Analysis of Patient Treatment Procedures
The BPI Challenge Case Study

R.P. Jagadeesh Chandra Bose[1,2] and Wil M.P. van der Aalst[1]

[1] Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands
[2] Philips Healthcare, Veenpluis 5–6, Best, The Netherlands
{j.c.b.rantham.prabhakara,w.m.p.v.d.aalst}@tue.nl

Abstract. A real-life event log, taken from a Dutch Academic Hospital, is analyzed using process mining techniques. The log contains events related to treatment and diagnosis steps for patients diagnosed with cancer. Given the heterogeneous nature of these cases, we first demonstrate that it is possible to create more homogeneous subsets of cases (e.g. patients having a particular type of cancer that need to be treated urgently). Such preprocessing is crucial given the variation and variability found in the event log. The discovered homogeneous subsets are analyzed using state-of-the-art process mining approaches. In this paper, we report on the findings discovered using *enhanced fuzzy mining* and *trace alignment*. A dedicated preprocessing ProM[3] plug-in was developed for this challenge. The analysis was done using recent, but pre-existing, ProM plug-ins. As the evaluation shows, this approach is able to uncover many interesting findings and could be used to improve the underlying care processes.

1 Introduction

Process mining provides a new means to improve processes in a variety of application domains. There are two main drivers for this new technology. On the one hand, more and more events are being recorded thus providing detailed information about the history of process instances. On the other hand, stakeholders would like to get insights into their operational processes. These insights should be based on facts and also provide actionable information.

There are basically three types of process mining [1]. The first type of process mining is *discovery*. A discovery technique takes an event log and produces a model without using any a-priori information. An example is the α-algorithm that is able to discover a Petri net based on sequence of events [2]. The second type of process mining is *conformance*. Here, an existing process model is compared with an event log of the same process. Conformance checking can be

[3] ProM is an extensible framework that provides a comprehensive set of tools/plug-ins for the discovery and analysis of process models from event logs. See http://www.processmining.org for more information and to download ProM.

used to check if reality, as recorded in the log, conforms to the model and vice versa [3]. The third type of process mining is *enhancement*. Here, the idea is to extend or improve an existing process model using information about the actual process recorded in some event log [1]. Whereas conformance checking measures the alignment between model and reality, this third type of process mining aims at changing or extending the a-priori model.

Over the last decade event data has become readily available and process mining techniques have matured. Moreover, process mining algorithms have been implemented in various academic and commercial systems. Given the increasing maturity it becomes important to compare the different approaches. It is no longer acceptable to provide a new process discovery algorithm without some form of evaluation or comparison. The *Business Process Intelligence Challenge* (BPI Challenge 2011) is an initiative to show the applicability of existing approaches. The event log provided (doi:10.4121/uuid:d9769f3d-0ab0-4fb8-803b-0d1120ffcf54) serves as a benchmark for comparing the different approaches.

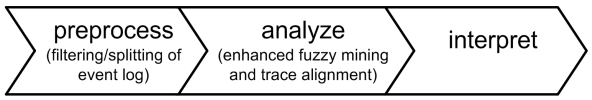


Fig. 1. Overview of the approach followed.

Figure 1 provides a high-level view on the approach used to obtain the results reported in this paper. We spent considerable efforts and time on preprocessing the event log provided for the BPI Challenge 2011. Given the diversity of process instances in the event log and the presence of data attributes related to diagnosis and treatment, we developed a dedicated filtering approach (added as a new plug-in to ProM). Based on properties such as the treatment code, diagnosis code, specialism, organizational data, trace length, time-perspective, and urgency we filter and split the event log into smaller, more homogeneous, event logs. After preprocessing the event log, we apply various process mining techniques (available as plug-ins in ProM 6). In particular, we apply the *enhanced fuzzy miner* [4] and *trace alignment* [5]. The enhanced fuzzy miner provides more control over the discovery process than the classical fuzzy miner [6]. The result is a hierarchical process model where the properties mentioned before (e.g., organizational unit) are exploited. Trace alignment can be used in the initial phase of analysis, i.e., traces are inspected to see the dominant patterns. Moreover, trace alignment can be used to find deviations and it is possible to answer specific questions by inspecting process instances. As Figure 1 shows, the analysis results need to be interpreted. In this paper, we discuss some of our findings. However, because we were not in contact with the stakeholders of the Dutch Academic Hospital, we are unable to interpret and validate some of our findings. Note that process mining projects were typically iterative; stakeholders provide

feedback on findings thus triggering a new round of analysis.

The remainder of this paper is organized as follows. Section 2 shows how the event log was filtered and split based on a variety of properties. Section 3 discusses the various plug-ins developed and used for our analysis. The experimental results are discussed in Section 4. Section 5 concludes the paper.

2 Dissecting the Event Log

The event log contains information on the activities pertaining to the treatment procedures that are being administered on the patients in a hospital. The log contains events related to a *heterogeneous* mix of patients diagnosed with cancer (at different stages of malignancy) pertaining to the cervix, vulva, uterus and ovary. Each case in the event log corresponds to a patient. The raw event log[4] (provided for the challenge) contains 1143 cases and 150291 events referring to 624 distinct activities. A naive attempt at analyzing this raw event log using existing process mining techniques is bound to provide results that are incomprehensible and unsatisfactory e.g., control-flow analysis on the log generates a workflow that is a graph containing a few hundred nodes (each node corresponding to an activity) and edges connecting the nodes based on their dependency. Graphs become quickly overwhelming and unsuitable for human perception and cognitive systems even if there are a few dozens of nodes [7].

We advocate the *preprocessing of the log* as an *essential* step in gaining meaningful insights. The event log contains rich information stored as attributes both at the event level and at the case level. We exploit this information and propose a few perspectives for preprocessing.

2.1 Diagnosis Perspective

Each case contains a few attributes that provide information on the illness the patient is diagnosed with. These attributes can be broadly classified into two categories (i) diagnosis code and (ii) diagnosis. Each case may contain up to 16 attributes of each type (e.g., Diagnosis code, Diagnosis code:1, Diagnosis code:2, ..., Diagnosis code:15 of the type diagnosis code and Diagnosis, Diagnosis:1, Diagnosis:2, ..., Diagnosis:15 of the type diagnosis). 1136 of the 1143 cases have at least one of these attributes. The diagnosis code attributes take on values such as M11, M12, 106 and the diagnosis attributes take on values such as 'Squamous cell ca cervix st IIb' (Squamous cell carcinoma of the cervix at stage IIb of malignancy). We make an assumption that the diagnosis code and diagnosis attributes have a correspondence between them as illustrated in Figure 2.

[4] The activity names and attribute values provided in the log are in Dutch. We have translated these names and values from Dutch to English. Henceforth, we report our analysis on the translated log.

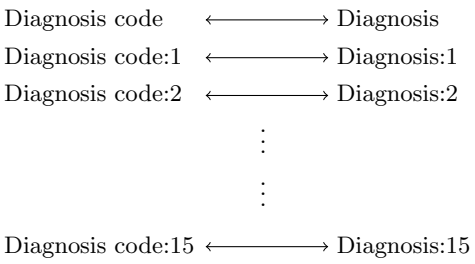


Fig. 2. Correspondence between diagnosis code and diagnosis attributes for each case.

The event log captures treatment procedures pertaining to 11 different diagnosis codes as summarized in Table 1. Figure 3 depicts the Venn diagram of the various diagnosis codes and the regions to which they are associated.

Table 1: Description of diagnosis of patients. The stage of malignancy is not known for some cases in all the diagnosis categories.

