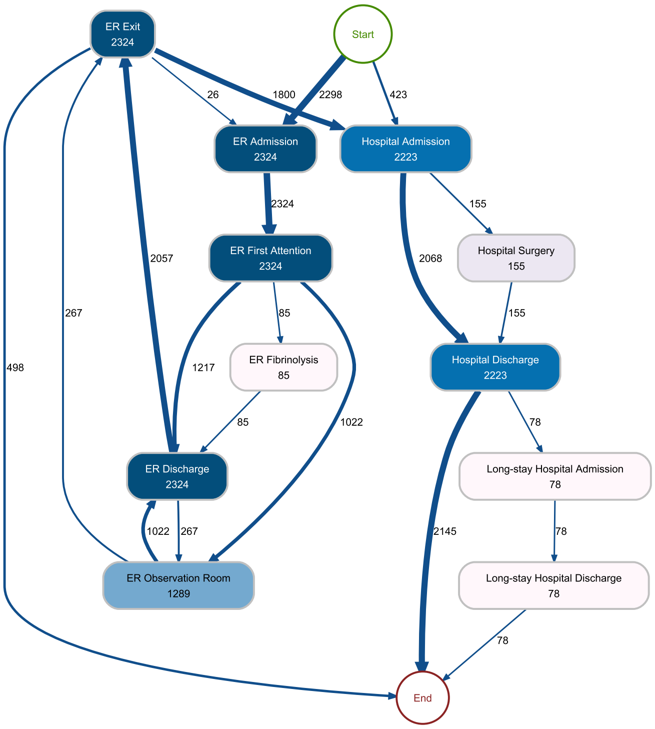

Fig. 4. Process map of episodes covering more than 95% of total episodes with transitions' frequency information.

solution does not entail any difference in the management of the Code Stroke.

The outputs presented here are divided in two different scopes: frequency considerations and time considerations.

1) CPW Frequency Considerations: The two main outputs to analyse the frequency elements of the Code Stroke CPW are process traces and process map with frequency information. In terms of health services research, information on the counts of crude rates of certain activities provides useful insight to understand the intensity of use, whether the patients follow the expected paths, and allows to discover outlier behaviors.

Process traces, see Fig. 3, are an account of the different sequences of activities observed in the activity log, in other words, the different types of episodes that took place in the care pathway. In this plot, the x-axis represent the sequence of the activities performed, the y-axis represents the different traces observed and the colours represent different activities. Process traces are ordered decreasingly with respect to the percentage of episodes that correspond to the given trace (the percentage in the grey box at the right side of the trace graph). In the plot of Fig. 3 only those traces that cover more than 95% of the total episodes have been depicted.

The process traces depicted in Fig. 3 clearly point that the Code Stroke CPW has a two major traces on its real implementation that cover nearly 50% of the most observed episodes. First trace represents those stroke episodes that start in the Emergency Room department and finish in a hospitalisation, with no other specific interventions (no ER fibrinolysis, no Hospital surgery). Second most observed trace is nearly equal to the first one with the main difference that the "observation room" activity in the ER is not present. The third most frequent trace correspond to those patients directly treated in the Hospital and, the fourth most frequent trace correspond to those patients only treated in Emergency Room.

This frequentist approach may be visually complemented with the frequency process map depicted in Fig. 4. A process map is a directed graph where the nodes are the activities, whose name appears in the node label, and an edge or an arc between activity A and B indicates that activity A is directly followed by activity B. Edges contain a figure indicating the number of episodes where the specific transition was observed in the activity log being the edge thickness proportional to this value. The coloring of the nodes is a gradient from white to dark blue, indicating the number of episodes where the specific activity appears, a number also indicated in the node label.

