Combining business process and data discovery techniques for analyzing and improving integrated care pathways

Jonas Poelmans[1], Guido Dedene[1,4], Gerda Verheyden[5], Herman Van der Mussele[5], Stijn Viaene[1,2], Edward Peters[1,3]
[1]K.U.Leuven, Faculty of Business and Economics, Naamsestraat 69, 3000 Leuven, Belgium
[2]Vlerick Leuven Gent Management School, Vlamingenstraat 83, 3000 Leuven, Belgium
[3]OpenConnect Systems, 2711 LBJ Freeway Suite 700, Dallas, TX 75234, United States of America
[4]Universiteit van Amsterdam Business School, Roetersstraat 11 1018 WB Amsterdam, The Netherlands
[5]Sint-Augustinus hospital, Oosterveldlaan 24, 2610 Wilrijk, Belgium
{Gerda.Verheyden, Herman.Vandermussele}@gza.be
{Jonas.Poelmans, Stijn.Viaene, Guido.Dedene}@econ.kuleuven.be
epeters@oc.com

Abstract. Hospitals increasingly use process models for structuring their care processes. Activities performed to patients are logged to a database but these data are rarely used for managing and improving the efficiency of care processes and quality of care. In this paper, we propose a synergy of process mining with data discovery techniques. In particular, we analyze a dataset consisting of the activities performed to 148 patients during hospitalization for breast cancer treatment in a hospital in Belgium. We expose multiple quality of care issues that will be resolved in the near future, discover process variations and best practices and we discover issues with the data registration system. For example, 25 % of patients receiving breast-conserving therapy did not receive the key intervention "revalidation". We found this was caused by lowering the length of stay in the hospital over the years without modifying the care process. Whereas the process representations offered by Hidden Markov Models are easier to use than those offered by Formal Concept Analysis, this data discovery technique has proven to be very useful for analyzing process anomalies and exceptions in detail.

Keywords: Breast cancer, process mining, data discovery, integrated care pathways

1 Introduction

An increasingly competitive health care market forces hospitals to search for ways to improve their processes in order to deliver high quality of care while at the same time reducing costs [1]. According to [14], the solution to poor quality is not to increase

the supply of physicians or specialists or hospital beds, but instead to improve health care systems and incentives to ensure that existing physicians and hospitals provide the best possible quality at the lowest cost. Integrated care pathways are structured multi-disciplinary care plans which detail the essential steps in the care process of a population of patients with a certain clinical problem [3]. The aims to achieve with care pathways are improving quality and efficiency of care, to standardize the outcomes of the provided care, to facilitate communication between healthcare professionals and to allow for systematic continuing audit. Care pathways are business process models which describe the expected progress of the patient through the care process and try to model the most standard frequent care pathway, based on expert prior knowledge.

Till date, the continuous monitoring, analysis and improvement of the care pathway's performance was performed in an ad hoc, manual and labor-intensive way. This approach however has some limitations. Modifications to the care process are performed in an ad hoc way and their success can only be measured by the impact of these modifications on the Key Performance Indicators (KPIs). This retrospective impact analysis can only be done after several months, which is an unacceptable long time window in healthcare management. Moreover, this standard model does not capture process variations, nor process exceptions and the root causes for inefficiencies are not known. Moreover, in practice there is often a significant gap between what is prescribed or supposed to happen and what actually happens. Process mining is an interesting method for gaining insight into what happens in a healthcare process for a group of patients with the same diagnosis.

In [6] the applicability of process mining in the healthcare domain was investigated, using Petri-Nets. The idea of process mining [12] is to extract, monitor and improve real processes by extracting knowledge from event logs.

In this paper, we use a unique combination of process discovery techniques and data discovery techniques to gain a deeper understanding of an existing breast cancer care process and the actual activities performed on the working floor to discover process inefficiencies, exceptions and variations immediately and to search for the root causes of inefficiencies. We propose and use a new approach based on Hidden Markov Models to discover a process model from event sequences. Formal concept Analysis (FCA) is used to analyze the characteristics of the clusters of patients that emerged from this process discovery exercise and vice versa to find groups of patients to feed into the process discovery methods.

