Getting a Grasp on Clinical Pathway Data: An Approach Based on Process Mining

Jochen De Weerdt[1], Filip Caron[1], Jan Vanthienen[1], and Bart Baesens[1,2]

[1] Department of Decision Sciences and Information Management, KU Leuven
Naamsestraat 69, B-3000 Leuven, Belgium
Jochen.DeWeerdt@kuleuven.be
[2] School of Management, University of Southampton
Highfield Southampton, SO17 1BJ, United Kingdom

Abstract. Since healthcare processes are pre-eminently heterogeneous and multi-disciplinary, information systems supporting these processes face important challenges in terms of design, implementation and diagnosis. Nonetheless, streamlining clinical pathways with the purpose of delivering high quality care while at the same time reducing costs is a promising goal. In this paper, we propose a methodology founded on process mining for intelligent analysis of clinical pathway data. Process mining can be considered a valuable approach to obtain a better understanding about the actual way of working in human-centric processes such as clinical pathways by investigating the event data as recorded in healthcare information systems. However, capturing tangible knowledge from clinical processes with their ad hoc and complex nature proves difficult. Accordingly, this paper proposes a data analysis methodology focussing on the extraction of tangible insights from clinical pathway data by adopting both a drill up and a drill down perspective.

Keywords: process mining, clinical pathways, healthcare information systems, event logs, fuzzy miner.

1 Introduction

Worldwide, the healthcare sector goes through a major reform. This has many reasons. First, the costs of healthcare are rising up to 15% in the United States (US) and close to 10% in Europe [1]. This is due to the increasing needs of a greying population, but also due to technological and pharmacological innovations that are really widening the possibilities for diagnosis and treatment. Second, there is a shift in the role of patients, going from a more passive role into a role of active consumers of care. Patients want to be informed and involved. Third, there is growing attention to quality and safety. The main drive comes from the *to err is human* report from the Institute of Medicine[1] [2]. This report indicated that as many as 44.000 to 98.000 US citizens die in hospitals each year as the result of medical errors. Even using the lower estimate, this would make medical

[1] http://www.iom.edu

T. Washio and J. Luo (Eds.): PAKDD 2012 Workshops, LNAI 7769, pp. 22–35, 2013.

errors the eighth leading cause of death in the US - higher than motor vehicle accidents (43.458), breast cancer (42.297) or AIDS (16.516). The report was publicly discussed in the Senate and was the start of an overall hospital reform. Most important in the discussion is that people are not blamed - *to err is human* indeed - but that the focus should be on improving the system.

An excellent way to do this is learning from past experience. Currently, it can be observed that with the growing implementation of integrated healthcare information systems, vast amounts of data are becoming available about the actual way of working in clinical pathways. These data form the cornerstone of this study. Accordingly, the notion of a clinical pathway is a crucial element. The terminology has its origins in methodologies such as PERT (Project Evaluation and Review Technique) and CPM (Critical Path Method), but transformed into "clinical" instead of "critical" pathways because of the very specific nature of healthcare. Clinical pathways are formally defined as *a complex intervention for the mutual decision making and organization of care for a well-defined group of patients during a well-defined period* by the European Pathway Association[2].

In this study, we propose an approach for deriving useful insights from clinical pathway data by making use of process mining techniques. Process mining is a relatively young research area [3], which lies at the intersection of data mining and Business Process Management (BPM) [4]. It consists of a family of analysis techniques for analyzing event logs as recorded by the logging infrastructures of information systems. Since these techniques rely on real data, they are capable of providing insight into the actual way of working in the context of a certain business process. In this paper, we will describe a methodology based on state-of-the-art process mining techniques for the analysis of clinical pathway data. The main contribution of this study is the development of solution strategies for dealing with the extremely unstructured nature of clinical pathway data.