Analysing this figure in depth, it is possible to obtain a finer level of granularity to understand the relationships between activities that take place in the real-world implementation of the Code Stroke pathway. Some examples are the following: the typical entry point to the Code Stroke pathway is ER Admission (2298 patients vs. 423 patients that went directly to Hospital Admission); the typical exit point is the Hospital Discharge (2145 patients vs. 576 from other activities); within the ER activities, nearly half of the patients move from ER Admission to ER Discharge (1217 patients of 2324 total), nearly the other half move to ER Observation Room (1022 patients of 2324 total), and a minority receive fibrinolysis treatment (85 patients of 2324 total); typically, ER Exit is followed of a Hospital Admission (1800 patients vs. 498 patients that leave pathway); and, typically, the patients leave the pathway after the Hospital

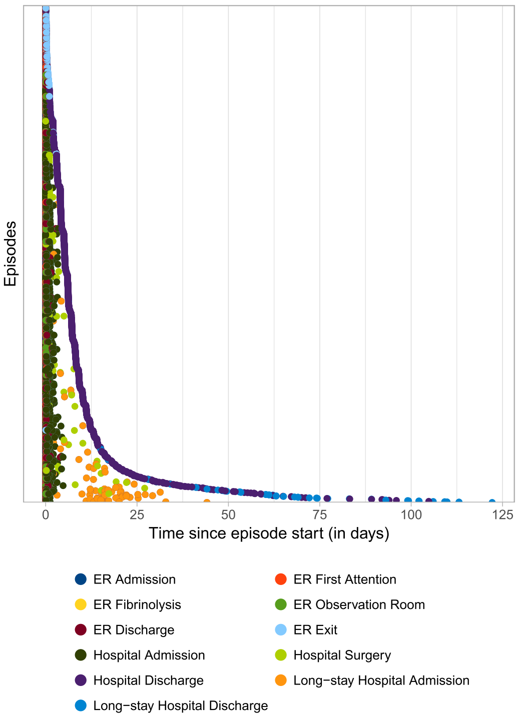

Fig. 5. Timeline of episodes covering more than 95% of total episodes.

Discharge (2145 patients vs. 78 patients that move to a Long-stay Hospital Admission).

2) CPW Time Considerations: when it comes to timing across the pathway, timeline plot, depicted in Fig. 5, offers an overall picture of all episodes. In this timeline the x axis represents time, the y axis represents the different episodes, and the color indicate the activity type. In the timeline of Fig. 5 the time expressed is the *relative* episode time, i.e. the time elapsed since the episode start,[2] as a coarse view of the time distribution.

Using this representation, it is possible to distinguish the initial set of episodes that finish in ER Discharge (light blue dots), whose duration is within a day. Then, there is the vast majority of the episodes, that finish in Hospital Discharge, with a duration between 1 day and 25 days. Finally, there is also a small set of episodes, that include Long-stay Hospital activities (orange dots and darker blue dots).

Complementing this coarse analysis, the process discovery also produces a process map with time information output which gives a fine grain knowledge of Code Stroke pathway. This plot, depicted in Fig. 6, is essentially the same directed graph depicted in Fig. 4, but the *extra* information refers to the time dimension: the edges contain the transitions' median time observed in the activity log, and the edges the median duration of the activities, always set to zero due to the characteristics of the activity log where activities have only a single time.

Analysing the process map of Fig. 6 it is possible to observe the time spent in the different transitions between activities as a straightforward approach to evaluate the time-conformance of the actual implementation of the pathway. Some interesting results observing the time information in Fig. 6 are: the median time till the ER First Attention is 18 minutes; the median time to fibrinolysis, adding time from ER admission to ER first attention activities and time from ER first Attention to ER fibrinolysis, is around 1 hour (see detailed information about fibrinolysis in the following section); the median time from ER first attention to ER observation room is around 1 hours and 48 minutes; the median time from ER discharge and actual ER exit is around 1 hour and 20 minutes; the hospital median length of stay is around 144 hours (6 days) in admissions without surgical procedure and 96 hours (4 days) when the patient underwent a surgery; finally, the longest transition observed in the pathway corresponds to stay at long-stay hospitals, whose median time is 708.5 hours (~29 days).