Diagnosis Code	Diagnosis Description
M11	Pertains to the cancer of the vulva. Describes information on cases diagnosed with – squamous cell carcinoma (stages I, II, III1, III2, IVa and IVb) – malignant neoplasms and melanoma – basal cell carcinoma – borderline malignancy
M12	Pertains to the cancer of the vagina. Describes information on cases diagnosed with – squamous cell carcinoma (stages II, III and IVb) – malignant neoplasms – adenocarcinoma (stage II) Certain metastases cases are also included.
M13	Pertains to the cancer of the cervix (uteri). Describes information on cases diagnosed with – squamous cell carcinoma (stages Ia1, Ia2, Ib, IIa, IIb, IIIb, IVa and IVb) – malignant neoplasms – adenocarcinoma (stages Ia1, Ib and IIa) – borderline malignancy – sarcoma

M14	Pertains to the cancer of the corpus uteri. Describes information on cases diagnosed with – adenocarcinoma (stages Ia, Ib, Ic, IIa, IIb, IIIa, IIIb, IVa and IVb) – malignant neoplasms and endometrium – clear cell carcinoma (stages Ib and IIIb) – borderline malignancy Certain metastases cases are also included.
M15	Primarily pertains to the cancer of the corpus uteri of type *sarcoma* (stages II and III according to the FIGO staging system). However, certain cases of colon cancer and myometrium are also classified into this category
M16	Pertains to the cancer of the ovary. Describes information on cases diagnosed with – adenocarcinoma of types • serous (stages Ia, Ic, IIa, IIIb, IIIc and IV) • endometroid (stages Ic, IIIc) • mucinous (stages Ic, IIc and IIIc) • non-differentiated (stages IIIc and IV) – non-epithelial malignancy (stages Ia, IIa, IIIa and IIIc) – neoplasms – borderline malignancy – clear cell carcinoma. Certain metastases cases are also included.
821	Pertains to the cancer of the ovary. Describes information on cases diagnosed with – adenocarcinoma of types • serous (stage IIIc) • mucinous (stage IIIc) – non-epithelial malignancy – neoplasms
822	Pertains to the cancer of the cervix (uteri). Describes information on cases diagnosed with – squamous cell carcinoma (stage Ib) – adenocarcinoma (stages IIa and IIb) – borderline malignancy – malignant neoplasms

106	Describes a heterogeneous mix of cases pertaining to the cancers of – cervix uteri, of types squamous cell carcinoma (stages Ia and IIa), malignant neoplasms and borderline malignancy – vulva, of types squamous cell carcinoma (stages III2, IVa and IVb) and malignant melanoma – corpus uteri, of types adenocarcinoma (stages Ib, Ic and IIa), malignant neoplasms and borderline malignancy – vagina, endometrium and ovarian tube
823	Describes a heterogeneous mix of cases pertaining to the cancers of – corpus uteri, of types adenocarcinoma (stages IVa and IVb), malignant neoplasms and sarcoma (stage IVb according to the FIGO staging system) – ovary, of type serous adenocarcinoma (stage IIIc) – endometrium
839	Describes a heterogeneous mix of cases pertaining to the cancers of – ovary, of types serous adenocarcinoma (stages IIIc and IV) and borderline malignany – uterine appendages, of type malignant neoplasms – vulva, of type malignant neoplasms

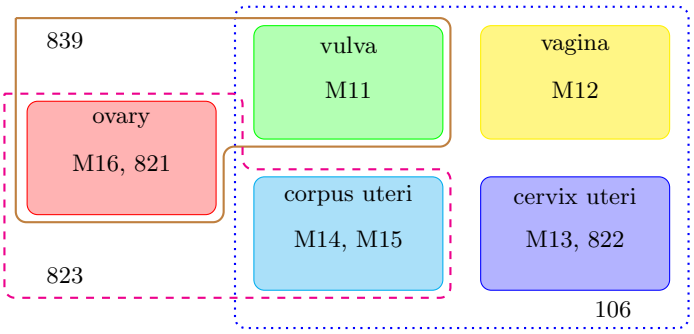


Fig. 3. Diagnosis codes and the regions to which they are associated.

Due to the different types of cancer and differences among patients, the cases found in the event log can be characterized as *heterogeneous*. This heterogeneity adds to the complexity of analysis and the results are often misleading. It has been argued that process mining results can be improved by partitioning an event

log into subsets of homogenous cases and analyzing these subsets independently [8]. We now propose a few means of segregating homogenous cases based on the diagnosis perspective.

- **Individual Diagnosis (Code):** As mentioned earlier, the cases in the event log contain multiple attributes of type 'Diagnosis code' and 'Diagnosis'. One can filter the event log based on a particular value for any of the diagnosis codes or diagnosis attributes e.g., consider cases where 'Diagnosis code:4 = M14', consider cases where the value for 'Diagnosis' is 'squamous cell carcinoma stage Ib'. One can also use a combination of both the diagnosis code and diagnosis to select cases for analysis e.g., cases where 'Diagnosis code = M11' and 'Diagnosis = basal cell carcinoma'.
- **Diagnosis (Code) Combination:** Since cases may contain multiple diagnosis code or diagnosis attributes, one can alternatively look at the combination of values for either of these attribute types. For example, a case can have {M13, 822, 106} as the set of values for the various diagnosis code attributes. If the event log is preprocessed for all such combinations, we can notice certain relationships between the diagnosis code combinations among the cases. For example, there might be distinct cases where the combinations {M13}, {M13, 822} and {M13, 822, 106} exist. We can clearly see a subsumption property between the values: {M13} $\subset$ {M13, 822} $\subset$ {M13, 822, 106}. The set of treatment procedures applied on patients with diagnosis code combination {M13, 822} typically includes the procedures applied to patients with diagnosis code {M13}. We can capture the relationships between the diagnosis code combinations manifested in the event log using a Hasse diagram by considering a partial ordering (with subsumption as the cover relation) on the code combinations. In the event log, there are a total of 38 distinct diagnosis code combinations. Table 2 in Appendix A depicts the distribution of cases over the different diagnosis code combinations. Figure 4 depicts the Hasse diagram corresponding to the code combinations involving M13 and other codes related to M13.

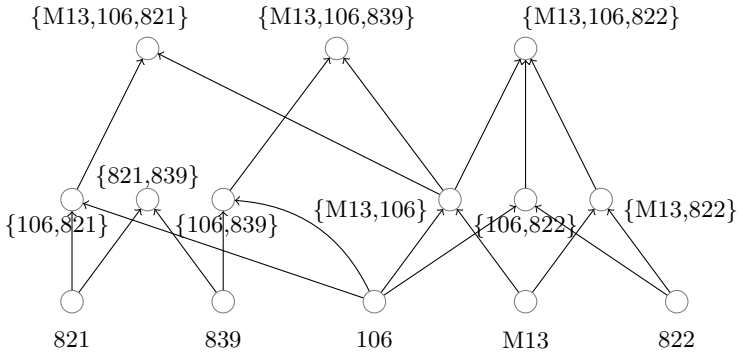


Fig. 4. Hasse diagram corresponding to the diagnosis code combinations involving M13 and related codes.

The nodes in the Hasse diagram form the basis for segregation of homogenous cases for analysis. Two types of node selection mechanisms can be adopted.

- *Single-node mode:* In this mode, one can consider all cases where the diagnosis code combination of the selected node is manifested e.g., choosing the node {M13, 106} implies considering only those cases pertaining to patients who have been diagnosed with both M13 and 106.
- *Sub-graph mode:* In this mode, multiple nodes can be selected. This implies the union of all cases where the diagnosis code combinations of the selected nodes are manifested are considered as homogenous e.g., selecting all the nodes subsumed under the maximal element {M13, 106, 822} implies considering all the cases pertaining to patients who have been diagnosed with either {M13, 106, 822} or {M13, 106} or {M13, 822} or {106, 822} or {M13} or {106} or {822} as homogenous.

We have used the 'Diagnosis code' attributes to explain the hierarchical grouping of cases. Obviously, one can use the 'Diagnosis' attributes in a similar fashion. One can also use both of them in tandem e.g., first segregate cases where the patients have been diagnosed for M11; partition these cases further into two clusters: one cluster containing cases where the 'Diagnosis' was either squamous cell carcinoma or neoplasms or melanoma and the other where the 'Diagnosis' was basel cell carcinoma or borderline malignancy.

2.2 Treatment Perspective

The event log also contains information corresponding to the treatment administered on the patients. However, unlike the diagnosis, only the 'treatment codes' are manifested in the log without the description for the codes. Each case may contain up to 16 treatment code attributes (e.g., Treatment code, Treatment code:1, Treatment code:2, ..., Treatment code:15). 1131 of the 1143 cases in the event log have at least one treatment code specified. There are a total of 46 distinct treatment codes and 236 distinct treatment code combinations in the event log. A vast majority of treatment code combinations are unique combinations.

One can also use the treatment codes as a basis for defining homogeneity of cases with the method illustrated using Hasse diagram shown in Figure 4. However, now treatment codes instead of diagnosis codes are used.

