Comparing Process Models for Patient Populations: Application in Breast Cancer Care

Francesca Marazza[1], Faiza Allah Bukhsh[2][0000−0001−5978−2754], Onno Vijlbrief[3], Jeroen Geerdink[3][0000−0001−6718−6653], Shreyasi Pathak[2][0000−0002−6984−8208], Maurice van Keulen[2][0000−0003−2436−1372], and Christin Seifert[2][0000−0002−6776−3868]

[1] Università di Genova, Italy
francesca.marazza93@gmail.com
[2] University of Twente, Enschede, Netherlands
[f.a.bukhsh,s.pathak,m.vankeulen,c.seifert]@utwente.nl
[3] Hospital Group Twente (ZGT), Hengelo, Netherlands
[o.vijlbrief, j.geerdink]@zgt.n

Abstract. Processes in organisations such as hospitals, may deviate from intended standard processes, due to unforeseeable events and the complexity of the organisation. For hospitals, the knowledge of actual patient streams for patient populations (e.g., severe or non-severe cases) is important for quality control and improvement. Process discovery from event data in electronic health records can shed light on the patient flows, but their comparison for different populations is cumbersome and time-consuming. In this paper, we present an approach for the automatic comparison of process models extracted from events in electronic health records. Concretely, we propose to compare processes for different patient populations by cross-log conformance checking, and standard graph similarity measures obtained from the directed graph underlying the process model. Results from a case study on breast cancer care show that average fitness and precision of cross-log conformance checks provide good indications of process similarity and therefore can guide the direction of further investigation for process improvement.

Keywords: process mining · process comparison · quality control · breast cancer care.

1 Introduction

Quality of health care can be assessed by observing the structure, processes and the outcomes of healthcare [7]. Processes in organisations, such as hospitals, may deviate from intended standard processes, due to unforeseeable events and the complexity of the organisation. For hospitals, the knowledge of actual patient streams for different patient populations (e.g., severe or non-severe cases) is important for quality control and improvement. Electronic health records (EHR) contain a wealth of information about patients, including timestamps of diagnoses and treatments. Thus, EHR can serve as input data to discover the as-is processes [2] in healthcare. To investigate patient processes for patient populations of interest (e.g., severe or non-severe cases) their process models have to be constructed and compared. Manual comparison of process models requires expertise in understanding process models and is time-consuming, which makes it unfeasible for many populations of interest. In this paper, we present an approach to automatically compare process models obtained from EHR event logs in order to have an initial quantification of the degree of similarity between different patient subgroups. More specifically, we compare three different approaches for obtaining a similarity between process models: i) visual inspection, i.e., human judgment, ii) cross-log conformance checking, and iii) similarity measures on directed graphs extracted from the process model. For cross-log conformance checking, we apply replay technique using the event log of one population against the process model discovered for a second population (and vice versa). Then, we evaluate which of the methods for measuring process similarities best approximate human judgment. The contribution of this paper is the following:

- We present new methods for quantitative process model comparison based on conformance checking and graph similarities
- We evaluate the methods in the domain of breast cancer care to compare different patient populations.

In the following, we first introduce the application domain breast cancer care (cf. section 2), review related work (cf. section 3). We then present the approach in detail (cf. section 4), and its evaluation for breast cancer populations (cf. section 5).

2 Application Background

Electronic health records (EHRs) can be a solution for improving the quality of medical care. EHRs represent digitally collected longitudinal data, like reports, images, sensitive data and clinical information about patients and their provided treatments [13]. The considered EHR of breast cancer patients contains 4 general types of reports. Radiology reports communicate the findings of imaging procedure by describing the radiology images (e.g., X-rays). In case of a patient with a suspicion of breast cancer, it also contains a BI-RAD score. Pathology reports are free medical texts where a diagnosis based on the pathologists examination of a sample of the suspicious tissue is given. The narrative operative surgery reports document breast cancer surgery. The multidisciplinary reports (MDO) are free-texts written during a multidisciplinary expert team meeting.

