PROCESS MINING IN HEALTHCARE
A Case Study

R.S. Mans, M.H. Schonenberg, M. Song, W.M.P. van der Aalst
Eindhoven University of Technology, P.O. Box 513, NL-5600 MB, Eindhoven, The Netherlands
{*r.s.mans, m.h.schonenberg, m.s.song, w.m.p.v.d.aalst*}*@tue.nl*

P.J.M. Bakker
Academic Medical Center, University of Amsterdam, Department of Innovation and Process Management, Amsterdam, The Netherlands
p.j.bakker@amc.uva.nl

Keywords: process mining, healthcare processes

Abstract: To gain competitive advantage, hospitals try to streamline their processes. In order to do so, it is essential to have an accurate view of the "careflows" under consideration. In this paper, we apply process mining techniques to obtain meaningful knowledge about these flows, e.g., to discover typical paths followed by particular groups of patients. This is a non-trivial task given the dynamic nature of healthcare processes. The paper demonstrates the applicability of process mining using a real case of a gynecological oncology process in a Dutch hospital. Using a variety of process mining techniques, we analyzed the healthcare process from three different perspectives: (1) the control flow perspective, (2) the organizational perspective and (3) the performance perspective. In order to do so we extracted relevant event logs from the hospitals information system and analyzed these logs using the ProM framework. The results show that process mining can be used to provide new insights that facilitate the improvement of existing careflows.

1 INTRODUCTION

In a competitive health-care market, hospitals have to focus on ways to streamline their processes in order to deliver high quality care while at the same time reducing costs (Anyanwu et al., 2003). Furthermore, also on the governmental side and on the side of the health insurance companies, more and more pressure is put on hospitals to work in the most efficient way as possible, whereas in the future, an increase in the demand for care is expected.

A complicating factor is that healthcare is characterized by highly *complex* and extremely *flexible* patient care processes, also referred to as "careflows". Moreover, many disciplines are involved for which it is found that they are working in isolation and hardly have any idea about what happens within other disciplines. Another issue is that within healthcare many autonomous, independently developed applications are found (Lenz et al., 2002). A consequence of this all is that *it is not known what happens in a healthcare process for a group of patients with the same diagnosis*.

The concept of process mining provides an interesting opportunity for providing a solution to this problem. Process mining (van der Aalst et al., 2003) aims at extracting process knowledge from so-called "event logs" which may originate from all kinds of systems, like enterprise information systems or hospital information systems. Typically, these event logs contain information about the start/completion of process steps together with related context data (e.g. actors and resources). Furthermore, process mining is a very broad area both in terms of (1) applications (from banks to embedded systems) and (2) techniques.

This paper focusses on the *applicability* of process mining in the healthcare domain. Process mining has already been successfully applied in the service industry (van der Aalst et al., 2007a). In this paper, we demonstrate the applicability of process mining to the healthcare domain. We will show how process mining can be used for obtaining insights related to careflows, e.g., the identification of care paths and (strong) collaboration between departments. To this end, in Section 3, we will use several mining techniques which will also show the diversity of process mining techniques available, like control flow discovery but also

the discovery of organizational aspects.

In this paper, we present a case study where we use raw data of the AMC hospital in Amsterdam, a large academic hospital in the Netherlands. This raw data contains data about a group of 627 gynecological oncology patients treated in 2005 and 2006 and for which all diagnostic and treatment activities have been recorded for financial purposes. Note that we did not use any a-priori knowledge about the care process of this group of patients and that we also did not have any process model at hand.

Today's Business Intelligence (BI) tools used in the healthcare domain, like Cognos, Business Objects, or SAP BI, typically look at aggregate data seen from an external perspective (frequencies, averages, utilization, service levels, etc.). These BI tools focus on performance indicators such as the number of knee operations, the length of waiting lists, and the success rate of surgery. Process mining looks "inside the process" at different abstraction levels. So, in the context of a hospital, unlike BI tools, we are more concerned with the care paths followed by individual patients and whether certain procedures are followed or not.

This paper is structured as follows: Section 2 provides an overview of process mining. In Section 3 we will show the applicability of process mining in the healthcare domain using data obtained for a group of 627 gynecological oncology patients. Section 4 concludes the paper.