2.3 Time Perspective

There exists up to 16 attributes pertaining to the time perspective at the case level in the event log. The names of the attributes indicate that they signify the start and end dates (Start date, Start date:1, Start date:2, ..., Start date:15, End date, End date:1, End date:2, ..., End date:15). Each of these attributes has a correspondence to the diagnosis/treatment code specified for the case. In other words, if a case has k diagnosis codes, it also contains k start dates

and k end dates. Apart from these attributes at the case level, each event in a case has a timestamp that indicates when that particular activity has been executed. *However, our investigation could not uncover any meaningful correlation between the start/end dates specified at the case level and the actual event timestamps.* Therefore, we do not use the information available in the start/end date attributes for our analysis. Instead, we rely on the event timestamps.

We define the span period (or duration) of a case to be time difference between the last and first events of the case. Figure 5 depicts the histogram of the span period of the cases. It can be seen that the cases typically run over a long period of time (the average span period is 386 days and the standard deviation is 338 days). Since each case in the event log corresponds to a patient, one can interpret the long durations to be the time under which the patient is being treated. During this period, the patient could have visited or consulted the hospital several times. Therefore, it is rather unusual to consider all the events in a case as belonging to a single process instance for analysis. This calls for the definition of appropriate notions of process instances, i.e., we will split cases into smaller more representative cases.

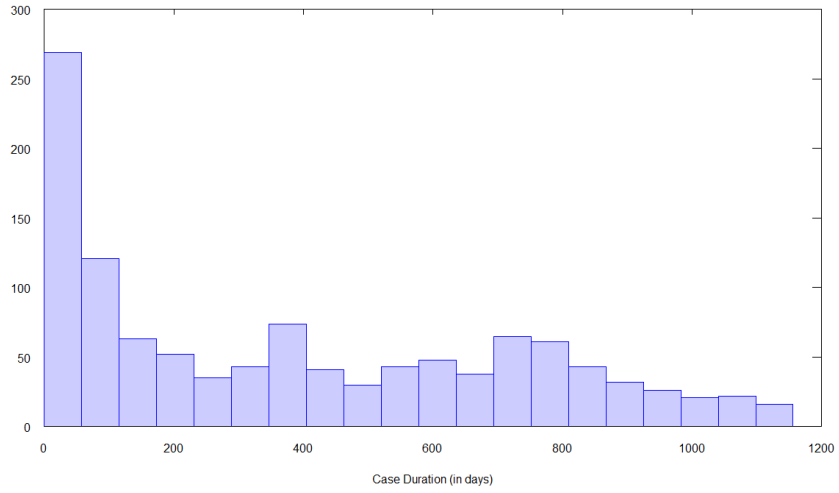


Fig. 5. Histogram depicting the span period of the cases.

A closer inspection of the log reveals that the activities in a case happen in bursts. Figure 6 depicts a typical scenario of activity bursts in cases. Bursts signify periods of consultation/treatment and idle periods denote recuperation e.g., a patient undergoing radiotherapy visits a hospital at regular intervals and there are recommended time intervals between successive applications of radiotherapy or between radiotherapy and surgery. We exploit this characteristic and define

a process instance to capture a burst of activities. One can use a parameter, say δ days, to demarcate the boundaries between process instances. Two events or event sequences with a time period between them greater than δ fall under two process instances as depicted in Figure 7.

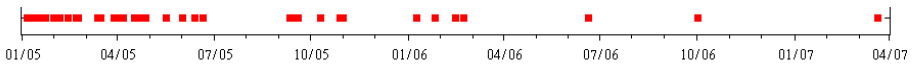


Fig. 6. The cases in the event log denote bursts of activities at certain points during the span period of the case. The x-axis denotes time with the format month/year. A dot can represent a multitude of events owing to the coarse-grained scale used in x-axis.

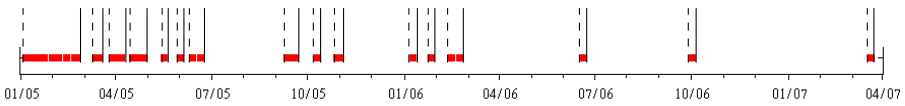


Fig. 7. Defining process instances within a case based on activity bursts and idle periods. The dotted lines indicate the start of a process instance and the immediate following solid line indicates the end of the process instance.

2.4 Using the Organizational Perspective to Derive Artifacts

Each event in the event log contains an attribute 'org:group' that captures the department/lab where the activity corresponding to the event was performed. There are 43 distinct org:group values (departments/labs) in the event log with one being 'unknown'. The process instances (bursts of events) defined above exhibit certain regularity with respect to the organization group. The regularity is often manifested as a related set of diagnosis tests in the form of a continuous series of activities, e.g., different diagnosis blood tests prescribed for a patient in the lab, as illustrated in Figure 8. This can be associated to the concept of *artifacts* in business processes. Nigam and Caswell [9] define an artifact to be any concrete, identifiable, self-describing chunk of information used in business processes.

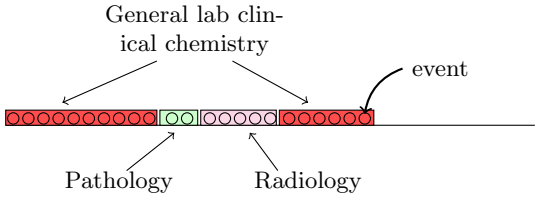


Fig. 8. Regularities in the form of series of activities (pertaining to related diagnosis tests) performed in a lab/department.

Exploiting this notion, we propose the transformation of the original log into an *abstraction log* where the activities correspond to the organization names. Each continuous sequence of one or more events pertaining to the same organization in the process instance of the original log is replaced by a single event with the organization name as its activity. At the same time, we create one sub-log for each organization whose process instances correspond to the replaced sequence of events. The process of transformation is illustrated in Figure 9. The process instance in Figure 8 is transformed into a process instance GPRG... where G, P and R are used as short-cuts for the organizations (departments) general lab clinical chemistry, pathology and radiology respectively. At the same time, we create three sub-logs, one for each of G, P and R. The process instances in these sub-logs correspond to the sequence of events replaced in the original process instance.

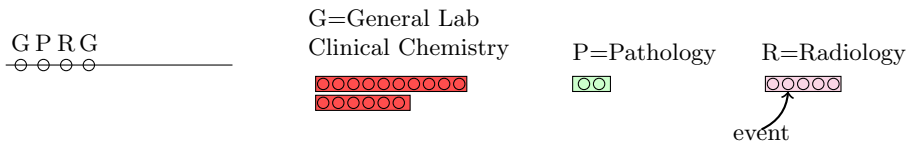


Fig. 9. Transformation of the original log into an abstraction log using the notion of artifacts on the organizational perspective. The activities in the abstraction log correspond to the organization names. Also, one sub-log is created for each organization.

2.5 Urgent and Non-urgent Cases

The event log contains certain activities that are classified as urgent. Ordinary counterparts to such activities also exist. For example, the activities 'haemoglobin photoelectric' and 'haemoglobin photoelectric-urgent' both exist (this activity corresponds to the estimation of haemoglobin using a photoelectric calorimeter). Similarly, the activities 'platelet count' and 'platelet count-urgent' both exist. There are a total of 28 urgent activities in the event log. This indicates that certain cases (patients) are considered as emergency cases and are treated in an expedited manner. We can exploit this information as well in order to segregate homogenous cases. We can partition a given log, say a log containing cases of patients with diagnosis code combination as {M11}, into two categories: urgent and non-urgent cases. Urgent cases are those cases where at least one activity of type urgent is manifested.

One can use the perspectives defined above either in isolation or in combinations as a preprocessing step to segregate homogenous cases based on the focus of analysis. We will give some examples of using this in Section 4.

3 Tools Used for Analysis

We focus on the control-flow and process diagnostics aspects and use the process mining tool, ProM, for our analysis. *We have built an event log preprocessing plug-in specifically for the challenge* and use other existing plug-ins for analysis.