One of the most important data included in the radiology report is the BI-RADS category. Breast Imaging-Reporting and Data System (BI-RADS) [18] is a classification system proposed by the American College of Radiology (ACR) to represent the malignancy risk of breast cancer of a patient in a standardized manner. A BI-RADS category can range from 0 to 6, with 0 being benign and 6 being the most malignant. Patients with different BI-RADS follow different processes e.g. patients with BI-RADS category 0 need additional imaging evaluation, BI-RADS category 3 needs initial short-interval follow-up and BI-RADS category 4 may be recommended for biopsy. In Netherlands, women between the ages of 50 and 75 are solicited for screening once every 2 years. The purpose of the screening program is to detect the breast cancer at an early stage, before symptoms appear.

3 Related work

Quality in health care and the corresponding reporting and evaluation is an issue of national and international importance. Donabedian [7] proposed that the quality of health care can be assessed by observing the structure, processes and the outcomes of healthcare. The Institute of Medicine (IOM) defines health care quality as "the degree to which health services for individuals and populations increase the likelihood of desired health outcomes and are consistent with current professional knowledge" [10]. Process-based quality measures are more suited to explain how and what is required to improve health care processes, as compared to outcome-based measures [14, 17]. Measuring the quality of various processes can also answer questions like accuracy of diagnosis, disease monitoring and therapy and percentage of patients, who received care as recommended [14]. Better process quality can also lead to better patient satisfaction with the series of transactions occurring during their hospital visit [12]. To summarize, the previous works state that quality of health care can be improved by measuring the health care processes, which can further lead to better health outcome.

The goal of **process mining** is to extract process models from event logs [6], also known as transactional logs or audit trails [1]. An event corresponds to an activity (i.e., a well-defined step in the process) affiliated with a particular case (i.e., process instance) [3] and particularly consist of a time stamp and optional information such as resources or costs. Process mining as a discipline consists of three dimensions, process discovery, conformance checking and process enhancement [19]. *Process discovery* refers to the construction of a comprehensive process model, e.g., Petri-Nets or State-charts, to reproduce the behavior seen in the log file [3]. We use the Inductive Visual Miner(IvM) [8] plug-in of ProM[4] since provides an user-friendly visualization, with an opportunity to investigate deviations. The deviations represent cases which do not follow the most common behaviours and thus correspond to event log traces that the process does not explain. *Conformance checking* is applied to compare process models and event logs in order to find commonalities and discrepancies between the modeled behavior and the observed behavior [16]. Our goal is to compare two processes, therefore we use the (ProM) plug-in Replay a Log on Petri Net for Conformance Analysis to play the event log of one patient population to the discovered process model of another population and use the obtained fitness, precision and generalization measures for our similarity analysis. A case study explored the applicability of **process mining in health care** and raised the concern that

[4] promtools.org last accessed 2019-05-06

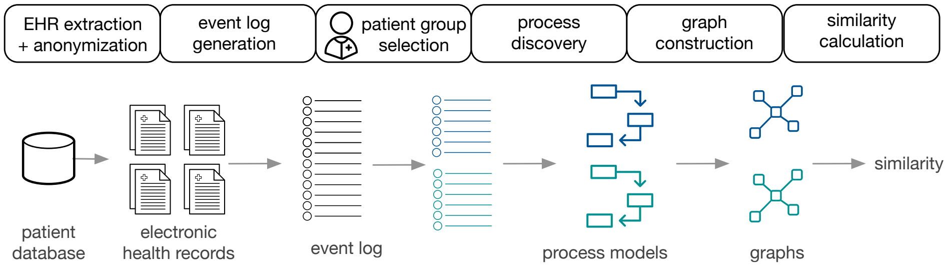


Fig. 1: Overview of the approach.

traditional process mining techniques have problems with unstructured processes as they can be found in hospitals [11]. In this paper, we will focus on a very small sub-domain, breast cancer care, to reduce the complexity of the EHR extraction and resulting process models.