2 Related Work

Process Aware Healthcare Information Systems. A fair share of information systems implemented in healthcare organizations can be described as process aware. This is because process aware information systems not only encompass traditional workflow management systems, but also include systems that provide much more flexibility. Accordingly, once an information system can be described as having an explicit notion of the process it supports, it can be described as process aware [5]. A such, many healthcare information systems fit within this definition. Since clinical pathways are inherently heterogeneous and multi-disciplinary in nature, the goal of IT support for healthcare processes is not to control the course of the process entirely, but to assist healthcare professionals by reducing cognitive overload and improving the basis for their decisions [6]. In this way, process orientation can be considered as a beneficial approach towards streamlining clinical pathways with the purpose of delivering high quality care while at the same time reducing costs [7]. While business process support

[2] `http://www.e-p-a.org`

for structured processes (e.g. manufacturing, logistics) has always been an important research topic, the growing importance of service organizations such as healthcare has triggered the need for different approaches towards business process support [8,9]. Because of the human-centric nature of such processes, they contain much more flexibility, alternative routings, loops, human judgement and variability than traditional business processes. Accordingly, the analysis of the actual way of working by making use of the data captured by healthcare information systems is promising. However, traditional business process analysis techniques come short in realizing this goal and therefore this paper proposes a novel analysis methodology based on process mining.

Process Mining. The field of process mining can be best defined as a broad family of techniques for the analysis of event logs as recorded by the logging infrastructures of information systems. These techniques can be broadly categorized into three groups according to three commonly distinguished process mining tasks: discovery, conformance and enhancement [3]. The most important learning task is called process discovery [10] which entails the extraction of control-flow models from such event logs. In the process mining literature, a lot of attention has been paid to the development of process discovery techniques [11,12,13]. However, discovery tasks can also focus on other aspects of an event log, for instance on organizational information [14]. Conformance [15] is a second important process mining task. Hereto, a process model is compared with the data in the event log with the purpose to verify whether reality conforms to the model and vice versa. Finally, the idea of enhancement tasks is to extend or improve process models with other information about the actual process as recorded in the event log. For instance, the addition of a performance perspective enhances a process model and provides the analyst with different insights.

Process Mining in Healthcare. Process mining techniques have been applied in a healthcare context. A first study by Mans et al. [16] shows how different process mining techniques such as HeuristicsMiner, social network analysis and dotted chart analysis allow for obtaining insights into care flow data. Another study by Rebuge and Ferreira [17] also describes a methodology for the analysis of business processes in a healthcare environment. The methodology consists of seven phases with its main asset being the application of sequence clustering techniques. Further, Bose and van der Aalst [18] propose the use of fuzzy mining and trace alignment for investigating clinical pathway data. Finally, Caron et al. [19] demonstrate the applicability of various process mining techniques to healthcare data by adopting both a department and a treatment based focus. This study differs from previous studies because it shows the benefits of both a drill up and a drill down perspective on the data relying on control-flow discovery with the Fuzzy Miner and networked graph visualizations.

3 Description of the Clinical Pathway Data

The data set concerns real data of a gynecological oncology process at the AMC hospital in Amsterdam, The Netherlands. It was first used in [16], but recently the data set was made publicly available (doi:10.4121/uuid:d9769f3d-0ab0-4fb8-803b-0d1120ffcf54) for the first Business Process Intelligence Challenge (BPIC'11). The data contains 150.291 events of 1.143 patient treatment processes related to individuals diagnosed with cancer pertaining to the cervix, vulva, uterus and/or ovary. Each case in the event log corresponds to a single patient and as such, the data presents a wide variety of care activity sequences. In the remainder of this section, the three most important dimensions of the data are outlined.

Diagnosis. Each case in the data set contains information on the type of disease the patient is diagnosed with. The related attributes are denoted *Diagnosis code* and *Diagnosis*. The data presents a total of eleven different diagnosis codes (e.g. M11, M12, 823, etc.). *Diagnosis* is a textual description which specifies the diagnosis code, taking values such as "adenocarcinoma stage Ia" or "clear cell carcinoma". It should be noted that the data contains up to 16 diagnosis code - diagnosis combinations for a single patient (denoted as *diagnosis code:1 to 16*). Accordingly, a single case might contain different codes. We observe 38 distinct diagnosis code combinations, for instance {M16, 821}. Table 2 presents an overview of the diagnosis codes detailing the region, example diagnoses and the number of cases showing this diagnosis code.

Treatment. Next to diagnosis information, one can also find details concerning the treatments of each of the patients. However, in contrast to diagnosis, the data only provides a treatment code and no further information. As such, the treatment perspective is more difficult to analyze. On top of this, there exist 46 distinct treatment codes which form 236 distinct treatment code combinations in similar fashion as the diagnosis code combinations.