2 PROCESS MINING

Process mining is applicable to a wide range of systems. These systems may be pure information systems (e.g., ERP systems) or systems where the hardware plays a more prominent role (e.g., embedded systems). The only requirement is that the system produces *event logs*, thus recording (parts of) the actual behavior.

An interesting class of information systems that produce event logs are the so-called *Process-Aware Information Systems* (PAISs) (Dumas et al., 2005). Examples are classical workflow management systems (e.g. Staffware), ERP systems (e.g. SAP), case handling systems (e.g. FLOWer), PDM systems (e.g. Windchill), CRM systems (e.g. Microsoft Dynamics CRM), middleware (e.g., IBM's WebSphere), hospital information systems (e.g., Chipsoft), etc. These systems provide very detailed information about the activities that have been executed.

However, not only PAISs are recording events. Also, in a typical hospital there is a wide variety of systems that record events. For example, in an intensive care unit, a system can record which examinations or treatments a patient undergoes and also it can record occurring complications for a patient. For a radiology department the whole process of admittance of a patient till the archival of the photograph can be recorded. However, frequently these systems are limited to one department only. However, systems used for billing purposes have to ensure that all services delivered to the patient will be paid. In order for these systems to work properly, information from different systems needs to be collected so that it is clear which activities have been performed in the care process of a patient. In this way, these systems within the hospital can contain information about processes *within* one department but also *across* departments. This information can be used for improving processes within departments itself or improving the services offered to patients.

The goal of process mining is to extract information (e.g., process models) from these logs, i.e., process mining describes a family of *a-posteriori* analysis techniques exploiting the information recorded in the event logs. Typically, these approaches assume that it is possible to sequentially record events such that each event refers to an activity (i.e., a well-defined step in the process) and is related to a particular case (i.e., a process instance). Furthermore, some mining techniques use additional information such as the performer or originator of the event (i.e., the person/resource executing or initiating the activity), the timestamp of the event, or data elements recorded with the event (e.g., the size of an order).

Process mining addresses the problem that most "process/system owners" have limited information about what is actually happening. In practice, there is often a significant gap between what is prescribed or supposed to happen, and what *actually* happens. Only a concise assessment of reality, which process mining strives to deliver, can help in verifying process models, and ultimately be used in system or process redesign efforts.

The idea of process mining is to discover, monitor and improve real processes (i.e., not assumed processes) *by extracting knowledge from event logs.* We consider three basic types of process mining (Figure 1): (1) *discovery*, (2) *conformance*, and (3) *extension*.

Discovery: Traditionally, process mining has been focusing on *discovery*, i.e., deriving information about the original process model, the organizational context, and execution properties from enactment logs. An example of a technique addressing the

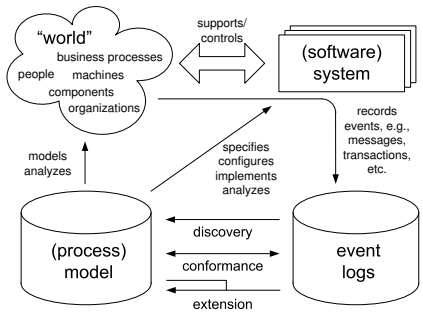

Figure 1: Three types of process mining: (1) Discovery, (2) Conformance, and (3) Extension.

control flow perspective is the α-algorithm (van der Aalst et al., 2004), which constructs a Petri net model describing the behavior observed in the event log. It is important to mention that there is no a-priori model, i.e., based on an event log some model is constructed. However, process mining is not limited to process models (i.e., control flow) and recent process mining techniques are more and more focusing on other perspectives, e.g., the organizational perspective, performance perspective or the data perspective. For example, there are approaches to extract social networks from event logs and analyze them using social network analysis (van der Aalst et al., 2005). This allows organizations to monitor how people, groups, or software/system components are working together. Also, there are approaches to visualize performance related information, e.g. there are approaches which graphically shows the bottlenecks and all kinds of performance indicators, e.g., average/variance of the total flow time or the time spent between two activities.