When **comparing two graphs** G_1 and G_2 one is either interested in exact matches of full or sub-graphs (graph homomorphisms) [5] or a measure of structural similarity (e.g. [15]), among which the graph edit distance (GED) [5] is most widely adopted. The GED is defined as the minimum number of operations (add/remove/substitute nodes and edges) needed to transform G_1 to G_2. The problem of calculating the GED is NP-hard in general, making it unfeasible to solve for larger graphs and giving rise to heuristic approximation approaches [20]. More recently, supervised machine learning approaches have been suggested. For instance, Li et. al use a combination of two neural networks to learn a similarity score for G_1 and G_2 [9]. As the authors demonstrate, supervised machine learning approaches can generate highly accurate similarity scores, but they require ground-truth graph-similarity data for training the models. In this work, a graph-similarity ground-truth is not available, therefore we use approximations of GED and an unsupervised machine learning approach based on hand-crafted features for G_1 and G_2 and standard distance metrics on these features.

4 Approach

In this work, we are interested in care processes followed by different patient populations and how these processes compare to each other. In this section we describe our approach for obtaining process models from EHR.

Figure 1 provides an overview of the approach. The available data are EHR, extracted and anonymized from a patient database. Then, event logs are constructed for populations of interest, the corresponding process models are constructed (cf. section 4.1) and transformed to weighted directed graphs 4.2. Two process models are then compared by i) visual inspection using obtaining a similarity measure based on human judgment, ii) cross-log conformance checking and iii) graph comparison methods (cf. section 4.3).

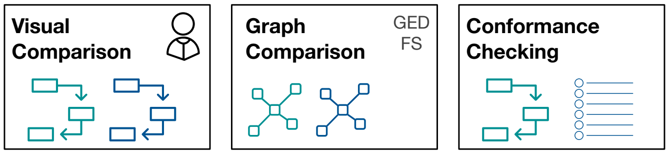


Fig. 2: Overview of methods for pairwise comparison of process models.

4.1 Process Discovery on EHR

In a first step, the EHR of patients subgroups have to be extracted from the hospital data base. From those EHRs, only information related to events of interest have to be extracted. Depending on the hospital data base, this process involves a combination of hand-crafted filtering and extraction rules. To preserve the privacy of hospital patients, personal data has to be anonymized.

In order to apply process mining techniques for analyzing models, data has to be transformed in event logs. EHR, a collection of reports, are our event logs. Each event is associated with a case, a patient, and they are chronologically ordered. Each case can have multiple events. Based on the question that we want to answer, data can be divided in different ways, creating various populations of interest. Consequently, event logs and process models are generated and they should reflect the logs. The algorithm used for producing process models is Inductive Visual Miner, that take care about the deviations. The deviations are patients whose behaviour is different from the most common path.

4.2 Graph Construction

In order to use graph comparison methods, we converted the process models to weighted directed graphs. We constructed $G(V1, V2, E)$ as follows: nodes in $V1$ represent activities and nodes in $V2$ correspond to logical operators in the process model. A directed edge is inserted between two nodes if it in between activities or activities and operators in the underlying process model. We used Boolean function operators to capture the semantic meaning of the process model in the directed graph. Operators can either be AND or XOR, the addition of ”-split” or ”-join” indicates the start and end of the respective paths. Thus, an ”AND-split” means that patients have to follow both paths that the operator outlines, without particular order. The ”AND-join” indicates the end of the paths that must be executed in parallel. The XOR operators work similarly, but states that the patient takes either one path, but not both. ”Loop” indicates a cycle in the process. Each edge has an associated weight that is set to the frequencies of the connection in the process model.

4.3 Process Model Comparison

Figure 2 illustrates the approaches for process model comparison.