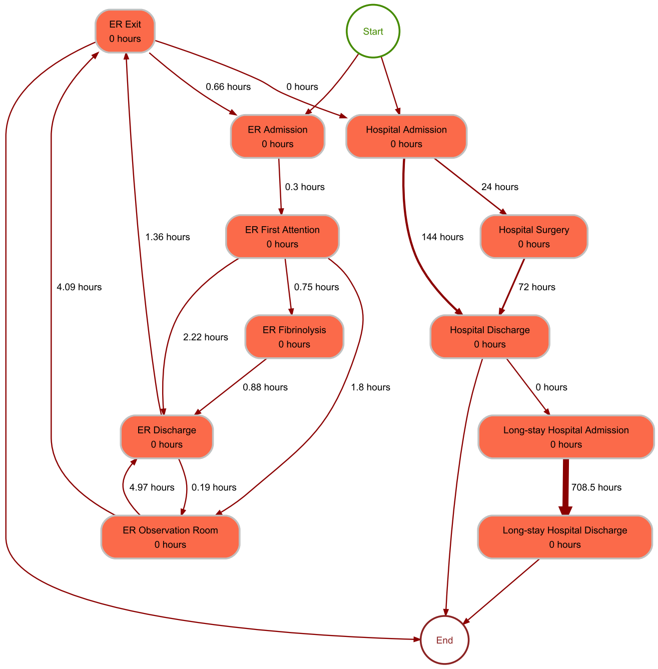

Fig. 6. Process map of episodes covering more than 95% of total episodes with transitions' median duration information.

F. Task 6: Process Improvement and Support. On the Fibrinolysis Treatment

Although it is not the objective of the present paper to present a detailed clinical approach of the Code Stroke CPW real-world implementation, the fibrionolysis treatment analysis is a clear example to demonstrate the usefulness of the approach to evaluate the uptake of symptoms, the adequacy of the health care provided and ultimately, the impact on health outcomes. Fibrinolysis is a clinical treatment with high influence in patients' survival but requires to be provided in tight time frame since the onset of the stroke symptoms (4.5 hours since the start of the symptoms according to Code Stroke Aragon guideline [28]). Due to its importance, the treatment with fibrinolysis is captured as a specific activity.

A quick evaluation of appropriateness may be done by combining the information of those traces that include "ER Fibrinolysis activity, depicted in Fig. 3 and the process map in Fig. 7. This

[2] It is also possible to depict the *absolute* time, i.e. wall-clock time of activities

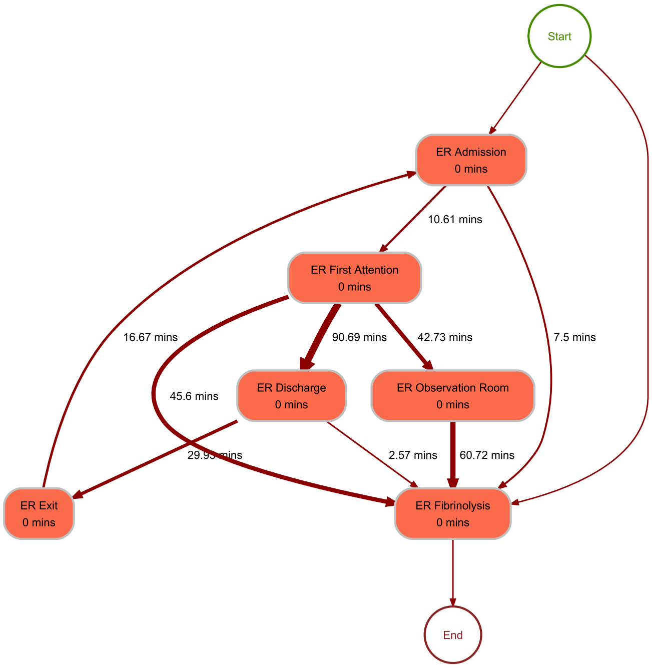

Fig. 7. Process map with transitions' median duration trimmed to end at fibrinolysis activity.

process map has been constructed by filtering those episodes including the "ER Fibrinolysis" activity and then trimming the episodes by discarding all those activities after the "ER Fibrinolysis" (a single filtering function of the bupaR packages). This filter increases the legibility of the process map enabling a clear view that, in all the possible episodes that include the treatment, the median time-to-fibrinolysis is below the desired 4.5 hours (within the accounted time frame). Transposing the traces that include the "ER Fibrinolysis" activity in Fig. 3 to the process map in Fig. 7, i.e. following the sequence of activities indicated in the traces within the process map, the largest median time observed is around 240 minutes.

In addition, thanks to the information present in process models obtained from the input datasets, such as the first ER that treated the patient, it is possible to easily create detailed analysis plots as the one in Fig. 8. This Figure exhibits the distribution of time-to-fibrinolysis grouped according the first ER of treatment, i.e. the ER facility where the patients entered to the pathway. The box-plots show that the median time-to-fibrinolysis for all hospitals is below 100 minutes. Hospitals 4, 7, 8, 11, 12 and 18 correspond to those smaller hospitals in the region where Code Stroke was recently deployed, have a median time-to-fibrinolysis between 50 and 100 minutes, being the hospitals 8 and 11 the ones with larger variability. This median time-to-fibrinolysis decreases below 50 minutes in hospitals 1, 9 and 10, the largest hospitals in the region with high experience in the Code Stroke CPW.

