Subpopulation Process Comparison with the Help of Ontological Foundation: A discussion

Faiza Bukhsh[0000−0001−5978−2754], Jeewanie Jayasinghe Arachchige[0000−0001−8619−6523], and Priya Naguine[0000−0002−1225−2040]

University of Twente, Enschede, The Netherlands f.a.bukhsh@utwente.nl, j.a.jayasinghearachchige@utwente.nl, p.v.naguine@student.utwente.nl,

Abstract. "Process modelling and mining frameworks have demonstrated their effectiveness across diverse domains, including healthcare. However, existing frameworks often lack explicit guidance on learning from best practices. For instance, the case of Frozen shoulder (FS), a condition with multiple treatment options and varying outcomes. Understanding how care paths differ among patient groups and determining the most effective approach remains a challenge. By identifying this gap, our research employs the Process Mining Project Methodology in Healthcare (PM^2HC) alongside the MIMIC-IV dataset to uncover distinctions in care paths among different age groups and genders. This experimental validation seeks to identify optimal strategies for addressing Frozen shoulders through ontological concepts. The study concludes by presenting a set of open challenges, aiming to guide future research in healthcare by integrating ontological concepts to learn from the best and optimal care paths. It is important to note that while this research doesn't offer a singular solution, it contributes significantly by opening a new dimension of ontological research. Specifically, it delves into how various care paths can be compared and aligned with the help of ontological foundation."

Keywords: subpopulation comparison · ontology · frozen shoulder · process mining

1 Introduction

Medical professionals often question whether there is a difference in the treatment procedures followed by subgroups of patients diagnosed with the same disease [10]. In this case, a subgroup refers to a group of patients with a common characteristic, e.g., all female patients diagnosed with frozen shoulder. When comparing subpopulations, experts' knowledge is essential. However, Ontology-based interpretation is a valuable technique for capturing better insight into a complex domain like healthcare.

In the context of information science and knowledge representation, an ontology is a formal and explicit specification of a shared conceptualization [6]. It provides a structured framework for representing knowledge in a particular domain by defining the entities, their properties, and the relationships between

them. Ontologies aim to capture a common understanding of a domain and facilitate communication and interoperability among different systems and applications [5].

This research aims to identify carepath for different subpopulations and learn from best practices. Since age and gender play a role in the development of a disease such as FS [9], these can be chosen as the subgroups. There has not been much research on comparing these sub populations. Ontological concepts are rich in nature and can provide a methodological way to compare subpopulations [11]

Therefore, the objective of this research is to find the differences and similarities between the care paths for different subpopulations and learn from best practices with the guidance of ontological foundations.

As an example scenario, we used the MIMIC-IV database and analysed the procedures followed by FS patients; this database contains data on approximately 300,000 patients that were admitted to a tertiary academic medical centre in Boston, the USA, between 2008 and 2019 [4,8]. Specifically, process mining takes data from hospital information systems (HIS) when applied in the healthcare domain [12]. Event logs are then created using the data from the HIS to show the sequence of processes followed by patients. The event logs created can then be used to find the differences in the care paths followed by subgroups of patients with FS. Further, to show comparison examples, the differences and similarities of the care paths were analyzed with BPMNDiffViz tool[1].

This paper is structured as follows. The state of the art will be described in section 2. Section 3 will describe the methodology used with the title of "An example Scenario: Frozen Shoulder Exploration with the Application of Process Mining (PM)". Section 4 will discuss the subpopulation comparison through ontological foundations. Finally, the paper concludes with opening insightful research directions to the reader in the last section 5.

2 State of the Art

Process mining techniques can be used for various purposes in the healthcare domain, e.g., with BPMN diagrams to get the graph edit distances. [11] used process mining techniques specifically for the process comparison of subgroups. There was a focus on the application of process mining for subpopulation process comparison between patients diagnosed with different types of cancer.

The tool BPMNDiffViz can be used to find graph similarity measures. It takes as input two BPMN diagrams and gives the minimal graph edit distance (GED) as a result. The GED can be defined as the minimum number of steps required to transform one graph into another [17]. This tool makes use of Business Process Model and Notation (BPMN) 2.0, which is one of the frequently used notations used for process modelling [7]. [16] states that although the use of BPMN diagrams in medicine is a recent development, it can be used to model clinical pathways to teach and train medical staff.