For **visual similarity assessment** of pairs of process models were judged by humans for their similarity. We used a 5-point Likert scale (0: Identical, 1: Slightly Different, 2: Somewhat Different, 3: Very Different, 4: Extremely Different). 3 of the co-authors (no medical background) of this paper judged pairs of process models independently, and were given the instruction to focus on the structure of the process and not on the numbers of the edges when assessing the similarity.

We applied **cross-log conformance checking** as follows: we used the (ProM) plug-in Replay a Log on Petri Net for Conformance Analysis to play the event log of one subpopulation with the process model generated by the second subgroup (and vice versa) and recorded standard conformance checking metrics fitness, precision and generalization.

For the **graph-based comparison** of process models, we used the networkx graph library for python for calculation graph metrics and similarities[5]. On the weighted directed graphs constructed from the process models (see section 4.2), we calculate the graph edit distance (GED) using an approximation algorithm [4]. For feature-based comparison of two graphs, we generated a feature vector for each graph with the following graph metrics as features: number of nodes, number of edges, average degree, average weighted degree, average clustering coefficient, average shortest path, average closeness and average betweenness centrality. We then obtained the similarity score for two graphs by first normalizing the feature vectors to unit length and then calculating the Euclidean distance of the two normalized vectors. The similarity score is then 1 minus the obtained distance.

5 Experiments

In our experiments, we collected EHRs from the hospital database (cf. section 5.1, created event logs for populations of interest (cf. section 5.2), and obtained process models for these populations (cf. section 5.3). We then compared these process models pairwise, by i) visual inspection, ii) cross-log conformance checking and iii) graph-based similarity measures and investigate how the obtained similarity measures reflect the similarity obtained by human judgment (cf. section 5.4).

5.1 EHRs Extraction and Event Log Preparation

Free-text reports on breast cancer patients from 2012 – 2018 were collected from the hospital database. The following rules defined whether a patient was included in the analysis: with the purpose of identifying the complete health path of the patients inside the hospital, patient has to be a “new patient”: in the range of time considered, the patient must have the first visit, that can not be described in the referring report by key words like ”MRI”, ”punctie” (biopsy),

[5] https://networkx.github.io/, last accessed 2019-05-20

Table 1: Overview of the selected sub-populations.

			#Reports in EHR			
Population	**#Events**	**#Cases**	**Radiology**	**Pathology**	**MDO**	**Surgery**
SVOB	17,677	5,793	10,429	6,987	199	62
NoSVOB	26,542	6,427	15,254	10,208	784	296
Age≥ 50	31,157	7,740	18,132	12,894	819	312
Age< 50	12,062	4,480	7,551	4,301	164	46
Birad12	23,393	8,612	15,356	7,874	131	32
Birad3-6	20,019	3,365	9,805	9,041	849	324

”mammatumor” (tumor in the breast). Also, it can not be represented by a MDO or Surgery report, because it is impossible that the first visit of a patient is one of that. So, the start event is the first report for each patient. On the other hand, the criteria to understand if a patient has finished the treatments in the hospital was not found due to the complexity of the problem. Therefore, there is no end condition point in health path analysis. The gathered patients are 12220, for a total of 44219 reports.

5.2 Patient Populations

We considered the following patient populations:

SVOB: Patients coming from a national breast cancer screening program,
NoSVOB: Patients sent to the hospital by the general practitioner,
Birad12: Patients with a BIRAD score 1 or 2 (0% likelihood of cancer),
Birad3-6 Patients with a BIRAD score of 3 (probably benign), 4 (suspicious), 5 (highly suggestive of malignancy) or 6 (known biopsy-proven).
Age ≥50: Patients of age 50 or older,
Age<50: Patients younger than 50.

The age boundary is set to 50 due to an empirically increased risk for breast cancer development at this age.