Conformance: There is an a-priori model. This model is used to check if reality conforms to the model. For example, there may be a process model indicating that purchase orders of more than one million Euro require two checks. Another example is the checking of the so-called "four-eyes" principle. Conformance checking may be used to detect deviations, to locate and explain these deviations, and to measure the severity of these deviations.

Extension: There is an a-priori model. This model is extended with a new aspect or perspective, i.e., the goal is not to check conformance but to enrich the model with the data in the event log. An example is the extension of a process model with performance data, i.e., some a-priori process model is used on which bottlenecks are projected.

At this point in time there are mature tools such as the ProM framework (van der Aalst et al., 2007b), featuring an extensive set of analysis techniques which can be applied to real-life logs while supporting the whole spectrum depicted in Figure 1.

3 HEALTHCARE PROCESS

In this section, we want to show the *applicability* of process mining in healthcare. However, as healthcare processes are characterized by the fact that *several organizational units* can be involved in the treatment process of patients and that these organizational units often have their own specific IT applications, it becomes clear that getting data, which is related to healthcare processes, is not an easy task. In spite of this, systems used in hospitals need to provide an integrated view on all these IT applications as it needs to be guaranteed that the hospital gets paid for every service delivered to a patient. Consequently, these kind of systems contain process-related information about healthcare processes and are therefore an interesting candidate for providing the data needed for process mining.

To this end, as case study for showing the applicability of process mining in health care, we use raw data collected by the billing system of the AMC hospital. This raw data contains information about a group of 627 gynecological oncology patients treated in 2005 and 2006 and for which all diagnostic and treatment activities have been recorded. The process for gynecological oncology patients is supported by several different departments, e.g. gynecology, radiology and several labs.

For this data set, we have extracted event logs from the AMC's databases where each event refers to a service delivered to a patient. As the data is coming from a billing system, we have to face the interesting problem that for each service delivered for a patient it is only known on which *day* the service has been delivered. In other words, we do not have any information about the actual timestamps of the start and completion of the service delivered. Consequently, the ordering of events which happen on the same day do not necessarily conform with the order in which events of that day were executed.

Nevertheless, as the log contains *real* data about the services delivered to gynecological oncology patients it is still an interesting and representative data set for showing the applicability of process mining in healthcare as still many techniques can be applied. Note that the log contains 376 different event names which indicates that we are dealing with a non-trivial careflow process.

In the remainder of this section we will focus on obtaining, in an explorative way, *insights into the gynecological oncology healthcare process*. So, we will

only focus on the *discovery* part of process mining, instead of the *conformance* and *extension* part. Furthermore, obtaining these insights should not be limited to one perspective only. Therefore, in sections 3.2.1, 3.2.2 and 3.2.3, we focus on the discovery of *care paths followed by patients*, the discovery of *organizational aspects* and the discovery of *performance related information*, respectively. This also demonstrates the diversity of process mining techniques available. However, as will be discussed in Section 3.1, we first need to perform some preprocessing before being able to present information on the right level of detail.

3.1 PREPROCESSING OF LOGS

The log of the AMC hospital contains a huge amount of distinct activities, of which many are rather low level activities, i.e., events at a low abstraction level. For example, for our purpose, the logged lab activities are at a too low abstraction level, e.g. determination of chloride, lactic acid and erythrocyte sedimentation rate (ESR). We would like to consider all these low level lab tests as a single lab test. Mining a log that contains many distinct activities would result in a too detailed spaghetti-like model, that is difficult to understand. Hence, we first apply some preprocessing on the logs to obtain interpretable results during mining. During preprocessing we want to "simplify" the log by removing the excess of low level activities. In addition, our goal is to consider only events at the department level. In this way, we can, for example, focus on care paths and interactions between departments. We applied two different approaches to do this.

Our first approach is to detect a *representative* for the lower level activities. In our logs, this approach can be applied to the before mentioned lab activities. In the logs we can find an activity that can serve as representative for the lab activities, namely the activity that is always executed when samples are offered to the lab. All other (low level) lab activities in the log are simply discarded.