[1] https://pais.hse.ru/en/research/projects/CompBPMN/

Visual comparison can be used to differentiate between the care paths followed by subgroups of patients and the tool BPMNDiffViz can be used for that. BPMNDiffViz allows for a choice between six comparison algorithms: Greedy, TabuSearch, Genetic, AStar, Ants and simulated annealing. [17] compares the algorithms mentioned except Genetic and concludes that the Greedy algorithm gives the best performance results while the TabuSearch algorithm gives more precise and accurate results. The Genetic algorithm only gives an approximation of the GED [15].

Subpopulation comparison based on visual aspects or graph edit distances provides us with an initial view. However, the robustness of these comparisons should be strengths beyond the statistical figures for making decisions in real situations, especially in complex domains like healthcare.

Ontological foundation is one of the potential approaches that can be used to ensure the accuracy of domain structures. Ontological concepts are hierarchical domain structures that provide a domain theory, have a syntactically and semantically rich language, and a shared and consensual terminology [3].

The work of [18] explores ontology learning, a dynamic research field crucial for effective ontology engineering. It distinguishes ontology-based definitions from conventional label-centric ones, emphasizing the interconnected nature of objects. This shift allows for advanced functionalities such as scenario search, ontology fusion, and recommendation through nuanced relation labelling. Moreover work of [3] discusses the potential of ontology-based process modelling (OBPM) to enhance business process management theoretically.

3 An example Scenario: Frozen Shoulder Exploration with the Application of Process Mining (PM)

In the following section, we will methodically elaborate on a specific scenario to illustrate the design and comparison of subpopulations. Throughout this example, we will highlight the potential role that could be played by ontological concepts in shaping and evaluating these subpopulations. The methodology to be used in this research is called Process Mining Project Methodology in Healthcare (PM^2HC) [14]. PM^2HC involves 6 phases: planning, extraction, data processing, mining and analysis, evaluation, and improvement and support.

3.1 Planning

During this phase, we chose specific subgroups to explore various care paths and organized the sequence of events. Additionally, we conducted thorough background research on frozen shoulder and process mining in healthcare, as detailed in the [13].

3.2 Extraction

In order to get access to and query the MIMIC-IV database, Google Cloud Platform BigQuery[2] was used. Since the MIMIC-IV database stores the diagnoses given to the patients at the end of their ICU stay using the International Classification of Diseases (ICD) Version 9 and 10 codes, the first step was to find the ICD codes associated with frozen shoulder. This was found in the **D_ICD_DIAGNOSES** table [1] using the keywords *frozen shoulder* and *adhesive capsulitis* for the *long_title*. The ICD codes are 7260, M750, M7500, M7501, and M7502 and their corresponding diagnoses are "Adhesive capsulitis of shoulder", "Adhesive capsulitis of shoulder", "Adhesive capsulitis of unspecified shoulder", "Adhesive capsulitis of right shoulder" and "Adhesive capsulitis of the left shoulder." It is important to note that there is a possibility that a patient is given more than one diagnosis associated with the frozen shoulder in a single hospitalization, e.g., a patient can be diagnosed with both M7501 and M7502.

To apply process mining algorithms to the data, the cases, events, start times and end times have to be defined. For both the subgroup process comparison and bottleneck analysis, a case is a patient's admission to the hospital and the events are the procedures that the patients were billed for.

Since the start and end times were not stored for the subgroup process comparison, the sequence number was used instead to indicate the order in which the procedures were carried out.

3.3 Data processing

In this phase, the CSV files on the subgroups were entered into ProM, converted into XES files and visualised using the **LogVisualiser (LogDialog)** plugin. Table 1 gives an overview of the number of cases and events per subgroup, given by the LogDialog. Also, further filtering was required to find the differences in care paths between the different patient groups. This was done using the **Filter Log on Event Attribute Values** plugin, where specific procedures were filtered out from the care paths.

3.4 Mining and Analysis

This phase involved finding the differences in the care paths between the different subgroups and the bottlenecks in the medications taken and the procedures followed by patients during their ICU stays. To do this, process models were created in ProM[3] and Disco[4].

The Inductive Miner plugin was chosen because it gives the best fitness, i.e., the degree by which the process models generated can recreate the cases in the

[2] https://cloud.google.com/bigquery
[3] https://promtools.org/
[4] https://fluxicon.com/disco/

Table 1: Number of cases and events per subgroup

Subgroup	#Cases	#Events
Female *	29	61
Male *	34	55
Age below 40 **	8	18
Age between 40 and 60 **	39	73
Age above 60 **	16	25

* Includes FS patients from all age groups
** Includes FS patients from both genders

event log [2]. At first, the plugin **Mine with Inductive visual Miner** was used because it can create animations showing the order in which the processes occur; it was used with the *activities* slider set to 1 and the *paths* slider set to 0.8. These settings were chosen so that the Petri net and the Inductive visual Miner models are equivalent. Secondly, **Mine Petri net with Inductive Miner** was used to create static process models that can be used for visual comparison, with a *noise threshold* of 0.2 to allow for slight deviations. Lastly, in order to convert the Petri net models into BPMN diagrams so that they can be loaded into BPMNDiffViz to get the GED, **Convert Petri net to BPMN diagram** was used.