Table 1 provides an overview for the selected populations. It shows the number of events with a breakdown on event type, i.e. the specific type of report and the number of cases, i.e., the number of patients. For each population, the event log contains at least 12,000 events, with radiology reports being the most frequent and surgery being the least frequent event type in all populations. Note that radiology report is always present for each case. This result will be confirmed graphically by process figures where an AND condition is always existing for radiology events. Surgery events are generally occurring less frequent than the others event types. In particular, the percentage of surgeries for Birad12 and Birad3-6 populations are significantly different.

Table 2: Quantitative population comparison. Reporting Jaccard similarity of sets of patient ids in the two groups.

Group 1	**Group 2**	**Jaccard Similarity**
SVOB	NoSVOB	0.00
Age$\geq$ 50	Age<50	0.00
Birad12	Birad3-6	0.00
NoSVOB	Age<50	0.30
SVOB	Birad12	0.42

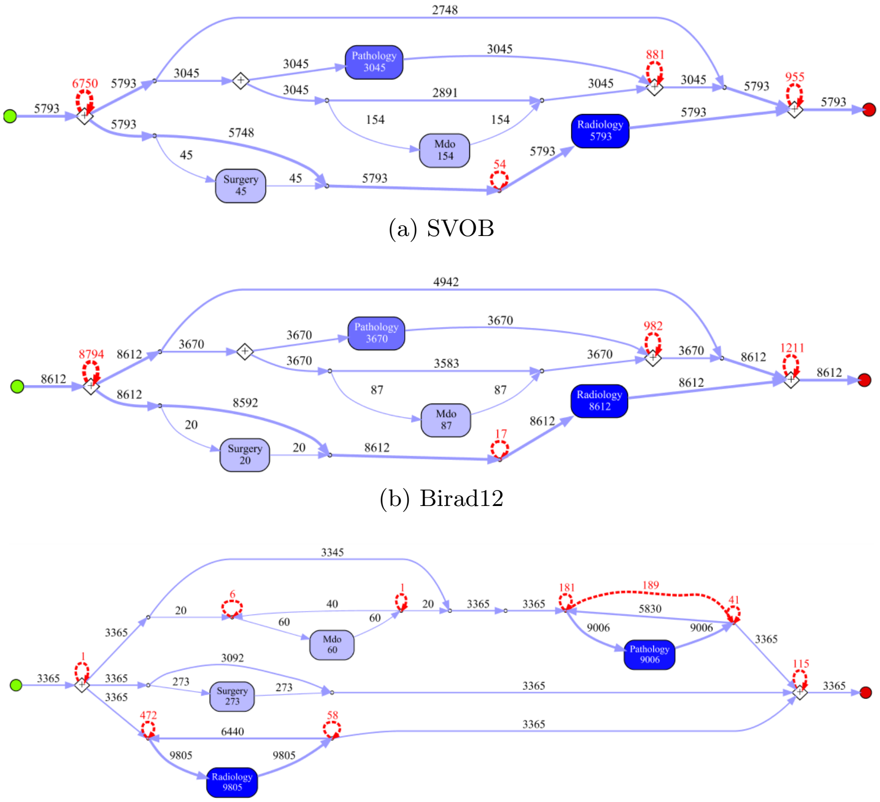
We compared process models of populations that are of clinical interest, namely screening vs. non-screening patients (SVOB/NoSVOB), low vs. high probability of malignancy (Birad12 vs. Birad3-6) and age groups before and after screening age (Age$\geq$ 50 vs. Age$<$ 50). We also included a comparison of before screening age patients and non-screening patients (NoSVOB vs. Age$<$ 50) and screening patients that were transferred to the hospital, but had a low probability of malignancy (SVOB vs. Birad12). Table 2 gives an overview of the compared groups and shows the Jaccard similarity for different populations pairs. To investigate the patient overlap between these populations, we calculated the Jaccard coefficient as follows: for each population, we created a set with anonymized patient ids. The Jaccard coefficient is the ratio of the number of patients two populations have in common (set intersection) and the total number of unique patient ids (set union). As can be seen, there is a 30% overlap between NoSVOB patients and patients younger than 50 as well as between SVOB patients and patients with low likelihood of cancer (Birad12).