3.1 Preprocessing

The concepts presented in this paper (segregation of homogenous cases using different perspectives) have been implemented as the 'BPI Challenge 2011' plug-in[5] in ProM. Figure 10 depicts the invocation of the plug-in. The plug-in takes an event log as input. The plug-in supports the case selection mechanisms based on the diagnosis and treatment code perspectives as discussed in Section 2. The plug-in also supports the creation of process instances based on the time perspective, transformation of the given log into an abstraction log based on the organizational perspective (selecting this option creates a parent log wherein the activity names correspond to the organization group (department/lab within the hospital) and sub-logs, one for each organization group) and the segregation of cases in the given log into urgent and non-urgent cases. These three features are embedded in the 'Transform Log' option and can be configured in the next step upon selecting this option. In addition, the plug-in also supports simple filters such as filtering traces based on their lengths (e.g., discard all traces that have less than 5 events) and filtering activities based on their frequency (e.g., remove all activities whose frequency of occurrence is 1). The plug-in also provides an option for merging a set of logs into a single log (the set of process instances in the merged log is the union of process instances in the input logs). This feature is handy if we want to merge previously filtered logs, e.g., logs filtered earlier based on code combinations {M13} and {M13, 106}. Figure 11 depicts the configuration panel for selecting the various modes for filtering.

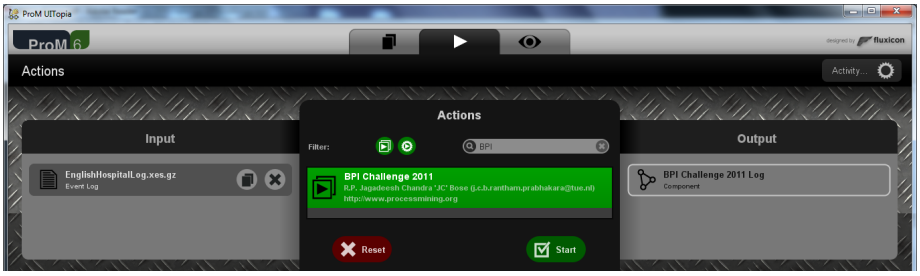


Fig. 10. The 'BPI Challenge 2011' plug-in

[5] This plug-in can be provided upon request.

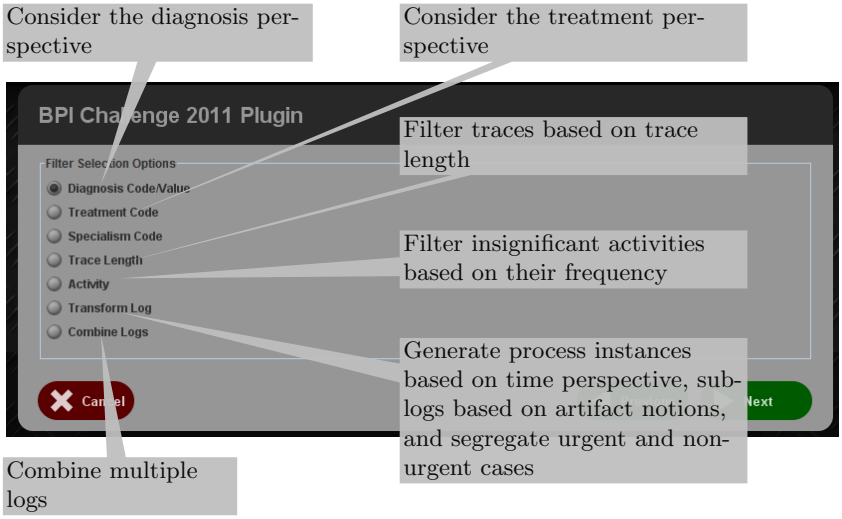


Fig. 11. Configuration options for selection of cases.

Figure 12 depicts the configuration panel for selecting cases based on the diagnosis perspective. The table in Figure 12 lists the individual diagnosis code combinations along with the number of cases with that code combination in the event log e.g., there are 201 patients diagnosed for M16 and there are 57 patients who have been diagnosed with the diagnosis code combination {M13, 106}.

3.2 Analysis

We use the enhanced Fuzzy Miner plug-in (to mine hierarchical workflow models) for control-flow analysis, the 'Guide Tree Miner' and the 'Trace Alignment With Guide Tree' plug-ins for process diagnostics. These plug-ins were developed in our earlier research [8, 10, 11, 4, 5].

– ***Enhanced Fuzzy Miner:*** The Fuzzy Miner [6] is a process discovery algorithm that mines an event log for a family of process models using a "map" metaphor (process models can be seen as the maps describing the operational processes of organizations). Recently, a two-phase approach to process discovery has been proposed to mine such process maps [4]. The first phase comprises of pre-processing the event log with abstractions at a desired level of granularity and the second phase deals with discovering the process maps with seamless zoom-in/out facility. Abstractions of lower level events are automatically formed by exploiting common patterns of execution in the event log [11]. The 'Pattern Abstractions' plug-in in ProM caters to the discovery of common execution patterns, the definition of abstractions over them, and the pre-processing of the event log with these abstractions. During the pre-processing phase, for each defined abstraction, the Pattern

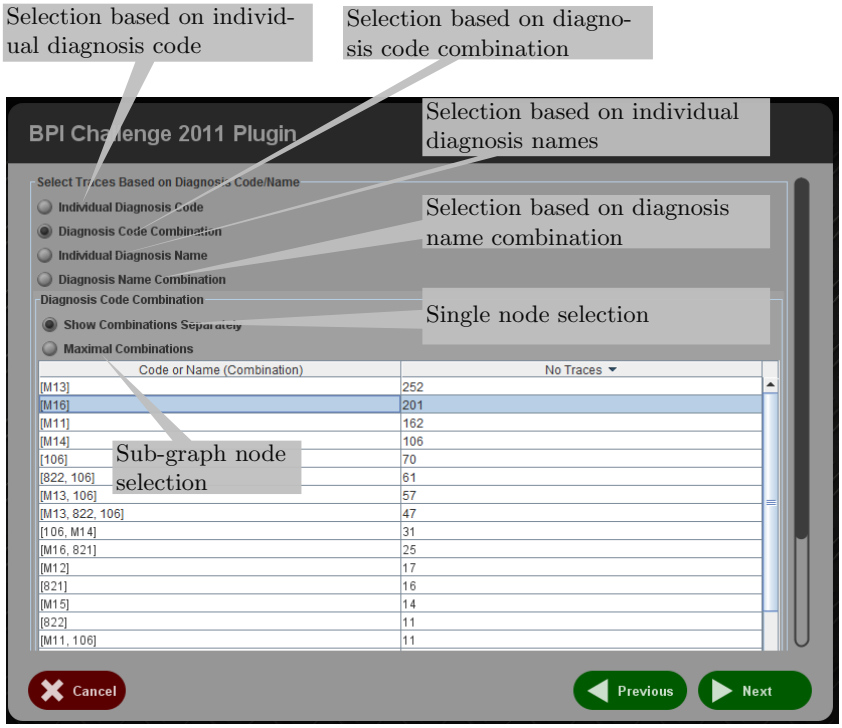


Fig. 12. Configuration options for selection of cases based on diagnosis code.

Abstractions plug-in generates a sub-log that captures the manifestation of execution patterns defined by that abstraction as its process instances. The Pattern Abstractions plug-in also generates a Hasse diagram for finding relationships.

The Fuzzy Miner plug-in in ProM has been enhanced to utilize the availability of sub-logs. Fuzzy models are discovered for each of the sub-logs and are displayed upon zooming in on its corresponding abstract activity in the parent model. The enhanced Fuzzy Miner in conjunction with the 'Pattern Abstractions' plug-in was shown to enable the discovery of hierarchical process models [12]. This enhancement is handy for our analysis based on the notion of artifacts on the organizational perspective.

– ***Trace Alignment With Guide Tree:*** Trace alignment has been proposed as a powerful technique for process diagnostics, especially when dealing with less structured processes [5]. If processes are less structured and event logs are far from complete, then it is better to *carefully inspect the event log by grouping and aligning the traces found in the event log.* By aligning traces we can see the common and frequent behavior, and distinguish this from the exceptional behavior. Trace alignment is a two step process: first, we

group similar traces in clusters [8, 10]; second, we visualize these clusters by aligning the traces [12].

The 'Trace Alignment With Guide Tree' plug-in in ProM implements this functionality. This plug-in depends on another plug-in, the 'Guide Tree Miner'. The Guide Tree Miner plug-in clusters the traces based on the techniques proposed in [8, 10].

4 Experimental Results and Discussion

In this section, we present and discuss our results/observations on the analysis of the event log.