The process models created in ProM and Disco for the subgroup process comparison and bottleneck analysis can be found in the author's GitHub repository[5].

When comparing the care paths of the subgroups, three keywords will be used. Firstly, **parallel** will be used when two procedures occur but the order in which they occur does not matter. Secondly, **sequence** is used when one procedure follows another. Lastly, **exclusive** will be used when only one of two procedures can occur.

Also, visual comparison is performed in BPMNDiffViz using the **TabuSearch** algorithm with *maximum expansions* and *tabu list size* set to 100 as this gives precise results faster than other algorithms [17]. The activities in the BPMN diagrams are encoded with different colours: blue denotes elements that match between the subgroups, green denotes elements that should be added to transform one diagram into the other and red denotes elements that should be deleted to transform one diagram into the other.

Visual comparison was made in BPMNDiffViz for the care paths followed by male and female FS patients, resulting in a final score of 167 using the TabuSearch algorithm. 37% of the elements matched between the care paths, 33% of the elements were deleted and 30% of the elements were added. Table 2 shows the procedures that are only performed on either female or male FS patients, but not both.

[5] https://github.com/PriyaNaguine/Complete-Process-Models-Frozen-Shoulder

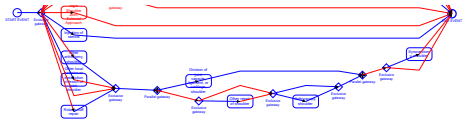
Table 2: Procedures performed on either male or female FS patients

Procedure	Female	Male
Drainage of Right Shoulder Joint, Percutaneous Approach, Diagnostic		✓
Excision of Left Shoulder Bursa and Ligament, Percutaneous Endoscopic Approach	✓	
Excision of Right Shoulder Joint, Percutaneous Endoscopic Approach		✓
Other total shoulder replacement	✓	
Release Right Shoulder Joint, External Approach		✓
Repair of recurrent dislocation of shoulder	✓	
Repair Right Shoulder Joint, Percutaneous Endoscopic Approach		✓
Repair Right Shoulder Tendon, Open Approach		✓

The procedure **"Other repair of shoulder"** can be done in parallel with "Division of joint capsule, ligament, or cartilage, shoulder" in male patients while in female patients, these procedures are performed in sequence. Furthermore, it is performed in sequence with "Rotator cuff repair" in male FS patients. However, in female patients, these processes are exclusive. This can be seen in figures 1a and 1b.

As can be seen in figures 1a and 1b, the procedure **"Synovectomy, shoulder"** is always the last process in male FS patients, in case it is performed. In female patients, it is exclusive to "Rotator cuff repair", while in male patients, they can occur in sequence, where "Rotator cuff repair" is the first procedure and "Synovectomy, shoulder" is the last procedure to take place.

(a) Snapshot of the BPMN diagram for female FS patients



(b) Snapshot of the BPMN diagram for male FS patients

3.5 Evaluation and improvements

In this phase, the insights obtained in the previous phase were used to suggest improvements and learn from care path by considering the best practices. In this phase, the stakeholders, e.g., medical professionals, decide on the path to be followed to implement the improvements.

This phase was conducted with an expert physiotherapist at Fysiotherapie Polman in Enschede, The Netherlands, in order to discuss and evaluate the

results of this research. Thereby, based on the discussion of the results found with the physiotherapist, which is based on his 8 years of experience working as a physiotherapist for FS, more insight was gained on patients of FS. In particular, there is a ratio of approximately 7:3 between female and male FS patients. This could be because female patients ask for help earlier on. Based on his experience, there is no difference in the care paths followed by male and female FS patients. Also, it was mentioned that the age group between 40 and 60 years old is more prone to developing FS and this applies to both genders. Furthermore, older people, i.e., those aged above 60, are more likely to experience FS after shoulder trauma. In this age group, they are less likely to get surgery as it is an invasive procedure. In general, depending on the health conditions of the patient, the older they are, the more they are at risk of developing complications.