The log also contains groups of low level activities for which there is no representative. For instance at the radiology department many activities can occur (e.g., echo abdomen, thorax and CT brain), but the logs do not contain a single event that occurs for every visit to this department, like a registration event for example. We apply *aggregation* for low level activities in groups without a representative by (1) defining a representative, (2) mapping all activities from the group to this representative and (3) removing repetitions of events from the log. For example, for the radiology department we define "radiology" as representative. A log that originally contains "`...,ultrasound scan abdomen, chest X-ray, CT scan brain,...`", would contain "`...,radiology,...`", after mapping low level radiology activities to this representative and removing any duplicates.

3.2 MINING

In this section, we present some results obtained through a detailed analysis of the ACM's event log for the gynecological oncology process. We concentrate on the discovery part to show actual situations (e.g. control flows, organizational interactions) in the healthcare process. More specifically, we elaborate on mining results based on three major perspectives (i.e. control flow, organizational, performance perspectives) in process mining.

3.2.1 CONTROL FLOW PERSPECTIVE

One of the most promising mining techniques is control flow mining which automatically derives process models from process logs. The generated process model reflects the actual process as observed through real process executions. If we generate process models from healthcare process logs, they give insight into care paths for patients. Until now, there are several process mining algorithms such as the α-mining algorithm, heuristic mining algorithm, region mining algorithm, etc (van der Aalst et al., 2004; Weijters and van der Aalst, 2003; van Dongen et al., 2007). In this paper, we use the heuristic mining algorithm, since it can deal with noise and exceptions, and enables users to focus on the main process flow instead of on every detail of the behavior appearing in the process log (Weijters and van der Aalst, 2003). Figure 2 shows the process model for all cases obtained using the Heuristics Miner. Despite its ability to focus on the most frequent paths, the process, depicted in Figure 2, is still spaghetti-like and too complex to understand easily.

Since processes in the healthcare domain do not have a single kind of flow but a lot of variants based on patients and diseases, it is not surprising that the derived process model is spaghetti-like and convoluted.

One of the methods for handling this problem is breaking down a log into two or more sub-logs until these become simple enough to be analyzed clearly. We apply clustering techniques to divide a process log into several groups (i.e. clusters), where the cases in the same cluster have similar properties. Clustering is a very useful technique for logs which contain many cases following different procedures, as is the

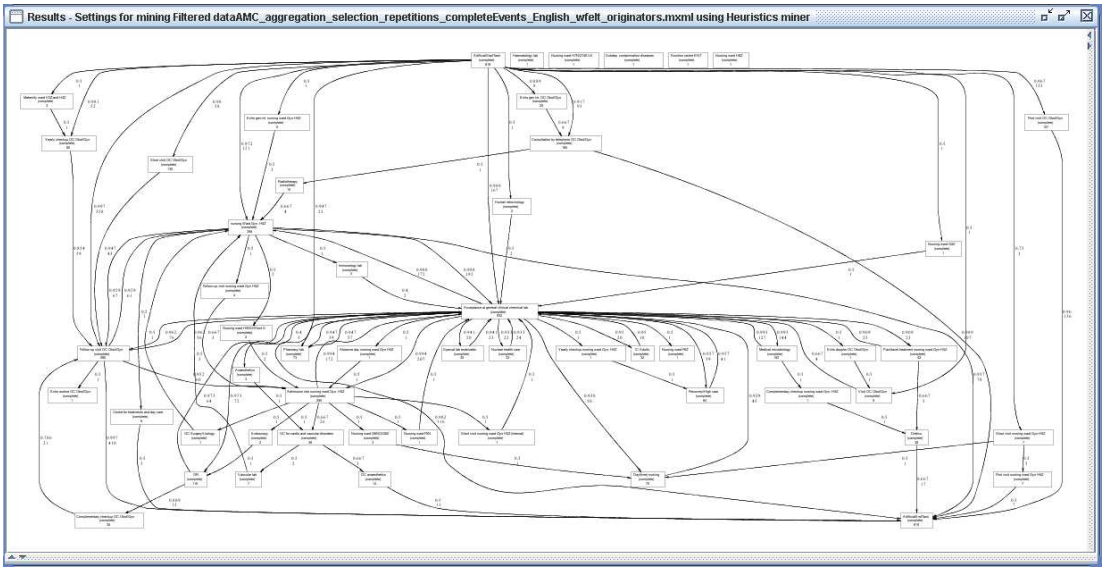

Figure 2: Derived process model for all cases.