4.1 Workflow of Treatment Procedures on Patients Diagnosed for M11

The event log provided for the challenge is subjected to the following preprocessing steps

- Select the cases whose diagnosis code combination is just {M11}. There are a total of 162 cases in the event log satisfying this criterion. This filtered event log contains 11280 events distributed over 207 activities.
- Segregate urgent and non-urgent cases from the filtered log obtained in the previous step. Of the 162 cases, 137 cases are non-urgent cases containing 6225 events referring to 143 activities while 25 cases are urgent cases containing 5055 events distributed over 173 activities. Let us call these two logs the *non-urgent cases log* and *urgent cases log*. It is interesting to note that though the number of urgent cases is just 25 (15%), they account for almost half (45%) of the total number of events.
- Transform the logs (separately for both the urgent and non-urgent cases log) based on the notion of artifacts over the organizational perspective. Selecting this perspective includes the selection of time perspective for the definition of process instances. These process instances are reflected in the sub-logs. The abstraction log for the non-urgent cases contains 136 cases[6] and 1561 events distributed over 21 activities (corresponding to the organization names). In addition 21 sub-logs are created, one for each abstract activity (organization name). The abstraction log for the urgent cases contains 25 cases and 1118 events referring to 18 distinct activities. In addition 18 sub-logs are created, one for each abstract activity (organization name).

[6] This log contains one case less than the non-urgent cases log because one case in the non-urgent cases log has certain events without the organizational group attribute/value. We ignored this exceptional case.

Figure 13 depicts the workflow of the abstraction log pertaining to the non-urgent cases mined using the enhanced Fuzzy Miner[7]. This corresponds to the patient flow across the various labs/departments. All the nodes in the workflow are colored in blue (signifying that they are abstract nodes). Abstract nodes can be seamlessly zoomed in to see the sub-processes underneath it. Figure 14(a) depicts the sub-process pertaining to the activities performed on the patients in the pathology department. The pathology department performs histopathological studies on the tissues of the patients. Some of the primary activities include identifying the compartment for inspection, resection of tissues (small and big) and performing biopsies on them. Figure 14(b) depicts the sub-process pertaining to the activities performed on the patients in the radiology department. The radiology department takes image scans of different regions. Different modalities of scanning are supported e.g., MRI, CT, ultrasound etc. The scans are primarily performed on the pelvis, thorax, abdomen and allied bones. A CT chest scan is also normally performed. There is one exceptional case for whom additional dental, brain, knee and leg veins scans were performed (the highlighted region in the figure) at different instances in time.

Figure 15 depicts the sub-process pertaining to the activities of the 'General Lab Clinical Chemistry'. This lab is concerned with various diagnosis tests on the blood and urine. A few classes of tests are highlighted in the figure. For example, tests that assess the levels of creatinine, sodium, phosphate, bicarbonate etc., are performed. As another example, tests are performed to assess the blood group, estimate Rh factor, the white blood cell counts (leukocyte), red blood cell counts (haematocrit etc.), platelet counts etc. Sediment and urine analysis tests are also performed on patients. Patients diagnosed with M11 (vulvar cancer) are subjected to the CEA (carcinoembryonic antigen) tumor marker test followed by the cancer antigen tests, CA 125 and CA 19.9. Figure 16 depicts the region corresponding to the blood count tests in Figure 15.

In a similar fashion, the procedures followed in other departments/labs can be analyzed.

Figure 17 depicts the workflow of the abstraction log pertaining to the urgent cases mined using the enhanced Fuzzy Miner. Figure 18(a) depicts the sub-process pertaining to the activities performed on the patients in the pathology department while Figure 18(b) depicts the sub-process pertaining to the activities performed on the patients in the radiology department. The activities performed are more or less the same as that of the non-urgent cases except that for some urgent cases, 'aspiration cytology' and 'cytology vulva' are performed in the pathology department. The CT scan of the retroperitoneal region was performed on some of the urgent cases while none of the non-urgent cases re-

[7] Legible versions of all the figures in this paper can be inspected at: http://www.win.tue.nl/~jcbose/bpichallenge2011. The username is jcbose and the password is bpichallenge2011

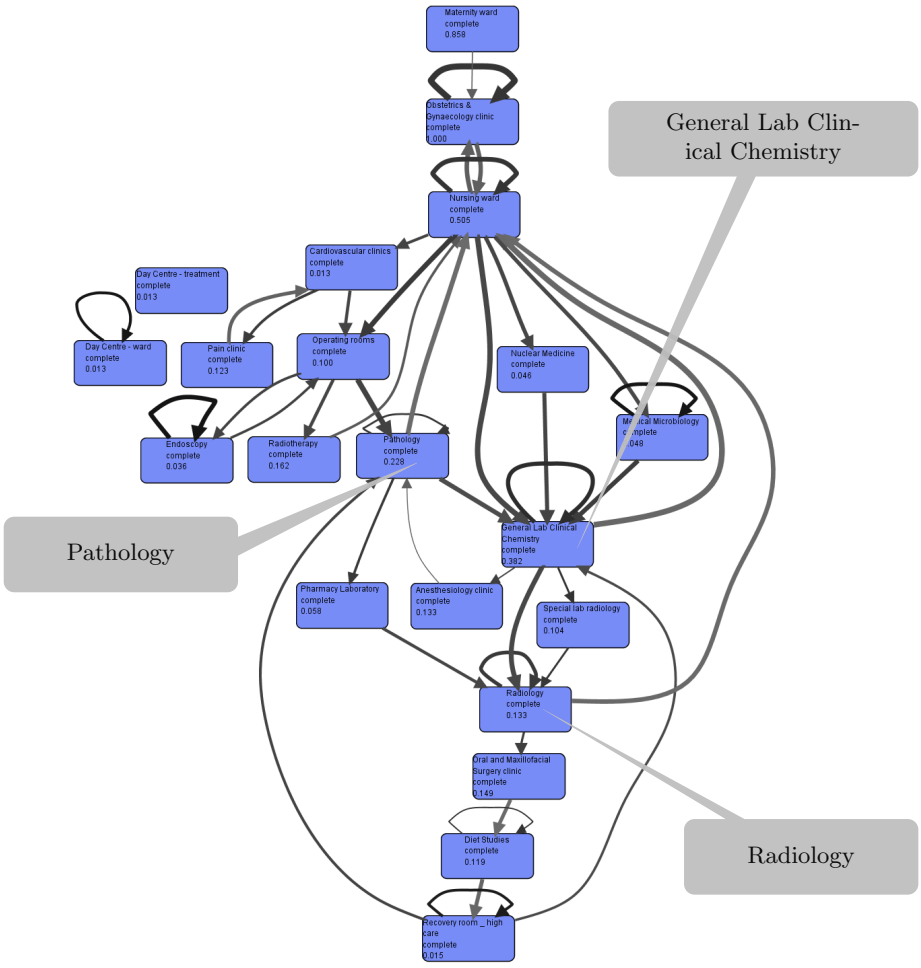
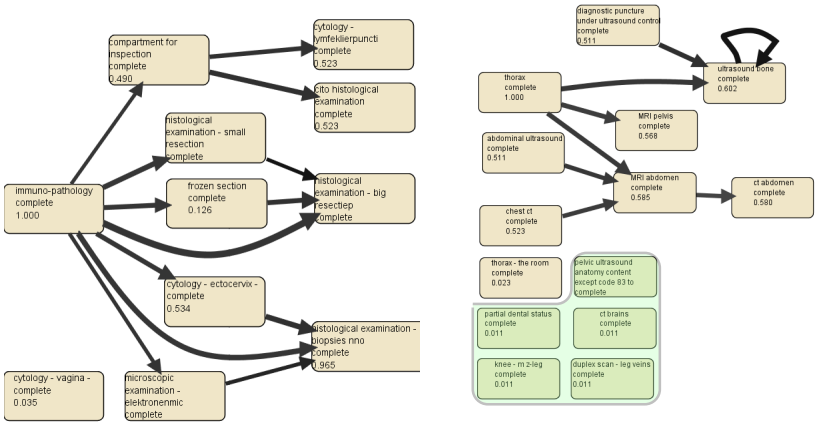


Fig. 13. The process model depicting the flow of patients across the different labs/departments in the non-urgent cases diagnosed with M11. The inside of the three highlighted nodes are detailed in Figure 14(a) (Pathology), Figure 14(b) (Radiology), and Figure 15 (General Lab Clinical Chemistry).

quired this. While only an MRI scan of the pelvis region was performed on the non-urgent cases, some urgent cases had a CT scan of the pelvis in addition to the MRI scans.