Departments. The final important data dimension is organizational in nature, i.e. the departments that are involved in the clinical pathways. Each event in the log contains an attribute "org:group" which denotes the department where the corresponding activity was performed. In the data, one can find 43 distinct organizational units. The most frequently observed departments are depicted in Table 3.

A very specific feature of the data is that events pertaining to certain departments occur in bursts. For instance, sets of blood diagnosis tests performed by the *General Lab Clinical Chemistry* department are often found. Similarly, bursts of radiotherapy-, nursing-, operating room-related and many other types of events can be observed. This data characteristic will be further employed in the next section.

Table 1. Excerpt of the event log

Case ID	Event name	Dept.	Timestamp	Diagn. code	Diagnosis	Treatm. code	Age	...
0	1st polyclinic consult	Radiotherapy	03/01/2005	M13	cervical malign.	23	33	...
0	administrative reg.	Radiotherapy	03/01/2005	M13	cervical malign.	23	33	...
0	gynec. cost assign.	Nursing ward	03/01/2005	M13	cervical malign.	23	33	...
0	ultrasonography	Obstr.&gyn. clinic	03/01/2005	M13	cervical malign.	23	33	...
0	1st consult	Nursing ward	03/01/2005	M13	cervical malign.	23	33	...
...	...	...	...	...	...	...	...	...

Table 2. Diagnosis codes in the clinical pathway data

Code	Region	Example diagnoses	# cases
M11	vulva	squamous cell carcinoma, borderline malignancy	176
M12	vagina	squamous cell carcinoma, adenocarcinoma	22
M13	cervix	squamous cell carcinoma, malignant neoplasms	368
M14	corpus uteri	adenocarcinoma, clear cell carcinoma	145
M15	corpus uteri, myometrium	sarcoma	17
M16	ovary	squamous cell carcinoma, non-epithelial malignancy	235
106	cervix, vulva, corpus uteri, vagina	squamous cell carcinoma, borderline malignancy	298
821	ovary	serous and mucinous squamous cell carcinoma, neoplasms	48
822	cervix	squamous cell carcinoma, adenocarcinoma	131
823	corpus uteri, ovary, endometrium	(serous) adenocarcinoma	16
839	ovary, vulva	serous adenocarcinoma, malignant neoplasms	21

Table 3. Number of events pertaining to organizational units by frequency

Organizational unit	# events
General Lab Clinical Chemistry	94917
Nursing ward	31066
Obstetrics & Gynaecology clinic	7065
Medical Microbiology	4170
Radiology	3171
Radiotherapy	2233
Internal Specialisms clinic	2146
Pathology	1975
Operating rooms	942
Pharmacy Laboratory	498
Recovery room / high care	495
Nuclear Medicine	281
Special lab radiology	279
...	...

4 Analysis Methodology and Results

Our data analysis approach combines two important strategies for extracting tangible knowledge from clinical pathway data. First, drill up is applied in order to get insight into the general behavior of the healthcare process. In a second phase, a drill down approach is described that centers on a certain part of the data.

4.1 Complexity of Clinical Pathway Data

The crucial challenge for data analysis in the context of clinical pathways is the complexity of the data. This is because clinical pathways are inherently ad hoc, multi-disciplinary and strongly human-centric. Because of these characteristics, almost every observed clinical pathway is unique, which is also the case for the the data set employed in this study. On top of that, the original data set contains 624 different activity types. Further, these activity types as registered in the different departments are not always of the same granularity. A final element that complicates the data analysis significantly is the fact that we can only observe care activities executed within the AMC hospital. However, it is undoubtedly reasonable to assume that other care activities in peripheral hospitals, by GP's, etc. are being executed but not registered in the data. Because of these data complexities, there is a need for versatile data analysis methods. In this study, it is shown how process mining techniques can be used, both in a drill up as well as in a drill down mode. Furthermore, we demonstrate the use of networked graphs for visualizing sets of cases and their respective characteristics.