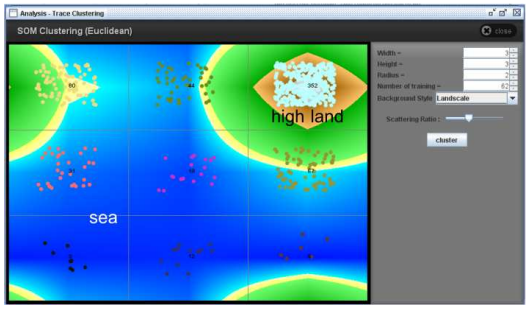

Figure 3: Log clustering result.

usual case in healthcare systems. Depending on the interest (e.g., exceptional or frequent procedures), a cluster can be selected. There are several clustering techniques available. Among these, we use the SOM (Self Organizing Map) algorithm to cluster the log because of its performance (i.e., speed). Figure 3 shows the clustering result obtained by applying the Trace Clustering plug-in. Nine clusters are obtained from the log. In the figure, the instances in the same cell belong to the same cluster. The figure also shows a contour map based on the number of instances in each cell. It is very useful to take a quick glance at the clusters – are there clusters with many similarities (high land), or are there many clusters with exceptional cases (sea).

By using this approach, we obtained several clusters of reasonable size. In this paper we show only the result for the biggest cluster, containing 352 cases all with similar properties. Figure 4 shows the heuristic net derived from the biggest cluster. The result is much simpler than the model in Figure 2. Furthermore, the fitness of this model is “good”. The model represents the procedure for most cases in the cluster, i.e., these cases “fit” in the generated process model. A closer inspection of this main cluster by domain experts confirmed that this is indeed main stream followed by most gynecological oncology patients.

When discussing the result with the people involved in the process, it was noted that patients, referred to the AMC by another hospital, only visit the outpatient clinic once or twice. These patients are already diagnosed, and afterwards they are referred to another department, like radiotherapy, for treatment and which is then responsible for the treatment process. Also, very ill patients are immediately referred to another department for treatment after their first visit.

3.2.2 ORGANIZATIONAL PERSPECTIVE

There are several process mining techniques that address organizational perspective, e.g., organizational mining, social network mining, mining staff assignment rules, etc. (van der Aalst et al., 2005). In this paper, we elaborate on social network mining to provide insights into the collaboration between departments in the hospital. The Social Network Miner allows for the discovery of social networks from process logs. Since there are several social network analysis techniques and research results available, the generated social network allows for analysis of social relations between originators involving process executions. Fig-

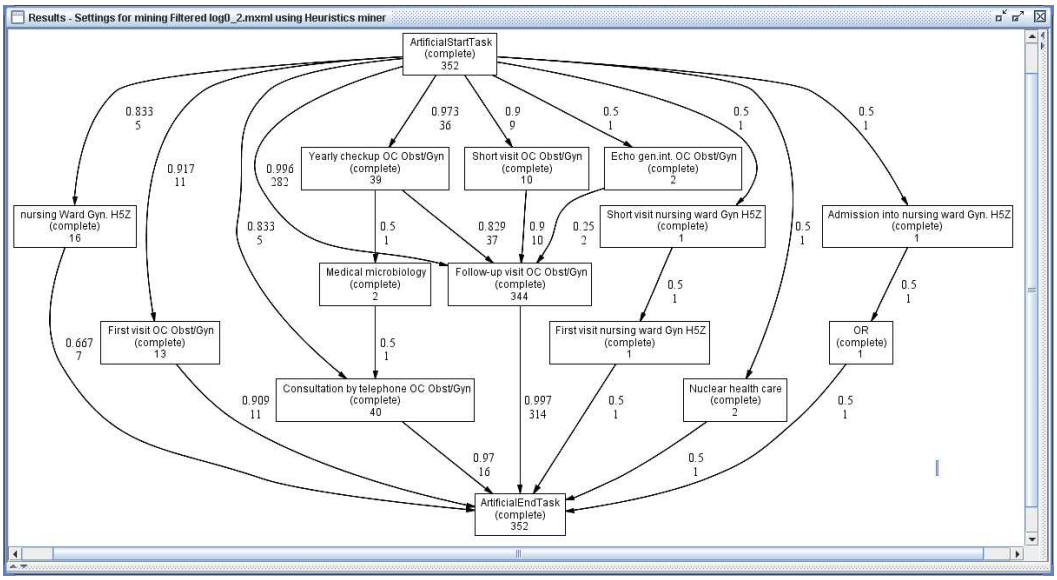

Figure 4: Process model from the biggest cluster.

ure 5 shows the derived social network. To derive the network, we used the *Handover of Work* metric (van der Aalst et al., 2005) that measures the frequency of transfers of work among departments.