The remainder of this paper is composed as follows. In section 2 we introduce the essentials of business process discovery, Hidden Markov Models and the HMM-based techniques that are proposed for process discovery. In section 3, we elaborate on FCA as a data discovery technique. In section 4, we discuss the dataset used. Section 5 describes the methodology and the results of our discovery exercise. Finally, section 6 rounds up with conclusions.

2. Business process discovery

In contrast to process modeling, which is developing a top-down representation of a "to-be" process reality, process discovery is a bottom up approach that tries to gain an understanding of the as-in process realities that are existing at the operational work floor. Discovering irregularities, exceptions and variations by means of analytics is essential in developing process and workforce intelligence. Statistical techniques often consider exceptions as nuisance information and eliminate them as noise. According to [8], statistical techniques are able to capture the general process model rather than the process model containing exceptional paths. For discovering process exceptions, anomalies and variations, the combination of learning techniques, mining and clustering is required to gain sufficient insights in the processes. Most workflow mining methods use Petri-Net like models. In [7], simulated process logs of hospital-wide workflows, containing events like "blood test" or "surgery" were used to build Petri-Net like models. In [2] a statistical approach, using Hidden Markov Models (HMMs) is taken to model the workflow inside the Operation Room. These probabilistic models offer a greater degree of flexibility and are a better option for healthcare, where traditional process mining techniques do not work well [4].

2.1 Hidden Markov Models

A Hidden Markov Model (HMM) is a statistical technique that can be used to classify and generate time series. A HMM [13] can be described as a quintuplet $I = (A, B, T, N, M)$, where N is the number of hidden states and A defines the probabilities of making a transition from one hidden state to another. M is the number of observation symbols, which in our case are the activities that have been performed to the patients. B defines a probability distribution over all observation symbols for each state. T is the initial state distribution accounting for the probability of being in one state at time $t = 0$. For process discovery purposes, HMMs can be used with one observation symbol per state. Since the same symbol may appear in several states, the Markov model is indeed “hidden”.

We visualize HMMs by using a graph, where nodes represent the hidden states and the edges represent the transition probabilities. The nodes are labelled according to the observation symbol probability.

2.2 HMM-based process discovery

There are multiple advantages of using HMMs for process discovery:

- A lot of (open source) algorithms have been published for analyzing and understanding HMMs (e.g. Expectation Maximization, Viterbi algorithm for most probable path for a given pattern of observations, etc.)
- Micro patterns of actor behavior (e.g. medical acts that belong together) can be easily aggregated into one single state in HMMs. Transitions of 100%. probability can be aggregated into one single state of activity.

- HMMs can be annotated with a variety of attributes, such as (risk and transition) probabilities, time duration, variances, etc.
- HMMs offer better possibilities to match the models obtained from process discovery with the training/learning datasets. In particular, parallel activities are filtered out in HMMs.

In this paper the standard HMM MATLAB toolbox developed by Kevin Murphy was used [9]. The patient data were transformed into sequences, and the Expectation Maximization (EM, also known as Baum-Welch) algorithm was used to produce the results for this paper. This algorithm combines both forward and backward learning techniques for training an HMM as a process model. The input data were organized according to the Event – Object – Actor standard for process mining input. In this case the input data were obtained from standard clinical patient reporting datasets, compatible with the Healthcare Level 7 record standard.

The only large scale commercial toolset for process discovery (including not only the process analytics, but also the automatic non-invasive gathering of input data) is provided by OpenConnect in its Comprehend product family.

3 Data discovery with Formal Concept Analysis

Formal Concept Analysis [5] is a data analysis technique that supports the user in analyzing the data and discovering unknown dependencies between data elements. In particular, the visualization capabilities are of interest to the domain expert who wants to explore the information available, but at the same time has not much experience in mathematics or computer science. The details of FCA theory and how we used it for KDD can be found in [11].

Traditional FCA is mainly using data attributes for concept analysis. In this paper the process activities (events) are used as the attributes, whereas the patients are used as the objects in the cross-table that is used as input for FCA. In analogy with [11] where coherent data attributes were clustered to reduce the computational complexity of FCA, coherent events have been clustered in this study.