4.2 Drill Up Analysis

Both in [16] and [19], it is shown that the straightforward application of existing process discovery techniques is infeasible for the data at hand. Even the stronger generalization capabilities of Fuzzy Mining [20] prove not very helpful. Therefore, we show how abstraction applied in a data preprocessing step can be beneficial in order to obtain general, but useful insights based on the entire data set.

Data Preprocessing. Realizing abstraction in a data preprocessing phase consists of replacing the bursts of events belonging to the same organizational unit by the name of the organizational unit itself. In this way, a clinical pathway in terms of the unique activities performed by different organizational units is transformed into sequences of departments.

A General View Using Process Discovery. As stated earlier, process discovery is the most important asset of the process mining domain. Process discovery is defined as the extraction of control-flow models from event logs. Note that these techniques make use of different process modeling notations (e.g. Petri nets, heuristic nets, fuzzy nets, etc.) in order to represent the discovered model. In this case, we applied the Fuzzy Miner to the transformed clinical pathways. The resulting fuzzy net is depicted in Figure 1. Note that the nodes contain a significance value between 0 and 1. Further, the figures on the edges indicate the edge significance and correlation, also ranging between 0 and 1.

With the purpose to increase the comprehensibility, we restricted the visualization to the eleven most frequent departments. Together with the abstraction power of Fuzzy Miner, the discovered graph provides some interesting insights with respect to the gynecological oncology process under investigation:

- A majority of the patients first visit the obstetrics and gynecology department.
- From top to bottom, we can clearly observe the diagnostic-therapeutic cycle characteristic to the majority of care processes. The nursing wards have a pivotal role in between diagnostics and therapeutics.
- From a diagnostic perspective, lab analyses (majorally blood sample tests) and to a lesser extent radiology are essential elements of disease typification in the context of gynecological oncology.
- Despite the fact that the data cover patients with very comparable diagnoses (i.e. gynecological cancers), streamlined clinical pathways cannot be observed, even in terms of involved departments.

4.3 Drill Down Analysis

As described in the previous section, drill up is a valuable approach for extracting general knowledge from care process data. Nevertheless, due to the characteristics of the data, intelligent drill down into specific parts of the data is bound