The primary difference between the treatment/diagnosis procedures followed between the non-urgent and urgent cases emanate in the general lab clinical chemistry sub-process. As mentioned earlier, 'urgent' variants of the activities



(a) Pathology (b) Radiology

Fig. 14. The sub-processes pertaining to the activities performed on the patients in the pathology and radiology departments for the non-urgent cases diagnosed with M11.

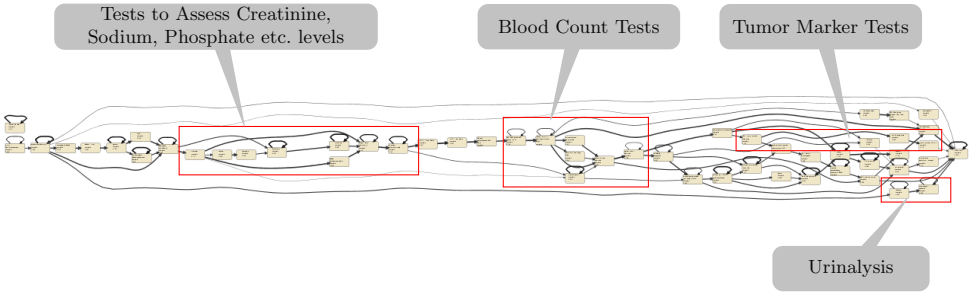


Fig. 15. The sub-process pertaining to the activities performed on the patients in the general lab clinical chemistry for the non-urgent cases diagnosed with M11.

(lab tests) are followed in the urgent cases. Figure 19 depicts the sub-process of the general lab clinical chemistry. The highlighted regions in the figure indicate the series of tests involving the urgent variants of the activities. The bottom half of the process is almost like that of the non-urgent cases. It is important to note that during the lifetime of a case, the lab tests can be performed multiple times. Some of these tests were conducted in an expedited manner whereas for other tests the normal flow was followed. That is the reason why we see both the normal and urgent variants of the activities.

We have also analyzed the workflow of cases diagnosed with other codes and code combinations. In all the scenarios, we are able to mine meaningful process models. At first sight though the log seems to be complex, using a systematic

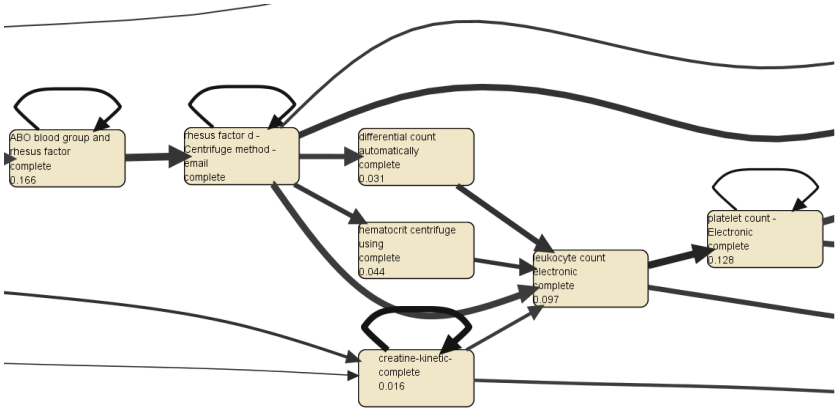


Fig. 16. A portion of the general lab clinical chemistry sub-process pertaining to the estimation of blood counts for non-urgent cases diagnosed with M11.

approach as discussed in this section, we have discovered that the models are rather simple and sequential.

4.2 Process Diagnostics Using Trace Alignment

We now analyze the log using trace alignment. Trace alignment can be used to explore the process in the early stages of analysis and to answer specific questions in later stages of analysis e.g., are there common patterns of execution, are there any anomalies, are there any distinguishing aspects with respect to the treatment procedures followed among cases, etc. As mentioned earlier, the 'Trace Alignment With Guide Tree' plug-in in ProM implements this functionality.

We will discuss the application of trace alignment and infer insights by using the cases diagnosed for cervical cancer of the uteri (diagnosis code M13). The raw event log is subjected to the following pre-processing steps:

- Select the cases whose diagnosis code combination is just {M13}. There are a total of 252 cases in the event log satisfying this criterion. This filtered event log contains 14611 events distributed over 272 activities.
- From the filtered log obtained in the previous step, select the cases who have been subjected to a treatment code combination of {803}. There are 23 cases satisfying these criteria. This filtered event log contains 3329 events referring to 135 distinct activities. Though only 9% of the cases diagnosed with {M13} are treated with the treatment code {803}, nearly 23% of the events happen in these 9% of the cases.
- Segregate urgent and non-urgent cases from the filtered log obtained in the previous step. Of the 23 cases, 15 cases are non-urgent cases containing 1961 events distributed over 110 activities while 8 cases are urgent cases

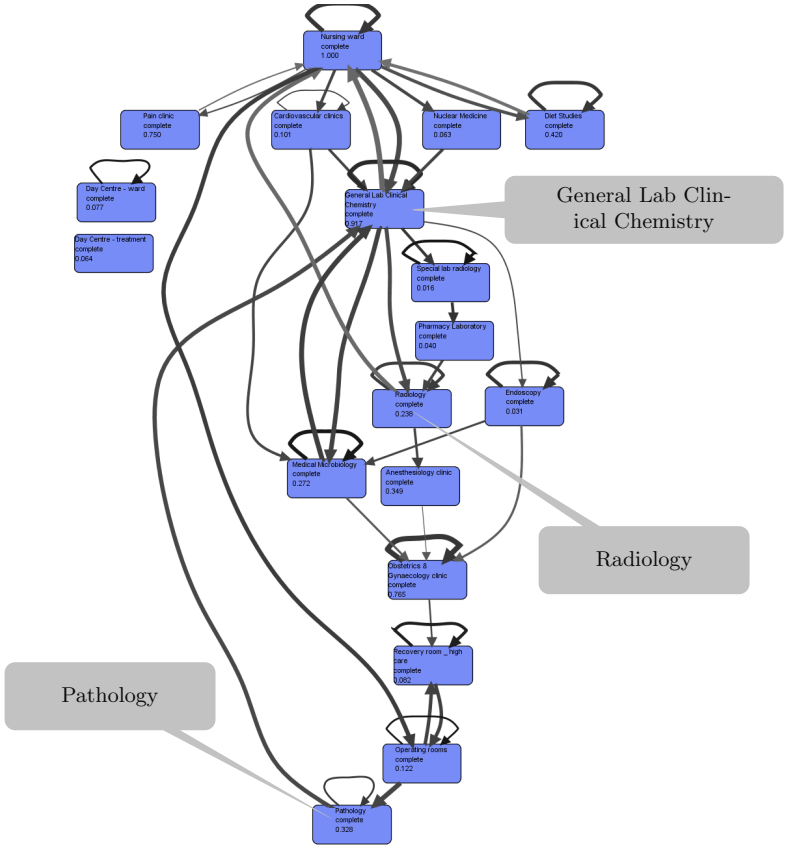
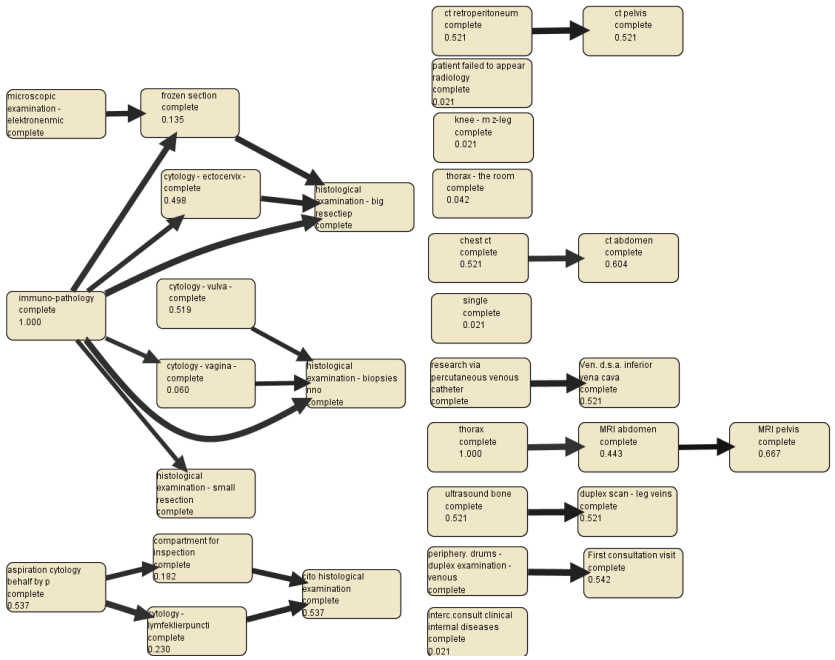


Fig. 17. The process model depicting the flow of patients across the different labs/departments in the urgent cases diagnosed with M11. The inside of the three highlighted nodes are detailed in Figure 18(a) (Pathology), Figure 18(b) (Radiology), and Figure 19 (General Lab Clinical Chemistry).

containing 1368 events distributed over 94 activities. Let us call these two logs the *non-urgent cases log* and *urgent cases log*.