IV. Methodological Caveats and Limitations

Once the utility of the process model construction based on RWD sets has been documented, it is important to highlight here a set of elements that should be considered when facing this task in other "experimental" settings.

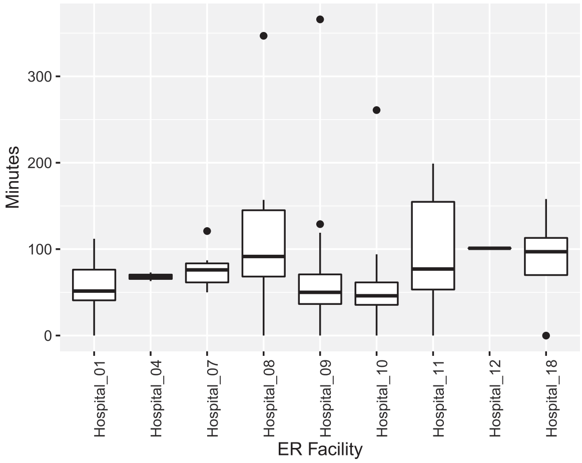

Fig. 8. Time to fibrinolysis treatment per ER facility of first contact.

A. Data Model

It has been mentioned in the text the importance of the data model for the analysis purposes. The data model may limit the possibilities of further analyses or the comparability among different process models obtained from different providers or health systems.

B. Data Availability

To the purpose of this work, three RWD sets were used. The RWD sets correspond to structured administrative data available in the Aragón health system. Undoubtedly, the addition of more RWD sets could have enriched the result, for example including the unstructured data coming from the clinical notes or reports written in natural language. Works such as [31] and [32] serve as an illustration on how to use natural language processing (NLP) approaches to capture unstructured data from electronic health records (EHR) and clinical reports, the most typical RWD sets available in health systems nowadays.

In order to asses whether not considering the inclusion of unstructured data could have implied a major limitation in the current study, an ad hoc exploration of a number of stroke discharge reports revealed two additional timestamped activities not present in the RWD sets used in this paper: functional analysis of patients and CT imaging time. However, these activities were not consistently in all reports. In any case, the inclusion of unstructured data via NLP techniques should be a key point in the future work of the current research.

C. Data Quality

As in any other data analysis research, the data quality is a crucial aspect to take into account. Closely related to the data model, it is strictly necessary to gain a high expertise of data semantics, previously cited, in order to ensure a high data quality. In the present work, this expertise is reflected in multiple steps of the methodology: first, the ability to capture the information from the multiple RWD sets, detecting the sources but also translating the case definition to extract the desired data from these datasets; second, the synthesis of the rules that capture the care continuity in the episode linking algorithm, an algorithm that deals with the datasets coverage and harmonizes the semantics; and third, the evaluation of the adequacy of the resulting episodes and the process mining outputs with the expert knowledge of the acute stroke CPW.

D. Data Granularity

Even being a data quality element, the data granularity should be treated on its own. This is one of the most sensitive elements when dealing with timestamped data and event logs. As detailed in the episode linking, Section III-C, the more granularity the data has, the better to link the RWD sets. The Episode linking algorithm synthesizes all the logic required to deal with the time coherence among datasets with different time granularity. However, an additional cavet should be considered when it comes to comparability: those datasets with less granularity will result in worse or impossible comparisons, for example, in the current paper the measurement of the time lagging from admission to treatment with fibrinolysis has been possible due to the availability of a high degree of granularity (hour and minute) in the timestamps of these activities.

V. Application to Health Services Research

For health services research, the construction of empirical process models from RWD sets provides a unique point-of-view to understand both how patients and health system interact and also how health systems organize in the reality to guide these interactions. Some potential uses of process discovery in health services research are the following.

A. Conformance Checking

The conformance checking is the evaluation of adequacy of the actual practices observed in the empirical process model as compared to those described in the theoretical or normative CPW. Conformance checking is one of the most useful applications of the resulting process models when focusing on the organizational point-of-view.