5.3 Process Discovery and Graph Construction

Process models obtained by the IvM plug-in of ProM (noise filtering set to 90%) for three populations are shown in figures 3 and 4 (top)[6]. Blue rectangles represent events, the intensity of colour is associated with the number of events. It can be seen that the process models for SVOB and Birad12 have the same structure (but different frequencies) although their patient populations are not the same (Jaccard similarity of 0.42, cf. table 2). The process model for Birad3-6 patients has a complicated structure, indicating that patients with non-zero probability of malignancy follow complex paths in the hospital. Many deviations, represented by red dashed lines, exist in all processes.

Figure 4 shows the process model for NoSVOB patients and the correspondent translated graph. Circles denote the logic operators while rectangles represent events. The un-normalized feature vector f obtained from the graph for NoSVOB patients is $f = (18, 28, 2.89, 3.09, 0.19, 0.81, 0.13, 0.03)$, the features are in the order as described in section 4.3.

[6] Process models for the other populations were omitted due to space constraints.



(a) SVOB

(b) Birad12

(c) Birad3-6

Fig. 3: Discovered process models for 3 sub-populations.

5.4 Process Model Comparison

Table 3 summarizes the comparison of process models. For visual assessment, we reported the average similarity score and standard deviation obtained by human judgment on a 5-point Likert scale (0 - very similar). Further we reported the fitness and precision values obtained by cross-log conformance checking and their averages. We also reported the GED and the similarity obtained by feature-based graph comparison (FS). As FS values are generally all above 0.9 and similar, they do not seem useful for comparing the underlying process models. Fitness and precision varies more and interestingly differs depending on whether the log of group 1 is played on the process model of group 2 or vice versa.

Figure 5 shows the Pearson correlations between the similarity measures for both, the original conformance-checking values and their averages. The GED shows the highest positive correlation with the visual judgment ($\rho = 0.95$) and the average fitness is also strongly correlated ($\rho = 0.67$). We observed a strong negative correlation between average precision and human judgment ($\rho = 0.82$).

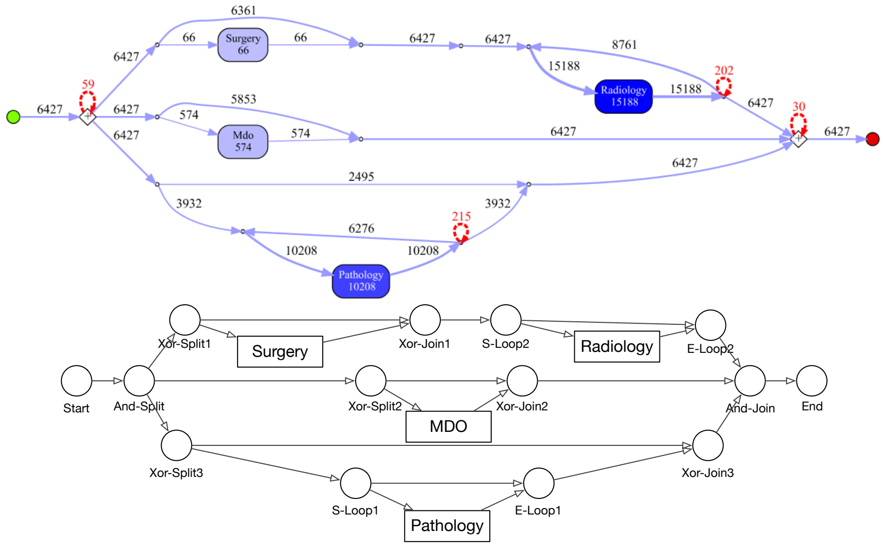


Fig. 4: Process model for NoSVOB population (top) and corresponding directed graph (bottom). Edge weights and loops omitted in the graph for readability.