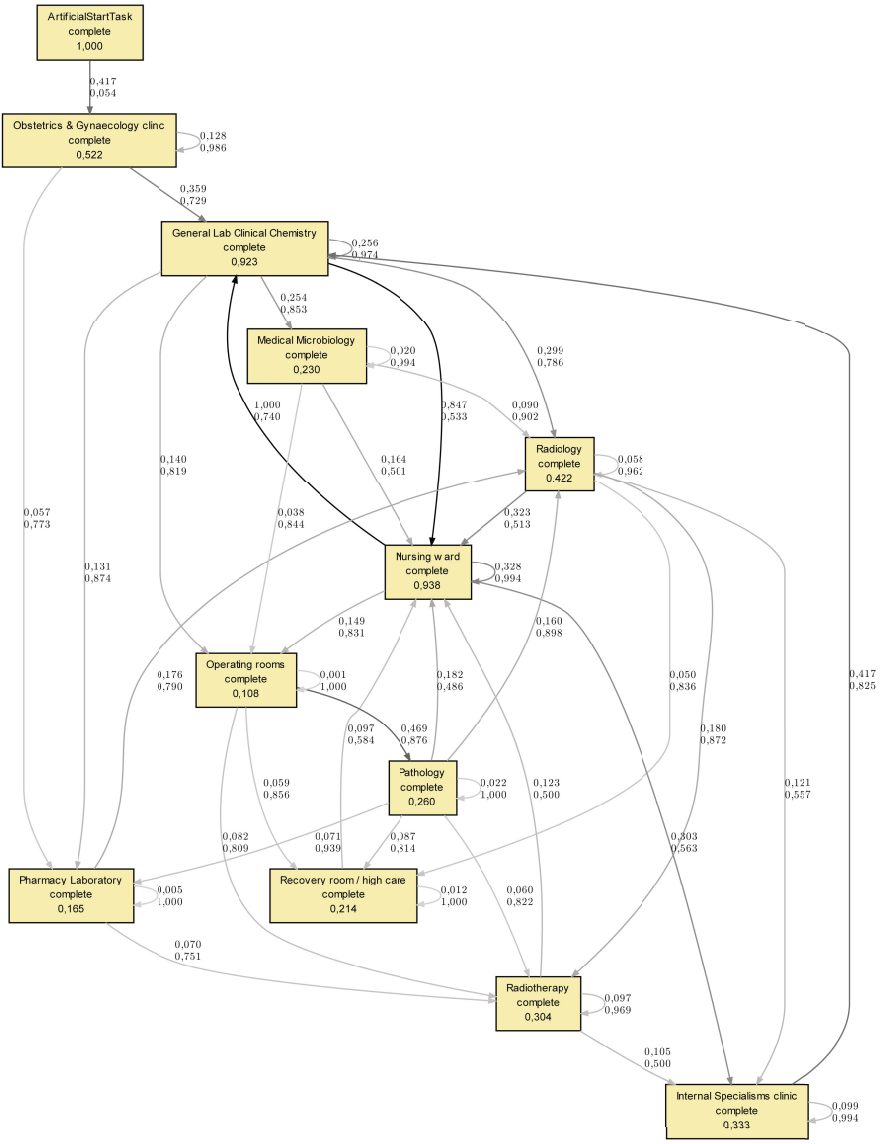

Fig. 1. Fuzzy process model showing the relations between frequently occurring organizational units

to provide even more interesting insights. Hereto, we demonstrate a particular focus on the therapeutic side of the clinical pathways. Since the data provides only limited information on specific treatments (only treatment codes, without any explanation), it is opted to investigate this perspective in more detail.

Focalizing on Therapeutic Activities. To adopt the focus on treatment, therapeutic activities need to be singled out. By inspection of the cases, it can be noticed that three different types of treatments can be identified: radiotherapy, chemotherapy and surgery. For radiotherapy, the selection of activities was rather straightforward since the granularity of the events is relatively coarse-grained. The data contains events such as *teletherapie - megavolt fotonen bestrali* and *brachytherapie - interstitieel - intensi*. Further, the radiotherapy department also carries out hyperthermia treatments, which are strictly speaking not radiotherapeutic, but often used in combination with radiation therapy. For chemotherapy, the selection of appropriate events was slightly more difficult due to the fact that this type of therapy is scattered between different organizational units. Nonetheless, we identified two important chemotherapeutic activities, viz. *paclitaxel* and *doxorubicine*. Finally, also surgical treatments should be taken into account. However, looking at the events pertaining to the Operating rooms department, there clearly exist two different types of procedures. On the one hand, the data set shows a multitude of diagnostic surgical procedures, whereas on the other hand only therapeutic operations are of interest given our current focus. Nonetheless the thin line between both, we were able to distinguish between the diagnostic or therapeutic nature of procedures by investigating the names of the events. For example, hysterectomies and vulvectomies were considered as therapeutic surgical activities, while hysteroscopies and urethrocystoscopies were not.

After singling out these therapeutic activities, 477 cases could be observed for which at least one therapeutic activity has taken place. The event log consisting of all these events was used to visualize the therapeutic activities by means of a process model. However, due to the large number of different surgical procedures, we renamed all these events to *surgery*, an abstraction which allows for more useful visualizations. The resulting process model as obtained with fuzzy mining is depicted in Figure 2. In contrast to Figure 1, the fuzzy net is slightly adapted by replacing the original figures in the nodes and on the arcs by more insightful statistics. As such, the nodes contain their frequency of occurrence, while the figure on each of the edges of the graph shows the number of times a case followed the transition from the source node to the target node.

The analysis allows for the formulation of the following findings:

- The fuzzy net shows a number of possible therapeutic choices. However, because the process model does not contain any such paths, it can be concluded that the combination of surgery or chemotherapy with radiotherapy occurs highly infrequent.
- The use of chemotherapy is rather limited despite the fact that regimens, i.e. combinations of different chemotherapy drugs, are often recommendable.