4 Dataset

Our dataset consists of 148 breast cancer patients that were hospitalized during the period from January 2008 till June 2008. They all followed the care trajectory determined by the clinical pathway Primary Operable Breast Cancer (POBC), which structures one of the most complex care processes in the hospital. The treatment of breast cancer consists of 4 phases in which 34 doctors, 52 nurses and 14 paramedics are involved. Fig. 1 contains a high-level summary of the breast cancer care process. Before the patient is hospitalized, she ambulatory receives a number of pre-operative investigative tests. During the surgery support phase she is prepared for the surgery she will receive, while being in the hospital. After surgery she remains hospitalized for a couple of days until she can safely go home. The post-operative activities are

also performed in an ambulatory fashion. Every activity or treatment step performed to a patient is logged in a database and in the dataset we included all the activities performed during the surgery support phase to each of these patients.

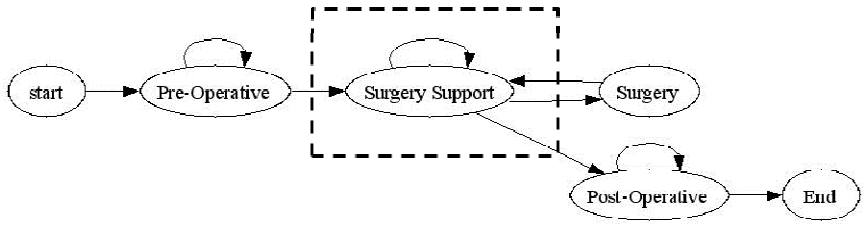

Fig.1. Breast cancer care process

Each activity has a unique identifier and we have 469 identifiers in total for the clinical path POBC. Using the timestamps assigned to the performed activities, we turned the data for each patient into a sequence of events. These sequences of events were used as input for the process discovery methods. We also clustered activities with a similar semantical meaning to reduce the complexity of the lattices and process models.

5 Analysis and results

One of the most important tasks of the care process manager is to gain insight into what's happening on the working floor. The goal was to develop an approach that optimally supports this manager's role. The synergy of process and data discovery techniques in healthcare we propose, has some major advantages over the traditional way of working:

- Significantly reduces the workload for the care process manager who has to monitor over 42 care processes.
- Many unknown data dependencies are revealed that stay hidden for traditional statistical analysis techniques, which typically only look at one or two aspects of the process simultaneously.
- Provides a structured method for finding knowledge gaps, outliers, quality of care issues, process anomalies and inefficiencies.
- Much more information is provided to the process manager, much more quickly. This allows for better analysis and real-time anticipation on potential problems, whereas in the past, this could be done only after a yearly, very time-consuming and labor-intensive retrospective data analysis.
- The method allows the user to zoom in on different aspects of the provided care.

The process models allow for the extraction and visualization of the most frequent standard care pathway. While analyzing these models, we observed many anomalies and process exceptions that were hard to explain. Therefore, we used FCA to zoom in on and analyze these observations in detail.

5.1 Quality of care analysis

Our initial process model was built from the full dataset with 148 patients and 469 activity codes. We observed a relatively linear process for the group of patients with a length of stay in the hospital less than 10 days. However, there were 12 patients for which the process model was very complex. They all had in common that their length of stay in the hospital was longer than 9 days. Fig. 2 contains screenshots from the output produced by the Comprehend toolset. The upper part displays the obtained process map on the set of patients with a length of stay lower than or equal to 10 days in the hospital and the lower part displays the obtained map for the patients with a length of stay lower than 10 days.

Fig.2. Comprehend process map for patients with a length of stay smaller than 10 days (upper part) and process map for patients with a length of stay larger or equal than 10 days (lower part)

We built an FCA lattice to explore their characteristics. This lattice gave us some first interesting insights in the problem. We will try to summarize the most important ones.