The network shows the relationships between originators above a certain threshold. Originators, for which all relationships are below the specific threshold, appear as isolated circles. The originators that were highly involved in the process appear as larger dots in the figure. These results are useful to detect whether there are frequent interactions between originators (departments, in our case). In hospitals there are many departments that interact and hand over work to each other. The mining result shows that the general clinical chemical lab is highly involved in the process and interacts with many departments. The outpatient clinic (OC) for gynecology and obstetrics is also often involved, but is not directly connected to all other departments. For instance there is no relationship (within this threshold) between this OC and the vascular lab. This means that there is no, or not much, interaction between these two departments.

When this result was presented to the people involved in the process, they confirmed the strong collaboration with the departments shown in Figure 5. However, they were surprised about the rather strong collaboration with the dietics department. Nevertheless, this can be explained by the fact that, when a patient has to go to several chemotherapy sessions, then a visit to the dietician is also often needed.

Moreover, they also noted that the many interactions between the lab and other departments is misleading as all the examinations are requested by gynecological oncology and not by the lab. This can be explained by the many lab tests and resulting interactions between the lab and other departments.

3.2.3 PERFORMANCE PERSPECTIVE

Process mining provides several performance analysis techniques. Among these, the dotted chart is a method suitable for case handling processes which are flexible and knowledge intensive business processes and focus not on the routing of work or the activities but on the case (e.g. careflows). In this paper, we use the dotted chart to show overall events and performance information of the log. Figure 6 shows the dotted chart. In the chart, events are displayed as dots, and the time is measured along the horizontal axis of the chart. The vertical axis represents case IDs and events are colored according to their task IDs. It supports several time options such as actual, relative, logical, etc. In the diagram, we use relative time which shows the duration from the beginning of an instance to a certain event. Thus it indicates the case duration of each instance. It also provides performance metrics such as the time of the first and of the last events, case durations, the number of events in an instance, etc. For example, in the figure (top right, average spread in seconds), the average case duration is about 49 days.

Users can obtain useful insights from the chart, e.g., it is easy to find interesting patterns by looking at the dotted chart. In Figure 6, the density of events on the left side of the diagram is higher than the density of those on the right side. This shows that initially pa-