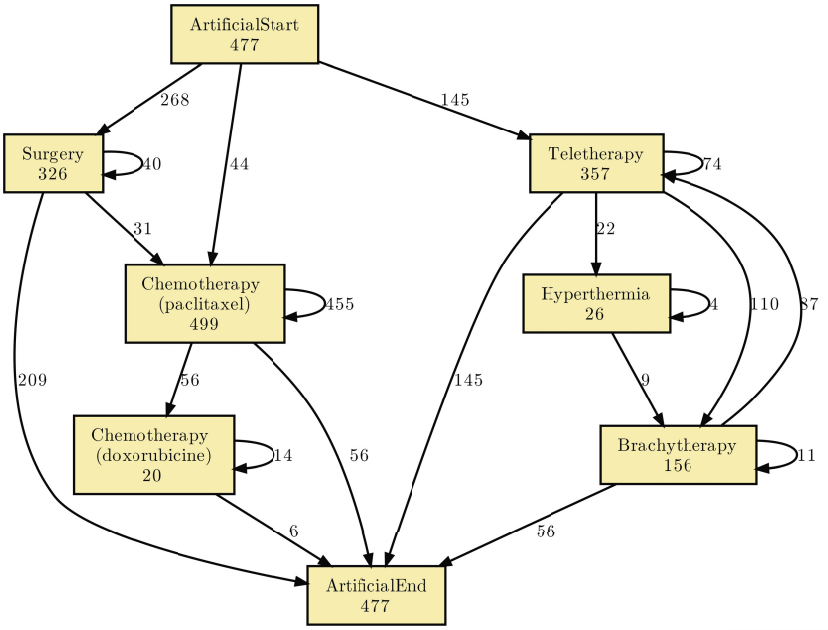

Fig. 2. Adapted fuzzy process model showing the relations between different therapeutic activities

In this way, we could only observe very few chemotherapeutic combinations of Paclitaxel with Doxorubicin. It should be further investigated why chemotherapy is underrepresented. One possible explanation might be that chemotherapeutic procedures might be carried out in peripheral hospitals, thus not captured in the data.

Visualizations Using Networked Graphs. A second drill down approach consists of visualizing a subset of cases by means of a networked graph. This methodology is useful because it allows to visualize cases from different angles by supplementing control-flow information with other perspectives. The construction of a networked graph consists of three steps. First cases are selected based on some criterion. In this case, we considered all cervical cancer cases. Secondly, a Euclidean distance matrix is constructed denoting the distance between each pair of cases. This matrix is built by making use of the MRA (Maximum Repeat Alphabet) technique as proposed in [21]. The MRA technique relies on the identification of specific patterns which characterize the traces. Notwithstanding the fact that the authors employ the method for clustering log traces, we use the underlying distance matrix to construct a networked graph. Such a network graph connects nodes which represent a case in the data. For comprehensibility reasons, sparsification is applied in order to reduce the number of connections between the nodes because otherwise a fully connected graph is obtained. In this case, we applied K-nearest-neighbors with $K = 2$.