4 Subpopulatoin comparison through ontological foundation

Section 3 of our research shows how different care paths of subpopulations derived using process mining and how they compare using BPMNDiffViz tool. Subpopulation comparison based on visual aspects or graph edit distances provides us with an initial view. There is no doubt, that we can argue the robustness of these comparisons in terms of statistical figures. However, the question is, whether these statistical figures are sufficient for making decisions in real situations, especially in complex domains like healthcare. Ontological foundation is one of the good approaches that can be used to ensure not only the structural correctness but also the accuracy of domain knowledge in the derived models.

In our research case study, we analyzed the treatment procedures and care paths for FS within two distinct subpopulations. The central focus of this study revolves around understanding the variations in care paths among different patient groups and determining the most effective approach. Naturally, the expertise of domain professionals serves as the primary and most fitting source of knowledge for these investigations. Secondly, the ontological foundation can be employed to determine the best care path. Surprisingly, research work is scarce on using ontologies for comparing (assessing similarities or differences) different care paths based on subpopulations.

To address this, our case study explores a research direction on establishing a method for comparing subpopulations within a given knowledge domain, along with defining appropriate evaluation criteria. These criteria encompass the ontological richness and the reliability of methodologies in conceptualization, shareability in terms of sources and granularity, explicitness and formality through implementation tools and formalization language, and adherence to design criteria within the methodological process of building ontologies.

In essence, our proposal leverages the significance of the ontology definition as a foundation for comparison features, ensuring a comprehensive evaluation that goes beyond traditional similarity metrics.

5 Food for thought

The use of ontology in subpopulation comparison involves various dimensions. Ontologies prove beneficial in comprehending and conceptualization. Below are key discussion points highlighting the ways in which ontology can be utilized for comparing subpopulations.

- Conceptual Clarity: Ontologies help define and clarify the concepts related to subpopulations. By establishing a common understanding of terms, attributes, and relationships, ontology ensures clarity in the representation of diverse sub-groups.
- Semantic Interoperability: Ontological representations facilitate semantic interoperability, allowing for the integration of diverse data sources and the comparison of subpopulations across different datasets. This is crucial for ensuring consistency and accuracy in comparisons.
- Granular Attribute Definition: Ontologies allow for the granular definition of attributes associated with subpopulations. This includes demographic information, medical conditions, or any relevant factors. This granularity enhances the precision of comparisons.
- Relationship Modeling: Ontologies capture relationships between entities, enabling the modelling of complex interactions within subpopulations. This is particularly valuable when comparing the influence of different factors on health outcomes or other relevant criteria.
- Automated Inference: Ontologies support automated reasoning and inference, allowing for the deduction of additional information based on the defined relationships. This capability aids in uncovering hidden patterns or correlations within subpopulations.
- Consistent Terminology:Ontologies promote the use of consistent and standardized terminology, reducing ambiguity in the description of subpopulations. Consistency in terminology is crucial for accurate and meaningful comparisons.
- Facilitating Data Integration: Ontologies provide a common framework for integrating data from diverse sources, making it easier to compare subpopulations across different studies or datasets. This promotes a more comprehensive understanding of variations and similarities.
- Enabling Query and Retrieval:Ontologies enhance the efficiency of querying and retrieving relevant information about subpopulations. Researchers can formulate queries using ontological terms, streamlining the comparison process.

While we acknowledge that this list may not be exhaustive, it represents our initial effort to address the multifaceted nature of this complex research. In essence, ontology serves as a powerful tool in subpopulation comparison by offering a structured, standardized, and semantically rich representation of entities and their relationships. This approach contributes to more meaningful, accurate, and efficient comparisons across diverse subsets of a population.

Acknowledgements We would like to thank the physiotherapist at Fysiotherapie Polman in Enschede for sharing important insights on frozen shoulder.