From this, we can conclude that at least for the tested event logs and process models, cross-log conformance checking measures provide an indication of process similarity, the higher the average fitness, the more similar the processes and the lower the average precision, the more dissimilar processes are. While the GED provides the highest correlation, the results need to be judged with care. We used an iterative algorithm for calculating the GED, and stopped the approximation after 1 hour of calculation if the values did not change anymore. This means, i) reported values might not be the true GED (and are for certain not in the case of SVOB - Birad12 comparison, where GED should be 0) and ii) this approach is not practical due to its computational inefficiency.

6 Summary

In this paper, we raised the problem of automatic quantitative comparison of process models generated from event logs with the same types of events. We proposed comparisons based on cross-log conformance checking and standard graph similarity measures obtained from the directed graph underlying the process model. We applied these methods on process models from different subgroups of breast cancer care patients. Results show that average fitness and precision of cross-log conformance checking provide good indications of process similarity and therefore can guide the direction of further investigation for process improvement. In our application scenario, the compared process models were rather small; they contained maximally 4 different event types. At a higher level

Table 3: Process comparison. Comparing visual impression (Visual), Process conformance checking metrics (F - Fitness, P - Precision) when group 1 event log is checked against the process model (PM) constructed for group 2 and vice versa. Graph similarities (GED - graph edit distance, FS - feature-based similarity)

Populations			**Conformance Checking**						**Graph Sim.**	
			Log1-PM2		Log2-PM1		Average			
Group 1	Group 2	Visual	F_{12}	P_{12}	F_{21}	P_{21}	$\overline{F}$	$\overline{P}$	GED	FS
SVOB	NoSVOB	3 ± 0	1.00	0.47	0.71	0.79	0.86	0.63	16	0.97
Age≥ 50	Age<50	4 ± 0	0.69	0.94	1.00	0.43	0.85	0.69	30	0.96
Birad12	Birad3-6	3 ± 1	0.81	0.82	0.96	0.52	0.89	0.67	24	0.90
NoSVOB	Age<50	3 ± 1	0.65	0.94	1.00	0.47	0.83	0.71	24	0.95
SVOB	Birad12	0 ± 0	0.80	0.82	0.78	0.79	0.79	0.81	3	0.98

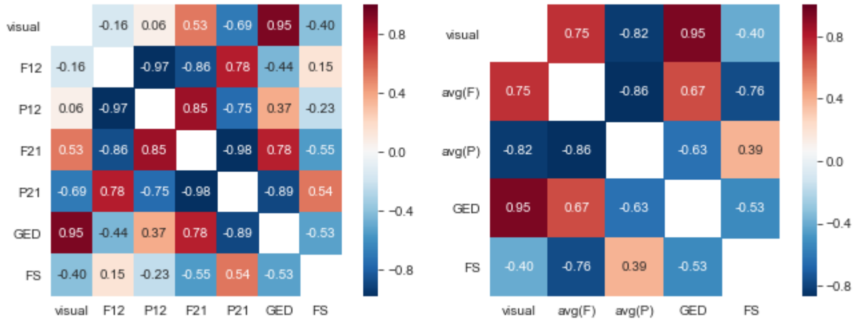


Fig. 5: Correlation between similarity measures. Based on the measures for the variables reported in table 3.

of event granularity (e.g. when differentiating between different types of radiology reports, such as Diagnostic, Biopsy) process models easily get more complex and thus automatic comparison becomes even more desirable. While the results indicate that cross-conformance checking metrics are indicative for process model similarity, further research in different domains and/or different event types in the same domain remains to be done.

Acknowledgement This work was supported by the Hospital Group Twente (ZGT) by providing data, secure server infrastructure and domain advice.