Figure 21 depicts the initial portion[8] of the aligned traces pertaining to the cases in the non-urgent event log. The event log is first encoded into traces where each trace is the sequence of activities corresponding to a case. The activities are represented in an encoded form in a trace with each activity encoded using two characters (e.g., `h4`, `i1`, `a2` etc.). We can clearly see some common patterns of

[8] The entire alignment is of length 375 and is not shown due to legibility issues. However, the entire alignment can be inspected at: http://www.win.tue.nl/~jcbose/bpichallenge2011. The username is jcbose and the password is bpichallenge2011



(a) Pathology (b) Radiology

Fig. 18. The sub-processes pertaining to the activities performed on the patients in the pathology and radiology departments for the urgent cases diagnosed with M11.

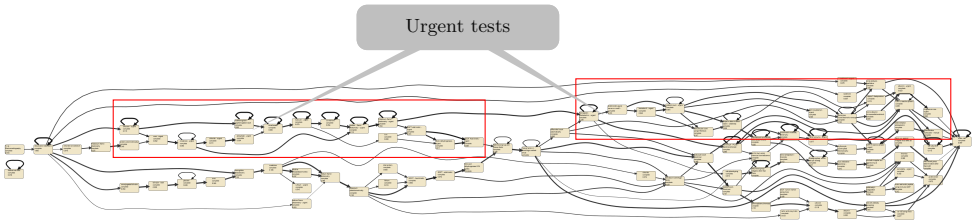


Fig. 19. The sub-process pertaining to the activities performed on the patients in the general lab clinical chemistry for the urgent cases diagnosed with M11.

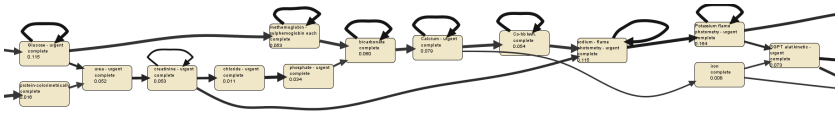


Fig. 20. A portion of the general lab clinical chemistry sub-process involving the urgent activities pertaining to the estimation of creatinine, sodium, phosphate etc. levels.

execution and exceptional/rare behavior in the alignment. For example, from the alignment, we can see that:

- some non-urgent cases start with `h4` (e.c.g. - Electrocardiography) while one case starts with `f4` (ultrasound - internal genitals). All the other cases start with `i1` (to particular laboratory).
- all cases have three instances of `i1` with the exception of one (trace 00001023), which has only two instances.
- all the cases with the exception of one (trace 00001023) have the activity sequence `e4c3` corresponding to the estimation of ABO blood group and Rh factor (`e4`) and Rh factor using centrifuge method (`c3`) respectively.
- only one of the cases (trace 000000928) required the execution of the activity `e8` corresponding to cephalin time-coagulation test.
- only in three cases was the activity `a0` (CEA - tumor marker using meia) performed before `e7` (squamous cell carcinoma using eia)

Such rare behavior/exceptions/deviations can be acceptable or can indicate a violation of normative behavior. For example, skipping the tests corresponding to the estimation of the ABO blood group and Rh factor is more likely to be a violation than an acceptable behavior. On the contrary, the rare execution of the cephalin time-coagulation test can be acceptable; based on the history of the patient, this test could have been recommended.

Figure 22 depicts the initial portion of the aligned traces pertaining to the cases in the urgent event log. Common patterns of execution and deviations can be clearly seen. For example, from the alignment, we can see that:

- unlike the non-urgent cases, some of the urgent cases start with `g8`, corresponding to 'nursing gynecology short-out card cost'.
- unlike the other urgent cases, the second trace (00000257) skips a lot of activities.
- two cases (traces 00000257 and 00000058) do not contain the activity `a0` corresponding to CEA - tumor marker using meia.
- for two traces (00000257 and 00000499), the activity sequence `f0c2` corresponding to 'first outpatient consultation' and 'administrative fee' happens quite early in the trace when compared to the rest. It could be that these two activities happen in parallel with the rest of the process.

Figure 23 depicts a portion of the alignment over the urgent cases where there are tandem repetitions of activities pertaining to the estimation of levels of sodium, bicarbonate, calcium etc. The activities involved are `i3` (Glucose - urgent), `d1` (methemoglobin - sulphemoglobin each), `b8` (bicarbonate), `a5` (Calcium - urgent), `b2` (Co-hb kwn), `h4` (sodium - flame photometry - urgent), `e7` (Potassium flame photometry - urgent), `g0` (hemoglobin photoelectric - urgent) and `b6` (Current ph - PCO2 - stand.bicarbonaat). It is also interesting to note that the third trace has one additional instance of execution of these activities and the last trace has two instances of execution less when compared to the other traces.

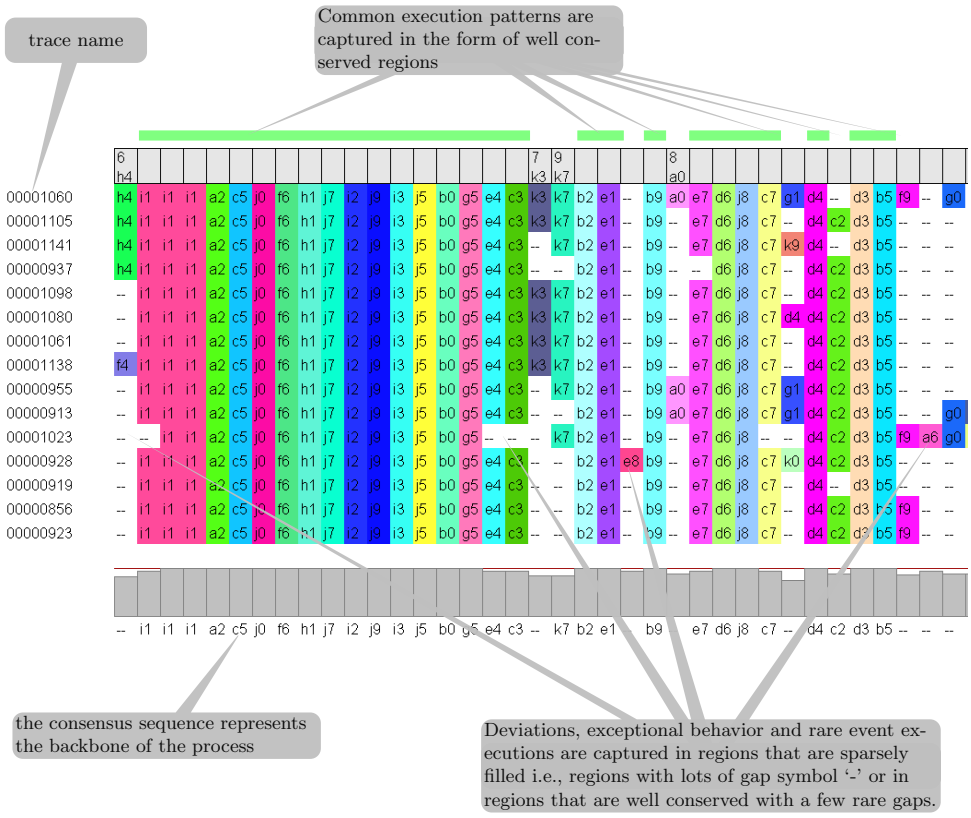


Fig. 21. Initial portion of the aligned traces pertaining to the non-urgent cases diagnosed with M13 and whose treatment code is 803. Each row refers to a process instance (a patient case). Columns describe positions in traces. Consider now the cell in row y and column x. If the cell contains an activity name `a`, then `a` occurred for case y at position x. If the cell contains no activity name (i.e., a gap "-"), then nothing happened for y at position x. Trace alignment aims a minimizing the number of gaps and maximizing the consensus.