The time-to-fibrinolysis evaluation presented in section III-F is an informal example of the conformance checking, contrasting the real-world observed times against the defined in the Code Stroke guideline [28]. This is an analysis that provides a high value when evaluating Code Stroke CPW implementation and was performed within minutes with few source code lines of the `bupaR` package.

A more sophisticated conformance checking approach will result by comparing the empirical process model discovered from the RWD sets to a manually codified model of the theoretical or normative CPW or the process model extracted from the clinical guidelines using NLP in a similar way to the work done in [33] for an archaeology manual. Independently on how the reference model is constructed, the comparison between the empirical process and the reference model may be done, using for example *event log replay* techniques [34], i.e. simulate how real-world episodes can follow the theoretical process model, or graph dissimilarity metrics, i.e. measure how different the real-world process maps are compared to theoretical process maps as described in [35].

B. Benchmarking of Providers

The benchmarking is a derived result of the conformance checking introduced in the previous point, and has been hinted in the comparison of the different hospital ER facilities presented when exposing the adequacy of the treatment with fibrinolysis in the Section III-F.

Thus, having the different degrees of adequacy and conformance between the real-world implementation of the CPW and theoretical or normative CPW in the different care providers can be used to compare and contrast the organizational approach of these providers. Those best, i.e. similarity to the normative path, will serve as benchmark providers to improve the quality of the healthcare services. The study presented in [5] proposed a similar benchmark.

It is interesting to note that the work presented in this paper is being used to perform a provider benchmarking within the European project ICTUSnet,[3] comparing CPW from 7 regions of 4 different European countries.

C. Comparative Effectiveness Research

Remains as next step of the current research the analysis of how the exposure to health system affect to health outcomes by exploiting individual and health system variables included in the RWD sets, and thus, in the inferred process model. This type of analyses aim at comparing effectiveness in two different facets.

First, measuring the effects of the traces that patients follow, comparing first health outcomes and/or costs of patients with different characteristics that follow the same trace and second comparing patients with same characteristics that follow different traces. Second, evaluating those decision points in the process model, i.e. those bifurcations in the process map. Applying decision mining techniques and using patients and health system variables, it is possible to measure which variables affect to the decision to follow one path or another and how the decision affect to health outcomes and/or costs.

D. Clinical Translation

Although the focus of this paper is on the health services research point of view, the results of the CPW analyses could ultimately be applied in the clinical practice. Serving as an example: 1) the results of the conformance checking may lead to the reorganization of those services that do not attend to the specific clinical practice guidelines; 2) the benchmarking of providers may serve to stablish new cross-organizational policies according to the best performing providers' structures and management; 3) the results of the comparative effectiveness research can be directly applied in regular medical practice as it will unveil those practices and interventions within the CPW that result in better health outcomes; and 4) process models enriched with clinical data may be used with other data mining/machine

[3] [Online]. Available: http://ictusnet-sudoe.eu/en/

learning techniques (e.g., random forests, neural networks) to predict adverse events within the CPW.

VI. Conclusion

It has been shown and discussed the viability of constructing care pathways' process models by linking RWD sets and using process mining analysis. To this end, there was a careful planning to define a data model and to develop a linking algorithm that guarantees the time coherence across the data available from the different datasets. The solution proposed solves the limitations pointed by W. Yang and Q. Su in [20] regarding the data availabilty. The actual source code of the solution is available in a GitHub repository, with a total guarantee of reproducibility for further experimentation.

The solution proposed has been shown to have great value for different stakeholders as it opens the door to a wide set of in-depth studies of how health systems organize, how this organisation may impact health outcomes and how to improve the organisation both at health system level and, in a final term, in the day-to-day clinical practice.

References

[1] H. Campbell, R. Hotchkiss, N. Bradshaw, and M. Porteous, "Integrated care pathways," *Brit. Med. J. (Clin. Research ed.)*, vol. 316, no. 7125, pp. 133–7, Jan. 1998. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2665398

[2] T. Rotter *et al.*, "Clinical pathways: Effects on professional practice, patient outcomes, length of stay and hospital costs," *Cochrane Database Systematic Rev.*, no. 3, Mar. 2010, doi: 10.1002/14651858.CD006632.pub2.