References

1. Van der Aalst, W.M., van Dongen, B.F., Herbst, J., Maruster, L., Schimm, G., Weijters, A.J.: Workflow mining: A survey of issues and approaches. Data & knowledge

engineering **47**(2), 237–267 (2003)

2. van Aalst, W.M., van Hee, K.M., van Werf, J.M., Verdonk, M.: Auditing 2.0: Using process mining to support tomorrow's auditor. Computer **43**(3), 90–93 (2010)
3. Van der Aalst, W.M., Weijters, A.: Process mining: a research agenda (2004)
4. Abu-Aisheh, Z., Raveaux, R., Ramel, J.Y., Martineau, P.: An Exact Graph Edit Distance Algorithm for Solving Pattern Recognition Problems. In: 4th Int Conf on Pattern Recognition Applications and Methods 2015. Lisbon, Portugal (Jan 2015). https://doi.org/10.5220/0005209202710278
5. Berretti, S., Del Bimbo, A., Vicario, E.: Efficient matching and indexing of graph models in content-based retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **23**(10), 1089–1105 (Oct 2001). https://doi.org/10.1109/34.954600
6. Bogarn, A., Cerezo, R., Romero, C.: A survey on educational process mining. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **8** (09 2017). https://doi.org/10.1002/widm.1230
7. Donabedian, A.: Evaluating the quality of medical care. The Milbank memorial fund quarterly **44**(3), 166–206 (1966)
8. Leemans, S.J.J., Fahland, D., van der Aalst, W.M.P.: Discovering block-structured process models from event logs - a constructive approach. In: Petri Nets (2013)
9. Li, Y., Gu, C., Dullien, T., Vinyals, O., Kohli, P.: Graph Matching Networks for Learning the Similarity of Graph Structured Objects. In: Proc. Intl. Conference on Machine Learning (2019)
10. Lohr, K.N., Schroeder, S.A.: A strategy for quality assurance in medicare. New England Journal of Medicine **322**(10), 707–712 (1990)
11. Mans, R.S., Schonenberg, M., Song, M., van der Aalst, W.M., Bakker, P.J.: Application of process mining in healthcare–a case study in a dutch hospital. In: International joint conference on biomedical engineering systems and technologies. pp. 425–438. Springer (2008)
12. Marley, K.A., Collier, D.A., Meyer Goldstein, S.: The role of clinical and process quality in achieving patient satisfaction in hospitals. Decision Sciences **35**(3), 349–369 (2004)
13. Noumeir, R., Pambrun, J.F.: Images within the electronic health record. pp. 1761 – 1764 (12 2009). https://doi.org/10.1109/ICIP.2009.5414545
14. Palmer, R.H.: Process-based measures of quality: the need for detailed clinical data in large health care databases. Annals of Internal Medicine **127**(8_Part_2), 733–738 (1997)
15. Raymond, J.W., Gardiner, E.J., Willett, P.: RASCAL: Calculation of Graph Similarity using Maximum Common Edge Subgraphs. The Computer Journal **45**(6), 631–644 (01 2002). https://doi.org/10.1093/comjnl/45.6.631
16. Rozinat, A., van der Aalst, W.M.P.: Conformance checking of processes based on monitoring real behavior. Inf. Syst. **33**(1), 64–95 (2008)
17. Rubin, H.R., Pronovost, P., Diette, G.B.: The advantages and disadvantages of process-based measures of health care quality. International Journal for Quality in Health Care **13**(6), 469–474 (2001)
18. Sickles, E.A., D'Orsi, C.J., Bassett, L.W., et al: ACR BI-RADS Atlas. American College of Radiology, Reston, VA (2013)
19. Van Der Aalst, W.: Process mining: discovery, conformance and enhancement of business processes, vol. 2. Springer (2011)
20. Zeng, Z., Tung, A.K.H., Wang, J., Feng, J., Zhou, L.: Comparing stars: On approximating graph edit distance. Proc. VLDB Endow. **2**(1), 25–36 (Aug 2009). https://doi.org/10.14778/1687627.1687631