In this fashion, trace alignment can assist in uncovering extremely interesting insights and act as probes when analyzing process execution behavior thereby giving cues on process improvement opportunities.

5 Conclusions

In this paper, we proposed a systematic approach for the analysis of the hospital event log provided for the BPI challenge. We focussed on two perspectives of pro-

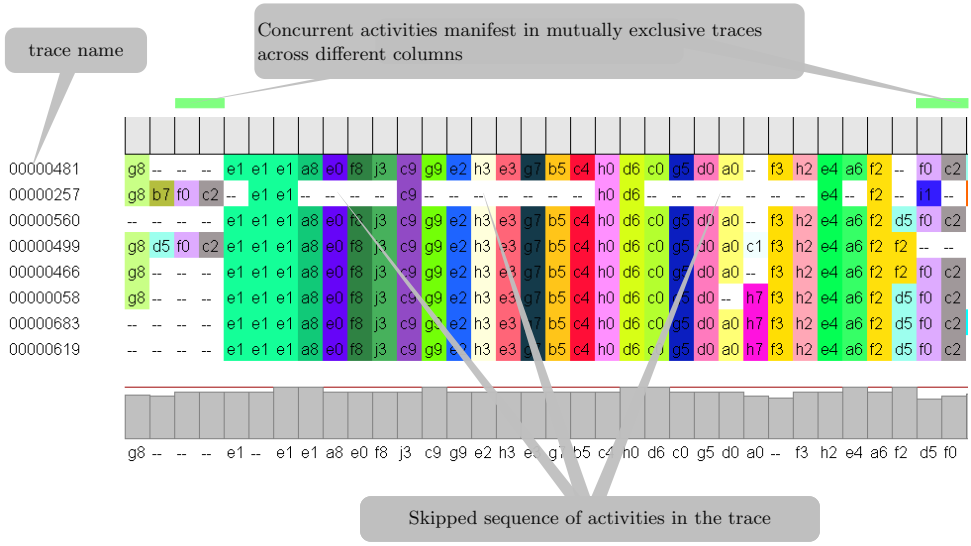


Fig. 22. Initial portion of the aligned traces pertaining to the urgent cases diagnosed with M13 and whose treatment code is 803.

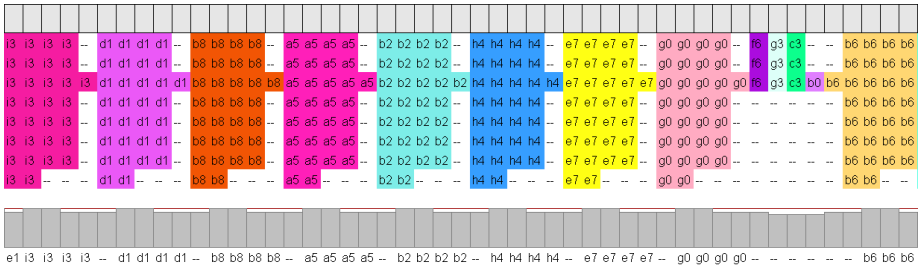


Fig. 23. Portion of the alignment where there are tandem repetitions of certain activities in the urgent cases diagnosed with M13 and whose treatment code is 803.

cess mining for our analysis viz., control-flow and diagnostics. We have reported the results (and observations) on these two perspectives for cases diagnosed with vulvar cancer (diagnosis code M11) and cervical cancer of the uteri (diagnosis code M13). Though the event log seems to be complex, in reality it is not. Adopting the systematic approach presented in this paper, we realized/showed that the processes are in fact rather simple and often sequential. Furthermore, based on trace alignment, we noticed that not only are the processes simple and sequential but also the cases share a lot in common with very little deviations from the main path.

In this paper, we presented just a small subset of our findings. One of the complications was that we were unable to discuss our findings with stakeholders and

domain experts. Process mining is typically an iterative activity driven by questions from stakeholders and surprising analysis results. Nevertheless, we were able to gain significant insights. Through enhanced fuzzy mining and trace alignment, we were able to understand the processes at hand. Moreover, we strongly believe that more accurate insights and interpretations can be obtained/inferred by performing the analysis in collaboration with the domain experts.

Acknowledgments The authors are grateful to Philips Healthcare for funding the research in process mining.

References

1. van der Aalst, W.M.P.: Process Mining: Discovery, Conformance and Enhancement of Business Processes. Springer-Verlag (2011)
2. van der Aalst, W.M.P., Weijters, A.J.M.M., Maruster, L.: Workflow Mining: Discovering Process Models from Event Logs. IEEE Transactions on Knowledge and Data Engineering **16**(9) (2004) 1128–1142
3. Rozinat, A., van der Aalst, W.M.P.: Conformance Checking of Processes Based on Monitoring Real Behavior. Information Systems **33**(1) (2008) 64–95
4. Li, J., Bose, R.P.J.C., van der Aalst, W.M.P.: Mining Context-Dependent and Interactive Business Process Maps using Execution Patterns. In zur Muehlen, M., Su, J., eds.: BPM 2010 Workshops. Volume 66 of LNBIP., Springer-Verlag (2011) 109–121
5. Bose, R.P.J.C., van der Aalst, W.M.P.: Trace Alignment in Process Mining: Opportunities for Process Diagnostics. In Hull, R., Mendling, J., Tai, S., eds.: Proceedings of the 8th International Conference on Business Process Management (BPM 2010). Volume 6336 of LNCS., Springer-Verlag (2010) 227–242
6. Günther, C., van der Aalst, W.M.P.: Fuzzy Mining: Adaptive Process Simplification Based on Multi-perspective Metrics. In: International Conference on Business Process Management (BPM 2007). Volume 4714 of LNCS., Springer-Verlag (2007) 328–343
7. Görg, C., Pohl, M., Qeli, E., Xu, K.: Visual Representations. In Kerren, A., Ebert, A., Meye, J., eds.: Human-Centered Visualization Environments. Volume 4417 of LNCS. Springer (2007) 163–230
8. Bose, R.P.J.C., van der Aalst, W.M.P.: Context Aware Trace Clustering: Towards Improving Process Mining Results. In: Proceedings of the SIAM International Conference on Data Mining (SDM). (2009) 401–412
9. Nigam, A., Caswell, N.: Business Artifacts: An Approach to Operational Specification. IBM Systems Journal **42**(3) (2003) 428–445
10. Bose, R.P.J.C., van der Aalst, W.M.P.: Trace Clustering Based on Conserved Patterns: Towards Achieving Better Process Models. In Rinderle-Ma, S., Sadiq, S., Leymann, F., eds.: BPM 2009 International Workshops, Ulm, Germany, September 7, 2009. Revised Papers. Volume 43 of LNBIP. (2009) 170–181
11. Bose, R.P.J.C., van der Aalst, W.M.P.: Abstractions in Process Mining: A Taxonomy of Patterns. In Dayal, U., Eder, J., Koehler, J., Reijers, H., eds.: Business Process Management. Volume 5701 of LNCS., Springer-Verlag (2009) 159–175
12. Bose, R.P.J.C., Verbeek, E.H.M.W., van der Aalst, W.M.P.: Discovering Hierarchical Process Models Using ProM. In Nurcan, S., ed.: Proceedings of the CAiSE Forum. Volume 734., CEUR-WS.org (2011) 33–40

A Appendix

Table 2. Distribution of cases based on diagnosis code combination

Diagnosis Code Combination	Number of Cases
{M13}	252
{M16}	201
{M11}	162
{M14}	106
{106}	70
{822, 106}	61
{M13, 106}	57
{M13, 822, 106}	47
{106, M14}	31
{M16, 821}	25
{M12}	17
{821}	16
{M15}	14
{822}	11
{M11, 106}	11
{M13, 822}	10
{839}	8
{823}	7
{839, M16}	6
{M12, 106}	4
{823, 106, M14}	4
{823, M14}	3
{106, 821}	3
{M11, 106, 839}	2
{106, M15}	2
{106, 839}	1
{M13, 106, 839}	1
{M12, 839}	1
{822, 106, M14}	1
{M13, 106, 821}	1
{839, M16, 821}	1
{823, 106}	1
{106, M16, 821}	1
{823, 106, M15}	1
{106, M16}	1
{822, 106, 821}	1
{839, 821}	1
{M11, 822}	1