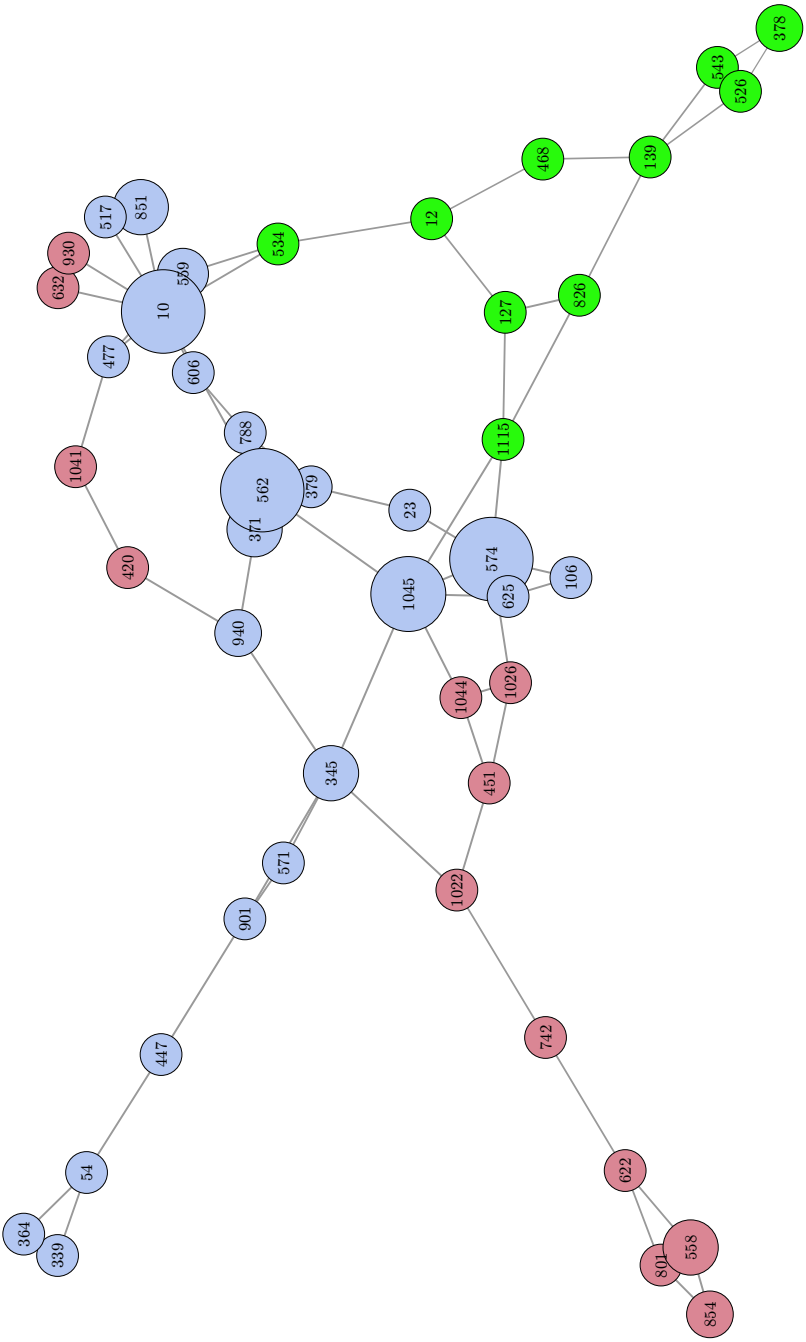

Fig. 3. Networked graph visualizing cervical cancer cases

Figure 3 shows one visualization created with this methodology. Note that for visualization of the graph, we employed the Yifan Hu algorithm [22] as implemented in Gephi[3]. The graph shows all patients diagnosed with some type of cervical cancer. The distance between the nodes is determined by the MRA distance, which entails that nodes which are closer together present similar execution paths in terms of the therapeutic events they contain. The figure in the nodes denotes one representative case ID, with the size of the nodes representing the frequency of a certain sequence of therapeutic activities. Note that it was infeasible to represent all ID's in each of the nodes of the graph. Furthermore, the node colors indicate the application of some specific therapeutic procedure. As such, the green nodes denote the occurrence of chemotherapy. In contrast, the red nodes are cases where hyperthermia treatment is applied. From this visualization, it can be seen that a vast majority of cases do not rely on hyperthermia or chemotherapy. Cervical cancer is majorally treated by either surgery or radiotherapy (teletherapy/brachytherapy).

5 Conclusion

In this study, it was shown how intelligent analysis of clinical pathway data based on process mining techniques can deliver valuable insights into the actual carrying out of a care process. In a practical case consisting of data on the clinical pathways of 1.147 gynecological oncology patients at the AMC hospital, it was demonstrated how both drill up as well as drill down approaches are useful for care flow knowledge discovery. We are convinced that data analysis based on the innovative techniques in the process mining domain is an ideal means for better streamlining and overall improvement of clinical care processes. In the future, we will focus on the development of novel methodologies for analyzing the complex data that is typically found in the logging infrastructures of healthcare information systems. As such, we will further elaborate the idea of networked graph visualizations and improve its integration with existing process discovery techniques. The major benefit of the technique is the enhancement of pure control-flow patterns with other data dimensions. Therefore, additional information, for instance on whether patients were cured or not, would instigate a wide variety of analysis possibilities.

Acknowledgments. The authors would like to thank the Flemish Research Council for financial support under Odysseus grant B.0915.09 and KU Leuven for grant OT/10/010.