[3] D. J. Zand, K. M. Brown, U. Lichter-Konecki, J. K. Campbell, V. Salehi, and J. M. Chamberlain, "Effectiveness of a clinical pathway for the emergency treatment of patients with inborn errors of metabolism," *Pediatrics*, vol. 122, no. 6, pp. 1191–1195, Dec. 2008.

[4] J. Lenkowicz *et al.*, "Assessing the conformity to clinical guidelines in oncology: An example for the multidisciplinary management of locally advanced colorectal cancer treatment Jacopo," *Manage. Decis.*, vol. 56, no. 10, pp. 2172–2186, Oct. 2018.

[5] A. Partington, M. Wynn, S. Suriadi, C. Ouyang, and J. Karnon, "Process mining for clinical processes: A comparative analysis of four Australian hospitals," *ACM Trans. Manage. Inf. Syst.*, vol. 5, no. 4, pp. 1–18, Jan. 2015.

[6] W. M. P. van der Aalst, *Process Mining: Data Sience in Action*, 2nd ed. Berlin, Germany: Springer, 2016.

[7] L. P. Garrison, P. J. Neumann, P. Erickson, D. Marshall, and C. D. Mullins, "Using real-world data for coverage and payment decisions: The ISPOR real-world data task force report," *Value Health*, vol. 10, no. 5, pp. 326–335, Sep. 2007.

[8] M. L. van Eck, X. Lu, S. J. J. Leemans, and W. M. P. van der Aalst, "PM^2 : A process mining project methodology," in *Proc. Int. Conf. Adv. Inf. Syst. Eng*, 2015, pp. 297–313.

[9] W. van der Aalst, T. Weijters, and L. Maruster, "Workflow mining: Discovering process models from event logs," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 9, pp. 1128–1142, Sep. 2004.

[10] A. J. M. M. Weijters, W. M. P. Van Der Aalst, and A. K. Alves De Medeiros, "Process mining with the heuristics miner-algorithm," Technische Universiteit Eindhoven, Eindhoven, The Netherlands, Tech. Rep. WP 166, pp. 1–3, 2006.

[11] C. W. Günther and W. M. Van Der Aalst, "Fuzzy mining - Adaptive process simplification based on multi-perspective metrics," in *BPM 2007: Business Process Management*, Lecture notes in computer science, vol. 4714. Berlin, Germany: Springer, 2007, pp. 328–343.

[12] S. J. J. Leemans, D. Fahland, and W. M. P. van der Aalst, "Discovering block-structured process models from event logs—A constructive approach," in *Proc. Int. Conf. Appl. Theory Petri Nets Concurrency*, 2013, pp. 311–329.

[13] J. Munoz-Gama, *Conformance Checking and Diagnosis in Process Mining*, vol. 270, lecture notes in business information processing. Cham, Germany: Springer, 2016.

[14] A. Rozinat and W. van der Aalst, "Conformance checking of processes based on monitoring real behavior," *Inf. Syst.*, vol. 33, no. 1, pp. 64–95, Mar. 2008.

[15] B. F. van Dongen, A. K. A. de Medeiros, H. M. W. Verbeek, A. J. M. M. Weijters, and W. M. P. van der Aalst, "The ProM framework: A new era in process mining tool support," in *Proc. Int. Conf. Appl. Theory Petri Nets*. 2005, pp. 444–454.

[16] C. W. Günther and A. Rozinat, "Disco: Discover your processes," in *Proc. Demonstration Track 10th Int. Conf. Business Process Manage*, 2012, pp. 40–44.

[17] G. Janssenswillen, B. Depaire, M. Swennen, M. Jans, and K. Vanhoof, "bupaR: Enabling reproducible business process analysis," *Knowl.-Based Syst.*, vol. 163, pp. 927–930, Jan. 2019.

[18] W. van der Aalst, "Process mining: Discovering and improving Spaghetti and Lasagna processes," in *Proc. IEEE Symp. Comput. Intell. Data Mining*, Apr. 2011, pp. 1–7.

[19] L. Maruster and R. Jorna, "From data to lnowledge: A method for modeling hospital logistic processes," *IEEE Trans. Inf. Technol. Biomedicine*, vol. 9, no. 2, pp. 248–255, Jun. 2005.

[20] W. Yang and Q. Su, "Process mining for clinical pathway: Literature review and future directions," in *Proc. 11th Int. Conf. Service Syst. Service Manage*, Jun. 2014, pp. 1–5.

