J. Mantas et al. (Eds.)

doi:10.3233/SHTI210279

Process Mining to Explore Variations in Endometrial Cancer Pathways from GP Referral to First Treatment

Angelina Prima KURNIATI[a,1], Eric ROJAS[b], Kieran ZUCKER[c], Geoff HALL[c], David HOGG[d] and Owen JOHNSON[d]
[a] *School of Computing, Telkom University, Bandung, Indonesia*
[b] *School of Medicine, Dept of Clinical Lab, Pontificia Universidad Católica de Chile, Chile*
[c] *School of Medicine, University of Leeds, Leeds Teaching Hospitals NHS Trust, UK*
[d] *School of Computing, University of Leeds, Leeds, UK*

Abstract. The main challenge in the pathway analysis of cancer treatments is the complexity of the process. Process mining is one of the approaches that can be used to visualize and analyze these complex pathways. In this study, our purpose was to use process mining to explore variations in the treatment pathways of endometrial cancer. We extracted patient data from a hospital information system, created the process model, and analyzed the variations of the 62-day pathway from a General Practitioner referral to the first treatment in the hospital. We also analyzed the variations based on three different criteria: the type of the first treatment, the age at diagnosis, and the year of diagnosis. This approach should be of interest to others dealing with complex medical and healthcare processes.

Keywords. Process mining, care pathway, endometrial cancer, 62-day wait

1. Introduction

Cancer treatment is complicated due to the high variability in the nature of this disease [1]. One important performance indicator for cancer treatment is cancer waiting time [2], including, in the UK, a target that patients should receive their first definitive treatment for suspected cancer within 62-day after referral. This study focuses on the 62-day wait in endometrial cancer [3], as one example of common cancer pathway.

Our data source was Patient Pathway Manager (PPM), a large and comprehensive EHR that has been used to record cancer patient treatments for over a decade in the Leeds Teaching Hospital NHS Trust (LTHT), integrating data from multiple systems within the Trust [4] [5]. Each patient event is recorded with patient identification and a timestamp, which makes it possible to extract the historical pathways for every patient. Our earlier literature reviews [6,7] identified successful use of process mining in healthcare, including in cancer. Challenges include complexity caused by the multidisciplinary nature of healthcare, heterogeneity, and process changes [8]. In earlier work, we used

[1] Corresponding Author, Angelina Prima Kurniati, Panambulai Building 108, School of Computing, Telkom University, Bandung, Indonesia, 40257; Email: angelina@telkomuniversity.ac.id.

process mining to analyze cancer treatment in an open-access data (MIMIC-III) from the USA [9,10] and real data from PPM in the UK [5]. In this study, we extracted PPM data to analyze endometrial cancer patient pathway variants concerning the 62-day target.

2. Methods

Our method follows the question-driven methodology, developed in [11] and designed to answer specific research questions posed by clinicians (see Figure 1).

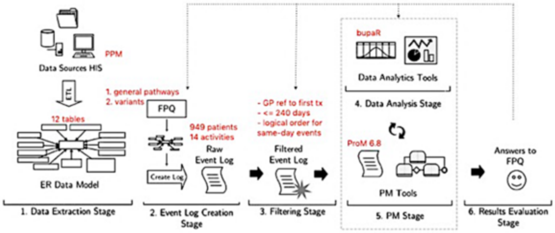


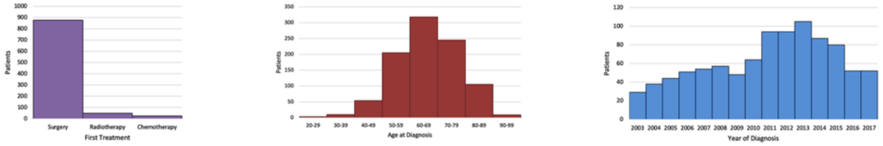
Figure 1. General methodology (based on [11]). The text in red shows the specific aspects for this study.

The case study aimed to examine process mining applicability and usefulness to answer our two research questions, which are: *"RQ-1. What is the general pathway of endometrial cancer from GP referral to the first treatment?"* and *"RQ-2: Are there differences in care paths followed by different patient groups?"*.