References

1. d_icd_diagnoses. https://mimic.mit.edu/docs/iv/modules/hosp/d_icd_diagnoses/ (Aug 2020), [Online; accessed 17. May. 2022]
2. Bogarín, A., Cerezo, R., Romero, C.: Discovering learning processes using inductive miner: A case study with learning management systems (lmss). Psicothema **30**, 322–329 (Aug 2018). https://doi.org/10.7334/psicothema2018.116
3. Corea, C., Fellmann, M., Delfmann, P.: Ontology-based process modelling-will we live to see it? In: Conceptual Modeling: 40th International Conference, ER 2021, Virtual Event, October 18–21, 2021, Proceedings 40. pp. 36–46. Springer (2021)
4. Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P., Mark, R., Mietus, J., Moody, G., Peng, C., Stanley, H.: PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation [Online] **101**(23), "e215–e220" (2000)
5. Gruber, T.: What is an ontology (1993)
6. Guarino, N., Oberle, D., Staab, S.: What is an ontology? Handbook on ontologies pp. 1–17 (2009)
7. Ivanov, S., Kalenkova, A., Aalst, W.: Bpmndiffviz: A tool for bpmn models comparison **1418**, 35–39 (Jan 2015)
8. Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L.A., Mark, R.: MIMIC-IV (version 1.0) (2021). https://doi.org/10.13026/s6n6-xd98
9. Koorevaar, R., Riet, E., Ipskamp, M., Bulstra, S.: Incidence and prognostic factors for postoperative frozen shoulder after shoulder surgery: a prospective cohort study. Archives of Orthopaedic and Trauma Surgery **137** (Mar 2017). https://doi.org/10.1007/s00402-016-2589-3
10. Mans, R.S., van der Aalst, W.M.P., Vanwersch, R.J.B., Moleman, A.J.: Process mining in healthcare: Data challenges when answering frequently posed questions. In: Lenz, R., Miksch, S., Peleg, M., Reichert, M., Riaño, D., ten Teije, A. (eds.) Process Support and Knowledge Representation in Health Care. pp. 140–153. Springer Berlin Heidelberg, Berlin, Heidelberg (2013). https://doi.org/10.1007/978-3-642-36438-9_10
11. Marazza, F., Bukhsh, F.A., Geerdink, J., Vijlbrief, O., Pathak, S., Keulen, M.v., Seifert, C.: Automatic process comparison for subpopulations: Application in cancer care. International Journal of Environmental Research and Public Health **17**(16). https://doi.org/10.3390/ijerph17165707
12. Munoz-Gama, J., Martin, N., Fernandez-Llatas, C., Johnson, O.A., Sepúlveda, M., Helm, E., Galvez-Yanjari, V., Rojas, E., Martinez-Millana, A., Aloini, D., Amantea, I.A., Andrews, R., Arias, M., Beerepoot, I., Benevento, E., Burattin, A., Capurro, D., Carmona, J., Comuzzi, M., Dalmas, B., de la Fuente, R., Di Francescomarino, C., Di Ciccio, C., Gatta, R., Ghidini, C., Gonzalez-Lopez, F., Ibanez-Sanchez, G., Klasky, H.B., Prima Kurniati, A., Lu, X., Mannhardt, F., Mans, R., Marcos, M., Medeiros de Carvalho, R., Pegoraro, M., Poon, S.K., Pufahl, L., Reijers, H.A., Remy, S., Rinderle-Ma, S., Sacchi, L., Seoane, F., Song, M., Stefanini, A., Sulis, E., ter Hofstede, A.H., Toussaint, P.J., Traver, V., Valero-Ramon, Z., van de Weerd, I., van der Aalst, W.M., Vanwersch, R.,

Weske, M., Wynn, M.T., Zerbato, F.: Process mining for healthcare: Characteristics and challenges. Journal of Biomedical Informatics **127**(103994) (2022). https://doi.org/https://doi.org/10.1016/j.jbi.2022.103994
13. Naguine, P.: Subpopulation Process Comparison and Bottleneck Analysis: A Case Study of Frozen Shoulder. B.S. thesis, University of Twente (2022)
14. Pereira, G., Santos, E., Maceno, M.: Process mining project methodology in healthcare: a case study in a tertiary hospital. Network Modeling Analysis in Health Informatics and Bioinformatics **9** (Dec 2020). https://doi.org/10.1007/s13721-020-00227-w
15. Riesen, K., Fischer, A., Bunke, H.: Improving approximate graph edit distance using genetic algorithms. pp. 63–72 (Aug 2014). https://doi.org/10.1007/978-3-662-44415-3_7
16. Scheuerlein, H., Rauchfuss, F., Dittmar, Y., Molle, R., Lehmann, T., Pienkos, N., Settmacher, U.: New methods for clinical pathways-business process modeling notation (bpmn) and tangible business process modeling (t.bpm). Langenbeck's archives of surgery / Deutsche Gesellschaft für Chirurgie **397**, 755–761 (Feb 2012). https://doi.org/10.1007/s00423-012-0914-z
17. Skobtsov, A., Kalenkova, A.: Efficient algorithms for finding differences between process models. pp. 60–66 (Dec 2019). https://doi.org/10.1109/ISPRAS47671.2019.00015
18. Somodevilla García, M., Vilariño Ayala, D., Pineda, I.: An overview of ontology learning tasks. Computación y Sistemas **22**(1), 137–146 (2018)