References

1. OECD: OECD health data 2010: Statistics and indicators (2010), http://www.oecd.org/health/healthdata
2. Kohn, L.T., Corrigan, J.M., Donaldson, M.S.: To Err Is Human: Building a Safer Health System. The National Academies Press, Washington DC (2000); Committee on Quality of Health Care in America, Institute of Medicine

[3] www.gephi.org

3. van der Aalst, W.M.P.: Process Mining - Discovery, Conformance and Enhancement of Business Processes. Springer (2011)
4. Weske, M.: Business Process Management: Concepts, Languages, Architectures. Springer (2007)
5. Dumas, M., van der Aalst, W.M.P., ter Hofstede, A.H.M.: Process-Aware Information Systems: Bridging People and Software through Process Technology. John Wiley & Sons, Inc. (2005)
6. Lenz, R., Reichert, M.: It support for healthcare processes - premises, challenges, perspectives. Data Knowl. Eng. 61(1), 39–58 (2007)
7. Anyanwu, K., Sheth, A.P., Cardoso, J., Miller, J.A., Kochut, K.: Healthcare enterprise process development and integration. Journal of Research and Practice in Information Technology 35(2), 83–98 (2003)
8. Lenz, R., Elstner, T., Siegele, H., Kuhn, K.A.: A practical approach to process support in health information systems. Journal of the American Medical Informatics Association 9(6), 571–585 (2002)
9. Reijers, H.A., Russell, N., van der Geer, S., Krekels, G.A.M.: Workflow for healthcare: A methodology for realizing flexible medical treatment processes. In: [24], pp. 593–604
10. van der Aalst, W.M.P., Weijters, A.J.M.M., Maruster, L.: Workflow mining: Discovering process models from event logs. IEEE Trans. Knowl. Data Eng. 16(9), 1128–1142 (2004)
11. Alves de Medeiros, A.K., Weijters, A.J.M.M., van der Aalst, W.M.P.: Genetic process mining: an experimental evaluation. Data Mining and Knowledge Discovery 14(2), 245–304 (2007)
12. Weijters, A.J.M.M., van der Aalst, W.M.P., Alves de Medeiros, A.K.: Process mining with the heuristicsminer algorithm. BETA Working Paper Series 166, TU Eindhoven (2006)
13. Goedertier, S., Martens, D., Vanthienen, J., Baesens, B.: Robust process discovery with artificial negative events. Journal of Machine Learning Research 10, 1305–1340 (2009)
14. Song, M., van der Aalst, W.M.P.: Towards comprehensive support for organizational mining. Decision Support Systems 46(1), 300–317 (2008)
15. Rozinat, A., van der Aalst, W.M.P.: Conformance checking of processes based on monitoring real behavior. Information Systems 33(1), 64–95 (2008)
16. Mans, R.S., Schonenberg, H., Song, M., van der Aalst, W.M.P., Bakker, P.J.M.: Application of Process Mining in Healthcare - A Case Study in a Dutch Hospital. In: Fred, A.L.N., Filipe, J., Gamboa, H. (eds.) BIOSTEC 2008. CCIS, vol. 25, pp. 425–438. Springer, Heidelberg (2008)
17. Rebuge, Á., Ferreira, D.R.: Business process analysis in healthcare environments: A methodology based on process mining. Information Systems 37(2), 99–116 (2012)
18. Bose, R.P.J.C., van der Aalst, W.M.P.: Analysis of patient treatment procedures. In: [23], pp. 165–166
19. Caron, F., Vanthienen, J., De Weerdt, J., Baesens, B.: Advanced care-flow mining and analysis. In: [23], pp. 167–168
20. Günther, C.W.: Process Mining in Flexible Environments. PhD thesis, TU Eindhoven (2009)
21. Bose, R.P.J.C., van der Aalst, W.M.P.: Trace clustering based on conserved patterns: Towards achieving better process models. In: [24], pp. 170–181

22. Hu, Y.: Algorithms for Visualizing Large Networks. In: Naumann, U., Schenk, O. (eds.) Combinatorial Scientific Computing (to appear)
23. Daniel, F., Barkaoui, K., Dustdar, S. (eds.): BPM Workshops 2011, Part I. LNBIP, vol. 99. Springer, Heidelberg (2012)
24. Rinderle-Ma, S., Sadiq, S.W., Leymann, F. (eds.): BPM 2009. LNBIP, vol. 43. Springer, Heidelberg (2010)