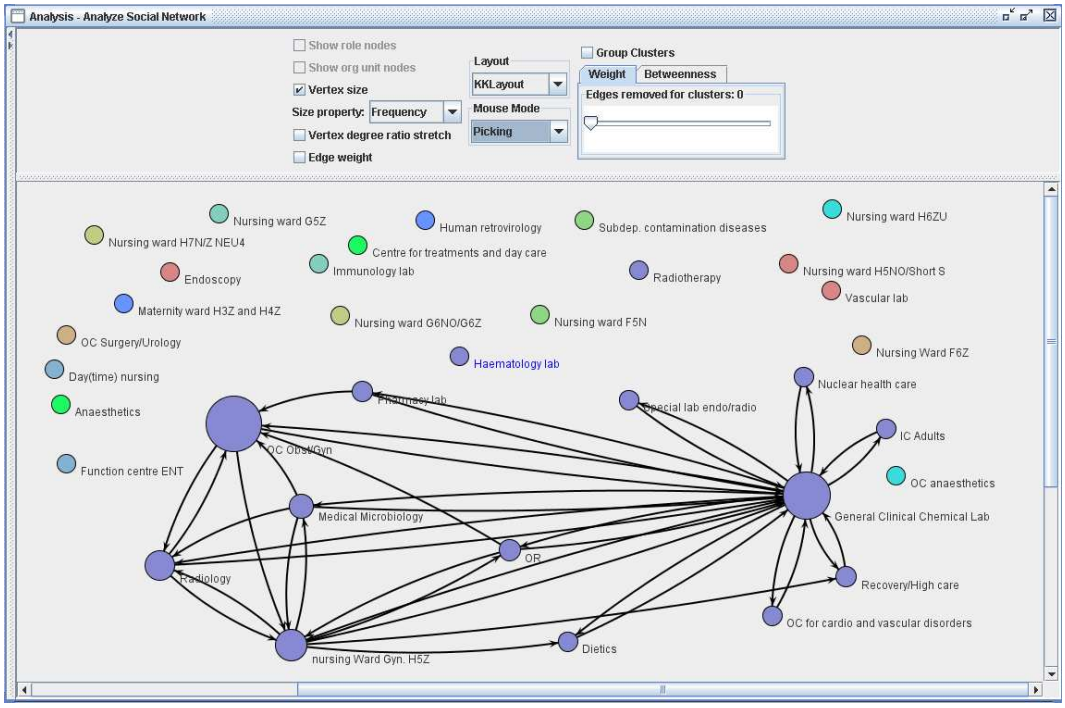

Figure 5: Social network (handover of work metrics).

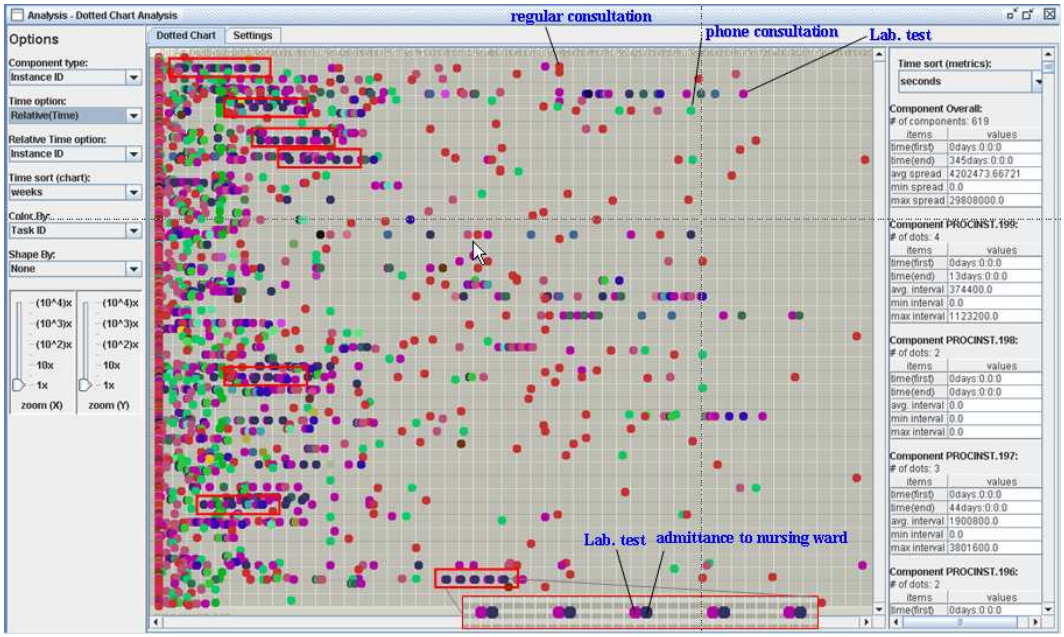

Figure 6: Dotted Chart.

tients have more diagnosis and treatment events than in the later parts of the process. When we focus on the long duration instances (i.e. the instances having events in the right side of the diagram), it can be observed that they mainly consist of regular consultation (red dot), consultation by phone (red dot), and lab test (violet dot) activities. It reflects the situation that patients have regular consultation by visiting or being phoned by the hospital and sometimes have a test after or before the consultation. It is also easy to discover

patterns in the occurrences of activities. For example, seven instances have the pattern that consists of a lab test and an admittance to the nursing ward activities.

When the results were presented to the people involved in the process, they confirmed the patterns that we found. Furthermore, for the last pattern they indicated that the pattern deals about patients who get a chemotherapy regularly. The day before, they come for a lab test and when the result is good, they get the next chemotherapy.