- One of our clinical indicators is the pain score which tells us at which days the pain experienced by patients reaches its highest level. We always saw peaks on 1 and day 4 of hospitalization however until now we had no idea why. The lattice gave us an interesting suggestion that this might be due to an overlooked connection between removal of the wound drains and insufficient pain medication. We were able to find that wound drains is probably the most contributing factor to an increased pain score experience

by patients and that pain medication should be administered before removing the drains (= improving quality of care).

- We were able to find a quality problem in the care provided to these 12 patients. For 1 patient the history record (containing amongst others clinical, psychosocial information) was not consulted prior to the start of treatment. This may result in an inappropriate nursing care thereby potentially neglecting physical and psychosocial patient needs.

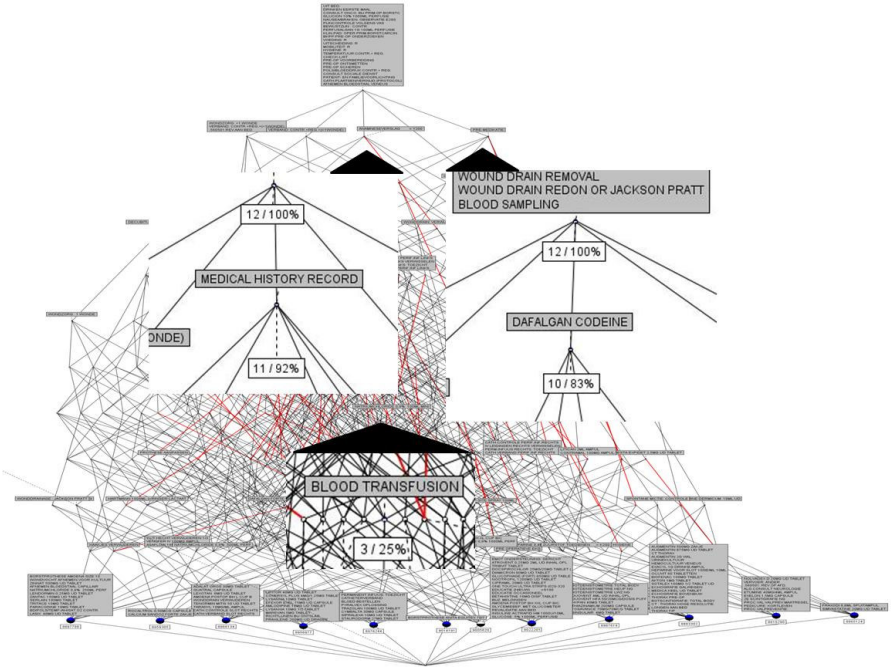

Fig.3. Lattice containing 12 patients with length of stay larger or equal to 10 days

- Probably one of the main reasons of the increased length of stay we found to be the following: neurological/psychiatric problems, wound infection, subsequent bleeding. This makes the care process more complex and result in more investigative tests. Since these additional morbidities are probably one of the root causes for this increased length of stay, there treatment should be anticipated on and optimalized during the preoperative phase.

5.2 Process variations

There are five types of breast cancer surgery: mastectomy, breast conserving surgery, lymph node removal and the combination of either mastectomy or breast conserving surgery with lymph node removal. For each of these surgery types, we extracted the

corresponding patients in the dataset and constructed a process model and an FCA lattice for in-depth analysis of the characteristics of these groups.

Mastectomy surgery consists of completely removing the breast and during breast conserving surgery only the tumor is removed. The process models showed that the complexity of the care process is much larger for the mastectomy patients. Since mastectomy is a more complex surgery type, we expected that the FCA lattices would also be more complex than for breast conserving surgery. Surprisingly we found out that this was not true. The complexity of the lattice was larger for the breast conserving surgery patients and we found that this was due to the less uniform structure of this care process, in which for many patients some essential care interventions were missing. Fig. 4 contains the interventions performed to the 60 patients receiving breast-conserving surgery with lymph node removal. The lattice shows that 3 of these patients did not receive a consultation from the social support service. 15 patients did not have an appointment with a physiotherapist and did not receive revalidation therapy. 1 patient did not receive a pre-operative preparation and 2 patients were missing emotional support before and after surgery.