Data Extraction and ***Event Log Creation***. Data was extracted using R code with embedded SQL queries to the PPM database. Patient data came from 12 out of 43 tables, including *Admissions, Investigations, Chemocycles, Referrals*, and *Surgery*. For each patient, events between the date of GP referral and the first treatment were included. The ***Filtering stage*** was done in several iterations through queries in R code [12] and ProM plugins [13]. We separated Surgery from Diagnostic Surgery and set a logical order of same-day activities. We included the following 14 activities: *Admission, Investigation, Chemotherapy, Consultation, Diagnosis, MDTReview, Outpatient, Pathology, Death, Radiotherapy, Referral, Diagnostic Surgery, Surgery,* and *Discharge* (question-driven filtering). The ***Data analysis stage*** characterised the data and discovered different patterns in the event logs. The ***Process mining stage*** addressed RQ-1 using the Inductive Miner infrequent algorithm [14]. Process analytics addressed RQ-2 by comparing process models from the patient groups using four Comparison Points (CPs) posed by clinical experts: *(CP1) Proportion of patients based on their first treatment, age at diagnosis, and year of diagnosis; (CP2) The variations of process based on the first treatment; (CP3) The variations based on the age at diagnosis;* and *(CP4) The variations based on the year of diagnosis.*

3. Results

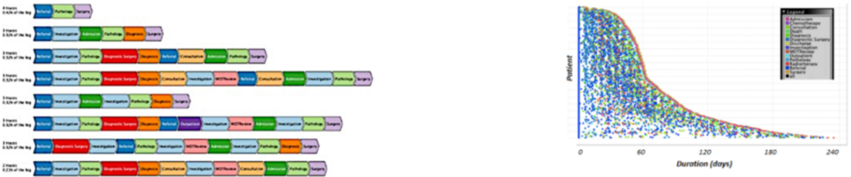
Data analysis stage. There were 949 patients included in the event log, with 17,413 events in total. The treatment duration ranged from 5 to 238 days (median = 62, mean = 79). There were 921 variants out of 949 patients, showing very high variability between individual pathways. Figure 2 summarises CP1 (Proportion of patient groups).



(a) based on the first treatment *(b) based on the age at diagnosis* *(c) based on the year of diagnosis*

Figure 2. Proportion of patients in each group. It shows that (a) the most common treatment is surgery, (b) patients were mostly diagnosed in their 50s-70s, and (c) the trend was increasing over years.

Process mining stage. The trace variants are presented in Figure 3. It shows the high variability of the traces, reflects the complexity of cancer treatment in real life and provides supportive evidence of the complexity of the pathway. Figure 4 shows the simplified process model. The model is highly representative of the general pathway with trace fitness = 0.83, precision = 0.8, and generalisation 0.995.



(a)trace variants *(b) dotted chart*

Figure 3. (a) Top 10 out of 910 trace variants of pathways from GP referral to the first treatment. (b) Dotted chart showing that around a half top of the patients had duration <= 62 days (the expected duration).

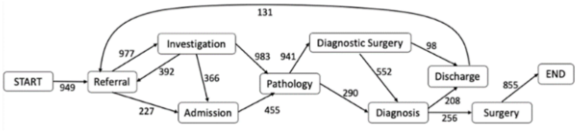
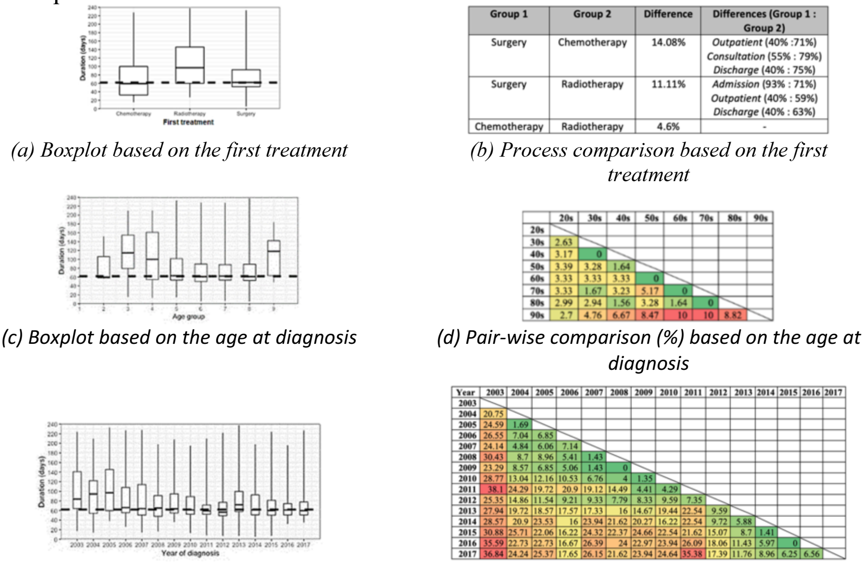


Figure 4. Process model (produced using ProM 6.8 plugin of iDHM, minimum frequency 0.1, redrawn to improve clarity). Boxes represent the activities. Arcs represent the flow from an activity to another with number shows the frequency of the flow. Not presented are the infrequent activities (Chemotherapy, Radiotherapy, and Death) and the highly frequent activities (Consultation, MDT Review, and Outpatient).