[21] E. Rojas, J. Munoz-Gama, M. Sepúlveda, and D. Capurro, "Process mining in healthcare: A literature review," *J. Biomed. Informat.*, vol. 61, pp. 224–236, Jun. 2016.

[22] E. Batista and A. Solanas, "Process mining in healthcare: A systematic review," in *Proc. 9th Int. Conf. Inf, Intell., Syst. Appl.*, Jul. 2018, pp. 1–6.

[23] Á. Rebuge and D. R. Ferreira, "Business process analysis in healthcare environments: A methodology based on process mining," *Inf. Syst.*, vol. 37, no. 2, pp. 99–116, Apr. 2012.

[24] S. Suriadi, R. S. Mans, M. T. Wynn, A. Partington, and J. Karnon, "Measuring patient flow variations: A cross-organisational process mining approach," in *Proc. Asia Pacific Business Process Manage.*, 2014, pp. 43–58.

[25] R. Mans *et al.*, "Process mining techniques: An application to stroke care," *Studies Health Technol. Informat.*, vol. 136, pp. 573–8, 2008.

[26] M. Binder *et al.*, "On analyzing process compliance in skin cancer treatment: An experience report from the evidence-based medical compliance cluster (EBMC^2)," in *Proc. Int. Conf. Adv. Inf. Syst. Eng.*, 2012, vol. 7328, pp. 398–413.

[27] I. Litchfield, C. Hoye, D. Shukla, R. Backman, A. Turner, M. Lee, and P. Weber, "Can process mining automatically describe care pathways of patients with long-term conditions in UK primary care? A study protocol," *BMJ Open*, vol. 8, no. 12, Dec. 2018.

[28] O. Alberti González *et al.*, *Plan de Atención al Ictus en Aragón. Actualización 2019-2022*, M. Bestué Cardiel, J. Marta Moreno, and G. Martínez Borobio, Eds., Zaragoza, Spain: Gobierno de Aragón. Dirección General de Asistencia Sanitaria, 2018. [Online]. Available: https://www.aragon.es/estaticos/GobiernoAragon/Departamentos/SanidadBienestarSocialFamilia/Sanidad/Documentos/Programa_Ictus_actualizacion2019.pdf

[29] P. Conde-Espejo, *Evaluación de la Eficiencia de Modelos Organizativos Para el Abordaje del Ictus (Unidades de Ictus)*. Madrid, Spain: Consejería de Sanidad - D. G. de Planificación, Investigación y Formación, 2013. [Online]. Available: http://www.comunidad.madrid/publicacion/1354388439231

[30] K. Chodorow and M. Dirolf, *MongoDB: The Definitive Guide*. Sebastopol, CA, USA: O'Reilly Media, 2010. [Online]. Available: http://cds.cern.ch/record/1359920

[31] E. S. Chen and I. N. Sarkar, "Mining the electronic health record for disease knowledge," *Biomed. Literature Mining*, vol. 1159, pp. 269–286, 2014.

[32] T. Delespierre, P. Denormandie, A. Bar-Hen, and L. Josseran, "Empirical advances with text mining of electronic health records," *BMC Med. Inform. Decis. Making*, vol. 17, p. 127, Dec. 2017, doi: 10.1186/s12911-017-0519-0.

[33] E. V. Epure, P. Martin-Rodilla, C. Hug, R. Deneckere, and C. Salinesi, "Automatic process model discovery from textual methodologies," in *Proc. of 2015 IEEE 9th Int. Conf. Res. Challenges Inf. Sci.*, vol. 2015, P. Loucopoulos, C. Gonzalez-Perez, C. Rolland, and D. Anagnostopoulos, Eds., Los Alamitos, CA, USA: IEEE Comput. Soc. Press, May 2015, pp. 19–30.

[34] W. van der Aalst, A. Adriansyah, and B. van Dongen, "Replaying history on process models for conformance checking and performance analysis," *Wiley Interdisciplinary Rev.: Data Mining Knowl. Discovery*, vol. 2, no. 2, pp. 182–192, Mar. 2012.

[35] S. Montani, G. Leonardi, S. Quaglini, A. Cavallini, and G. Micieli, "Mining and retrieving medical processes to assess the quality of care," in *Proc. Int. Conf. Case-Based Reasoning*. Saratoga Springs, NY, USA, 2013, pp. 233–240.