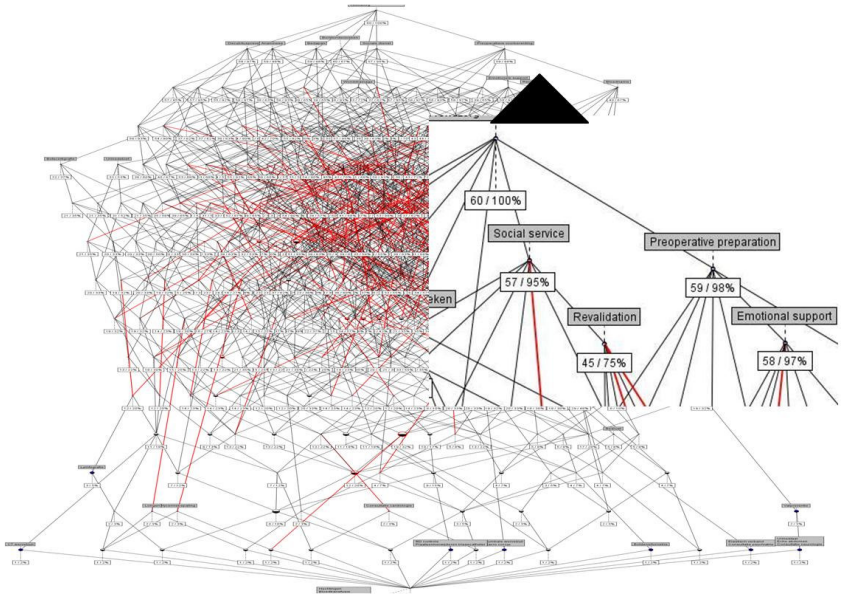

Fig.4. Lattice containing 60 patients receiving breast-conserving surgery with lymph node removal

The originally developed breast-conserving surgery care pathway was written for a certain length of stay for the patients in the hospital. This length of stay was significantly reduced over the past years without modifying the care process model. As a consequence, we found it became impossible to execute the prescribed process model in practice and patients are receiving suboptimal care. The activities performed

to the patients should be reorganized and a new care pathway, taking into account this time restriction, should be developed.

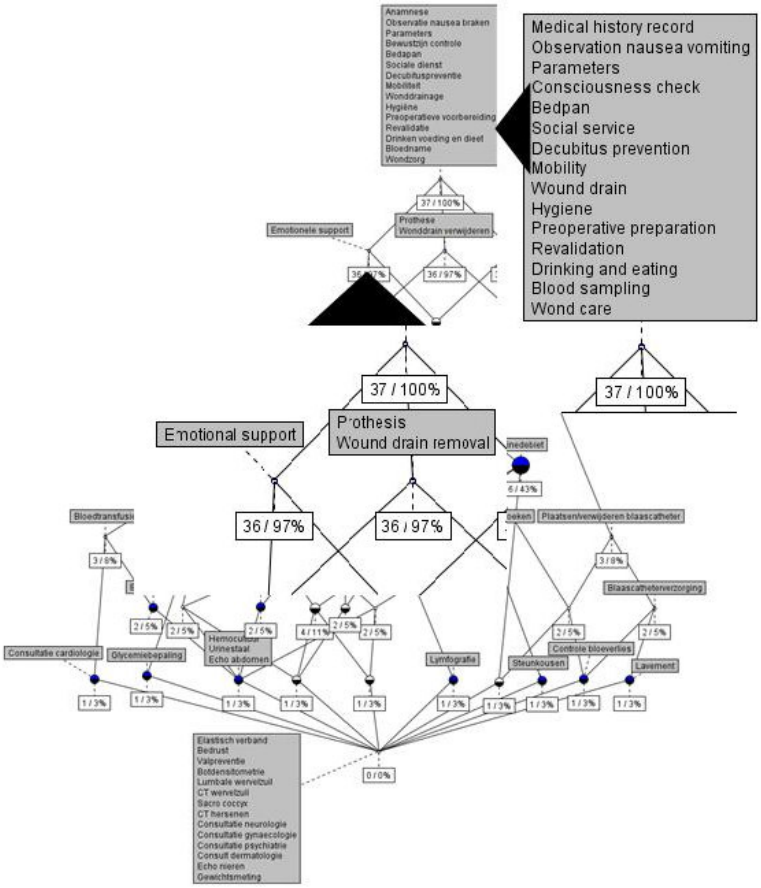

Fig.5. Lattice containing 37 patients receiving mastectomy surgery with lymph node removal