Variations of pathways based on the first treatment (CP2), the age at diagnosis (CP3), and the year of diagnosis (CP4). We compared the grouped treatment durations using boxplots and Process Comparator plugin in ProM 6.8 to produce Figure 5 which shows (a) similar median durations for patients with Chemotherapy and Surgery, and a longer duration for patient with Radiotherapy, (b) the most significant differences on the Outpatient, Consultation, Admission, and Discharge activities, (c) similar median durations for patients diagnosed at their 50s to 80s, (d) group 90s as the group with the most significant difference to other groups, (e) potentially possible changes in 2005-2006 and in 2013 which might have impacted the treatment duration, and (f) group 2003 as the most significant difference compared to others.

Result evaluation stage. The results were discussed with clinical experts at each stage of the study. In the *Data Extraction* stage, experts confirmed the 12 tables included in this study containing events in patient treatments. In the *Event Log Creation*, clinicians suggested that the *GP referral* was the referral from a GP to one of four oncology specialists; while the first treatments were Surgery, Chemotherapy, or Radiotherapy. In the *Filtering* stage, clinicians recommended separating diagnostic and therapeutic surgeries, based on the main procedure label. In the *Data Analysis* stage, clinical experts suggested to include only patients having a pathway duration of no more than 240 days, because a longer time period would be too long for the events to be related. In the *Process Mining* stage, clinical experts suggested to handle the same-day events by following a logical sequence of activities.



(a) Boxplot based on the first treatment

Group 1	Group 2	Difference	Differences (Group 1 : Group 2)
Surgery	Chemotherapy	14.08%	*Outpatient* (40% :71%) *Consultation* (55% : 79%) *Discharge* (40% : 75%)
Surgery	Radiotherapy	11.11%	*Admission* (93% : 71%) *Outpatient* (40% : 59%) *Discharge* (40% : 63%)
Chemotherapy	Radiotherapy	4.6%	-

(b) Process comparison based on the first treatment

(c) Boxplot based on the age at diagnosis

	20s	30s	40s	50s	60s	70s	80s	90s
20s								
30s	2.63							
40s	3.17	0						
50s	3.39	3.28	1.64					
60s	3.33	3.33	3.33	0				
70s	3.33	1.67	3.23	5.17	0			
80s	2.99	2.94	1.56	3.28	1.64	0		
90s	2.7	4.76	6.67	8.47	10	10	8.82	

(d) Pair-wise comparison (%) based on the age at diagnosis

(e) Boxplot based on the year of diagnosis

Year	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
2003															
2004	20.75														
2005	24.59	1.69													
2006	26.55	7.04	6.85												
2007	24.14	4.84	6.06	7.14											
2008	30.43	8.7	8.96	5.41	1.43										
2009	23.29	8.57	6.85	5.06	1.43	0									
2010	28.77	13.04	12.16	10.53	6.76	4	1.35								
2011	38.1	24.29	19.72	20.9	19.12	14.49	4.41	4.29							
2012	25.35	14.86	11.54	9.21	9.33	7.79	8.33	9.59	7.35						
2013	27.94	19.72	18.57	17.57	17.33	16	14.67	19.44	22.54	9.59					
2014	28.57	20.9	23.53	16	23.94	21.62	20.27	16.22	22.54	9.72	5.88				
2015	30.88	25.71	22.06	16.22	24.32	22.37	24.66	22.54	21.62	15.07	8.7	1.41			
2016	35.59	22.73	22.73	16.67	26.39	24	22.97	23.94	26.09	18.06	11.43	5.97	0		
2017	36.84	24.24	25.37	17.65	26.15	21.62	23.94	24.64	35.38	17.39	11.76	8.96	6.25	6.56	

(f) Pair-wise comparison (%) based on the year of diagnosis

Figure 5. Comparison of the treatment duration. For each boxplot, a box shows the first quantile, median and third quantile; lines show the variability from the minimum to the maximum duration; while the dashed line shows target duration (62 days). For each pair-wise difference matrix, percentage is color-coded from red (the highest) to green (the lowest) differences and alpha significant level is 5%.