4 CONCLUSION

In this paper, we have focussed on the applicability of process mining in the healthcare domain. For our case study, we have used data coming from a non-trivial care process of the AMC hospital. We focussed on obtaining insights into the careflow by looking at the control-flow, organizational and performance perspective. For these three perspectives, we presented some initial results. We have shown that it is possible to mine complex hospital processes giving insights into the process. In addition, with existing techniques we were able to derive *understandable* models for large groups of patients. This was also confirmed by people of the AMC hospital.

Furthermore, we compared our results with a flowchart for the diagnostic trajectory of the gynaecological oncology healthcare process, and where a top-down approach had been used for creating the flowchart and obtaining the logistical data (Elhuizen et al., 2007). With regard to the flowchart, comparable results have been obtained. However, a lot of effort was needed for creating the flowchart and obtaining the logistical data, where with process mining there is the opportunity to obtain these kind of data in a semi-automatic way.

Unfortunately, traditional process mining approaches have problems dealing with unstructured processes as, for example, can be found in a hospital environment. Future work will focus on both developing *new* mining techniques and on using *existing* techniques in an innovative way to obtain understandable, high-level information instead of "spaghetti-like" models showing all details. Obviously, we plan to evaluate these results in healthcare organizations such as the AMC.

ACKNOWLEDGEMENTS

This research is supported by EIT, NWO-EW, the Technology Foundation STW, and the SUPER project (FP6). Moreover, we would like to thank the many people involved in the development of ProM.

REFERENCES

Anyanwu, K., Sheth, A., Cardoso, J., Miller, J., and Kochut, K. (2003). Healthcare Enterprise Process Development and Integration. *Journal of Research and Practice in Information Technology*, 35(2):83–98.

Dumas, M., van der Aalst, W., and ter Hofstede, A. (2005). *Process-Aware Information Systems: Bridging People and Software through Process Technology*. Wiley & Sons.

Elhuizen, S., Burger, M., Jonkers, R., Limburg, M., Klazinga, N., and Bakker, P. (2007). Using Business Process Redesign to Reduce Wait Times at a University Hospital in the Netherlands. *The Joint Commission Journal on Quality and Patient Safety*, 33(6):332–341.

Lenz, R., Elstner, T., Siegele, H., and Kuhn, K. (2002). A Practical Approach to Process Support in Health Information Systems. *Journal of the American Medical Informatics Association*, 9(6):571–585.

van der Aalst, W., Reijers, H., and Song, M. (2005). Discovering Social Networks from Event Logs. *Computer Supported Cooperative Work*, 14(6):549–593.

van der Aalst, W., Reijers, H., Weijters, A., van Dongen, B., de Medeiros, A. A., Song, M., and Verbeek, H. (2007a). Business process mining : an industrial application. *Information Systems*, 32(5).

van der Aalst, W., van Dongen, B., Günther, C., Mans, R., de Medeiros, A. A., Rozinat, A., Rubin, V., Song, M., Verbeek, H., and Weijters, A. (2007b). ProM 4.0: Comprehensive Support for Real Process Analysis. In Kleijn, J. and Yakovlev, A., editors, *Application and Theory of Petri Nets and Other Models of Concurrency (ICATPN 2007)*, volume 4546 of *Lecture Notes in Computer Science*, pages 484–494. Springer-Verlag, Berlin.

van der Aalst, W., van Dongen, B., Herbst, J., Maruster, L., Schimm, G., and Weijters, A. (2003). Workflow Mining: A survey of Issues and Approaches. *Data and Knowledge Engineering*, 47(2).

van der Aalst, W., Weijters, A., and Maruster, L. (2004). Workflow Mining: Discovering Process Models from Event Logs. *IEEE Transactions on Knowledge and Data Engineering*, 16(9):1128–1142.

van Dongen, B., Busi, N., Pinnaand, G., and van der Aalst, W. (2007). An Iterative Algorithm for Applying the Theory of Regions in Process Mining. BETA Working Paper Series, WP 195, Eindhoven University of Technology, Eindhoven.

Weijters, A. and van der Aalst, W. (2003). Rediscovering Workflow Models from Event-Based Data using Little Thumb. *Integrated Computer-Aided Engineering*, 10(2):151–162.