Fig. 5 shows the lattice for the 37 patients receiving mastectomy surgery with lymph node removal, which has a much less complex structure than the lattice for the breast conserving surgery with lymph node removal. For the mastectomy patients, we found that most patients received all key interventions prescribed in the clinical pathway. Only for two patients there was a quality of care issue, namely 1 patient did not receive emotional support and 1 patient did not receive a breast prosthesis. These shortcomings in the provided care however may have serious consequences for her psychological well-being.

5.3 Workforce intelligence

We also made a lattice for each type of surgery in which we used as attributes the names of the surgeons and the length of stay of the patients in the hospital. We calculated the average length of stay of the patients and looked at how many patients stayed longer, equal or shorter than this average time of' stay. Fig. 6 contains the lattice for the 60 patients receiving breast conserving surgery with attributes length of stay and doctor performing the operation.

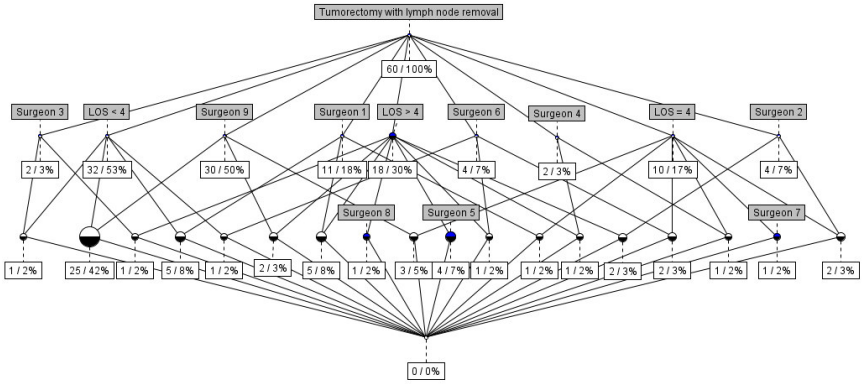

Fig.6. lattice for 60 patients receiving breast conserving surgery

We saw for the breast conserving surgery with lymph node removal that 25 patients with a length of stay smaller than 4 days were treated by "surgeon 9", whereas almost all patients treated by the other doctors had a longer length of stay.

We extracted these subsets of patients and constructed a process model for the groups of patients with a length of stay smaller than 4 days, equal to four days and larger than 4 days. This way, we were able to extract some best practices that could be used to improve the care provided to all patients. Fig. 7 contains the HMM process model extracted from the datasets with the 10 breast-conserving surgery patients with a length of stay in the hospital of 4 days (the average length of stay). This process model was chosen because of its simplicity in comparison with the other models and since it most closely resembles the standard care process as perceived by the domain experts.

Table 1 contains some of the complexity measures for these process variations. For each surgery type and length of stay, the number of patients, the average number of activities and the number of unique activities performed to these patients is given. For visualizing the process maps, we laid a cutoff point at 5%, i.e. all transitions with a lower probability of occurrence were removed from the process representation. The table also contains the number of remaining unique activities and the number of connections after filtering. The structural complexity measure after filtering is the sum of these two measures.