4. Discussion

Process mining perspective. The event log was extracted from the PPM database and filtered to improve the data quality. Patients were grouped based on their first treatment, their age at diagnosis, and the year of diagnosis; to understand the variations of the pathway. Intensive discussions between process miners and clinical experts are crucial for the success of process analytics. The question-driven methodology followed in this study provides good step-by-step guidance based on the pre-defined research questions. A straight-forward process mining approach would not be as useful as an iterative approach based on the feedback gathered from clinical experts.

Clinical perspective. Based on the question-driven methodology, clinical questions were the starting points of the analyzes to drive the direction of the study. Clinical experts were involved through discussion and evaluation throughout the stages. The process models that were discovered were found to be useful for intensive discussions with clinical experts. One limitation of the general process model (Figure 4) was that Chemotherapy and Radiotherapy were infrequent and were not presented. Clinical expert also confirmed several facts revealed in the results, including: Surgery was the central treatment in the pathway; patients diagnosed at their 90s are mostly treated with Radiotherapy; and PPM system was not recording complete events in 2003.

5. Conclusions

We used a real life hospital EHR to analyze a 62-day wait endometrial cancer pathway. The question-driven methodology supported an effective exploratory case study guided by the clinicians' questions. The general pathways from GP referral to the first treatment (RQ-1) produced an effective process model supported by trace variant and dotted chart; all showing the high variability of the pathway. The main challenge was to analyze variability in care pathways (RQ-2). This was approached by splitting the event log based on: first treatment, age at diagnosis, and year of diagnosis. The analysis showed that: Surgery was the most common first treatment; the most common age range at diagnosis was between 50s to 80s; and the change over time that was affected by the evolving EHR system. A limitation of this study was the need for expert clinical to categorize surgery types and to set the logical order for same-day events. Future work could automate those two activities with machine learning algorithms or statistical approaches. Another area of further work is to do more detailed analysis, such as time between events, critical events causing treatment delays, and the change over time affecting the treatment process.

Acknowledgment

This research is supported by the Yorkshire and the Humber NIHR Patient Safety Translational Research Centre and the Indonesia Endowment Fund for Education (LPDP). Data access: Health Research Authority (HRA) Approval Number 206843.

References

[1] Jemal A, Bray F, Center MM, et al. Global cancer statistics. CA Cancer J Clin [Internet]. 2011 [cited 2014 Jul 12];61:69–90. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21296855.

[2] National Health System. Delivering Cancer Waiting Times: A Good Practice Guide. 2015;0–67.

[3] National Cancer Institute. Uterine Cancer - Patient Version [Internet]. 2018 [cited 2007 Aug 20]. Available at: https://www.cancer.gov/types/uterine/patient/endometrial-treatment-pdq.

[4] Newsham A, Johnston C, Hall G. Development of an advanced database for clinical trials integrated with an electronic patient record system. Comput Biol Med. 2011;41:575–586.

[5] Baker K, Dunwoodie E, Jones RG, et al. Process mining routinely collected electronic health records to define real-life clinical pathways during chemotherapy. Int J Med Inform. 2017;103:32–41.

[6] Rojas E, Munoz-Gama J. Process mining in healthcare: A literature review. J Biomed Inform. 2016;61:224–236.

[7] Kurniati AP, Johnson O, Hogg D, et al. Process mining in oncology: A literature review. Proc 6th Int Conf Inf Commun Manag ICICM 2016; 2016. p. 291–297.

[8] Homayounfar P. Process Mining Challenges in Hospital Information Systems. Proc Fed Conf Comput Sci Inf Syst; 2012. p. 1135–1140.

[9] Kurniati AP, Hall G, Hogg D, et al. Process Mining in Oncology using the MIMIC-III Dataset. IOP J Phys Conf Ser 971. 2018;971:1–10.

[10] Kurniati AP, Rojas E, Hogg D, et al. The assessment of data quality issues for process mining in healthcare using MIMIC-III , a publicly available e-health record database. Health Informatics J. 2018;

[11] Rojas E, Sepúlveda M, Munoz-Gama J, et al. Question-Driven Methodology for Analyzing Emergency Room Processes Using Process Mining. Appl Sci. 2017;7:302.

[12] Janssenswillen G. bupaR: Business Process Analysis in R [Internet]. R Packag. version 0.4.2. 2019. p. 1–34. Available at: https://cran.r-project.org/package=bupaR.

[13] van der Aalst WMP, Van Dongen BF, Gunther C, et al. ProM: The process mining toolkit. CEUR Workshop Proc. 2009;489.

[14] Leemans SJJ, Fahland D, van der Aalst WMP. Discovering block-structured process models from event logs containing infrequent behaviour. Int Conf Bus Process Manag. Springer, Cham; 2013. p. 66–78.