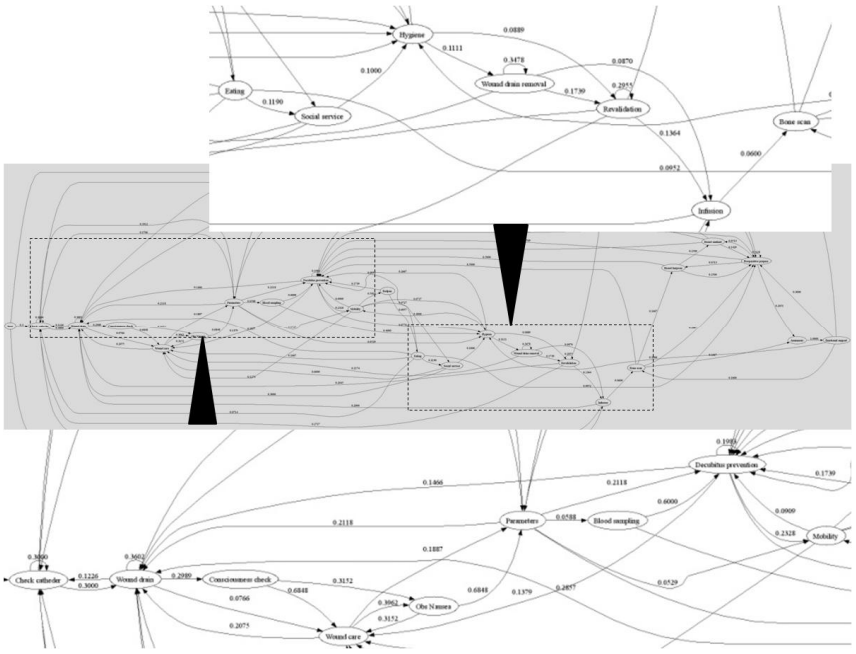

Fig.7. Process model for 10 breast-conserving surgery patients with length of stay of 4 days

Table 1. Complexity measures for the two process variations with the largest number of patients.

SURGERY	\ LOS	LOW	AVG	HIGH
Breast Conserving Therapy with Lymph Node Removal	Length of stay	< 4 days	= 4 days	> 4 days
	# patients	32	10	18
	Avg. # activities	97	146	184
	# unique activities	32	23	35
	# unique act filtered	24	22	22
	# connections filtered	98	80	92
	Struct. Complex. filtered	122	102	114
Mastectomy with Lymph Node Removal	Length of stay	< 7 days	= 7 days	> 7 days
	# patients	17	4	16
	Avg. # activities	187	206	268
	# unique activities	27	21	36
	# unique act filtered	19	20	24
	# connections filtered	83	78	100
	Struct. Complex. filtered	102	98	124

5.4 Data entrance quality issues

Using the process models, we also found some data entrance quality problems. For some patients, activities were registered after the day of discharge. We found that this was due to an error in the computer program combined with sloppy data entry by the nursing staff. We also found many semantically identical activities that had different activity numbers.

When we analyzed the process models, we found that some of the events typically were not ordered in the sequence that they are performed in real life. In other words, the timing of the events as can be found in the data does not always correspond to the timing at the real-life working floor. We found this is due to an error in the computer system which sometimes imposes a certain sequence of events and does not allow for a correct registration of activities.

There is a discrepancy between this built-in top-down developed model and the reality. This discrepancy is probably due to the insufficient insight into the reality of the working floor when the system was developed. The anomalies found during this process mining exercise will be used as input for the development of the new IT systems.

6 Discussion and conclusions

Neither process nor data discovery techniques alone are sufficient for discovering knowledge gaps in particular domains such as healthcare. In this paper, we showed that the combination of both gives significant synergistic results. Whereas FCA does not provide easy to use process representations, it has proven to be very useful for process analysis, i.e. to analyze anomalies and exceptions in detail.

Initially, we thought FCA would only be useful for post-factum analysis of the results obtained through process discovery, but in this case we also found that FCA can play a significant role in the discovery process itself. In particular, concept lattices were used to improve the detection and understanding of outliers in the data. These exceptions are not noise, but are the activities performed to human beings, so every exception counts and must be understood. Concept lattices were also used to reduce the workspace, to cluster related events together in an objective manner.

Using this combination of techniques, we exposed multiple quality of care issues. We gained a better understanding of the process variations and better understood where we should take action to improve our healthcare processes. The impact of co-morbidities of patients on the overall care process was found to be of importance and offers some opportunities for improving quality and efficiency of care. Further, reducing the length of stay of breast-conserving therapy patients was discovered to be the root cause for a suboptimal care, missing some key interventions, provided to patients. Finally, we found the length of stay for patients receiving breast-conserving surgery was significantly different for different surgeons. This situation may be improved by uniformization of discharge criteria.

Avenues for future research include the use of supervised clustering, mainly to obtain normalized process models, in which many-to-many transitions are eliminated

(as argued in [10]). The normalized clusters will give the best views on process variations. Again, a posterior data discovery (FCA) can be used to understand the meaning of the different clusters.

Acknowledgements

We would like to express our special thanks to Chris Houck (OpenConnect) for his help in the construction of the process models. Jonas Poelmans is aspirant of the ”Research Foundation - Flanders" or "Fonds voor Wetenschappelijk Onderzoek - Vlaanderen”. Ed Peters is special guest professor for the K.U.Leuven OPENCONNECT Research Chair on Process Discovery.

References

1. Anyanwu, K., Sheth, A., Cardoso, J., Miller, J., Kochut, K. (2003). Healthcare enterprise process development and integration. Journal of research and practice in information technology, 35 (2): 83-98.
2. Blum, T., Padoy, N., Feussner, H., Navab, N. (2008) Workflow mining for visualization and analysis of surgery. International journal of computer assisted radiology and surgery, 3. suppl. 1, June 2008.
3. Campbell, H., Hotchkiss, R., Bradshaw, N., Porteous, M. (1998) Integrated care pathways, British medical journal 316, 133-137.
4. Ferreira, D., Zacarias, M., Malheiros, M., Ferreira, P. (2007) Approaching process mining with sequence clustering: experiments and findings. BPM 2007, Lecture Notes in Computer Science, Vol. 4714, 360-374. Springer, Heidelberg.
5. Ganter, B., Wille, R. (1999), Formal Concept Analysis: Mathematical Foundations. Springer, Heidelberg.
6. Mans, R.S, Schonenberg, M. H., Song, M., Aalst, W.M.P. , Bakker, P.J.M. (2009) Application of process Mining in health care - a case study in a Dutch hospital. Biomedical engineering systems and technologies, International Joint conference, BIOSTEC 2008 Funchal, Madeira, Portugal, January 28-31, 2008. Springer Heidelberg.
7. Maruster, L., Van der Aalst, W.M.P., Weijters, A.J.M.M., Van der Bosch, A. Daelemans, W. (2002) Automated discovery of workflow models from hospital data. Proceedings of the ECAI workshop on knowledge discovery and spatial data, 183-190.
8. Maruster, L., Weijters, A.J.M.M., Van der Aalst, W.M.P., Van den Bosch, A. (2006). A rule-based approach for process discovery dealing with noise and imbalance in process logs. Data mining and knowledge discovery, 13.
9. Murphy, K. (1998) Hidden Markov Model (HMM) MATLAB toolbox http://people.cs.ubc.ca/~murphyk/Software/HMM/hmm.html
10. Peters, E. M., Dedene, G., Houck, C. (2009). Business Process Discovery and Workflow Intelligence Techniques with healthcare applications, Proceedings of the INFORMS ORAHS 2009 Conference, Leuven 2009 http://www.econ.kuleuven.be/eng/tew/academic/prodbel/ORAHS2009/page5.htm
11. Poelmans, J., Elzinga, P., Viaene, S., Dedene, G. (2009). A case of using formal concept analysis in combination with emergent self organizing maps for detecting

domestic violence. In : Lecture Notes in Artificial Intelligence, Vol. 5633(XI), (Perner, P. (Eds.)). Industrial conference on data mining ICDM 2009. Leipzig (Germany), 20-22 July 2009 (pp. 402).
12. Quaglini, S. (2009) Process Mining in Healthcare: A Contribution to Change the Culture of Blame. Lecture Notes in Business Information Processing. Vol. 17, 308-311. Book Business Process Management Workshops. Springer Berlin Heidelberg
13. Rabiner, L.R. (1989) A tutorial on Hidden Markov Models and selected applications in speech recognition. Proceedings IEEE 77 (2): 257-286.
14. Skinner, J. Chandra, A., Goodman, D., Elliot, S. F. (2008) The elusive connection between health care spending and quality. Health affairs - web exclusive w119-w123.