Process Mining in Healthcare

Opportunities Beyond the Ordinary

R.S. Mans[a,*], W.M.P. van der Aalst[a], Rob J.B. Vanwersch[b]

[a] *Department of Mathematics and Computer Science, Eindhoven University of Technology, P.O. Box 513, NL-5600 MB, Eindhoven, The Netherlands*
[b] *Maastricht University Medical Center, P.O. Box 5800, NL-6202 AZ, Maastricht, The Netherlands*

Abstract

Nowadays, in a Hospital Information System (HIS) huge amounts of data are stored about the care processes as they unfold. This data can be used for process mining. This way we can analyse the operational processes within a hospital based on facts rather than fiction. In order to enhance the uptake of process mining within the healthcare domain we present a *healthcare reference model* which exhaustively lists the typical types of data that exists within a HIS and that can be used for process mining. Based on this reference model, we elaborate on the most interesting kinds of process mining analyses that can be performed in order to illustrate the potential of process mining. As such, a basis is provided for governing and improving the processes within a hospital.

Keywords: healthcare, process mining, reference model

*Corresponding Author. Address: Eindhoven University of Technology, Department of Mathematics and Computer Science, Architecture of Information Systems (AIS), MF 7.062, Den Dolech 2, P.O. Box 513, 5600 MB Eindhoven, The Netherlands, Tel:+31-40-247-3686, Email address: r.s.mans@tue.nl.

Email addresses: r.s.mans@tue.nl (R.S. Mans), w.m.p.v.d.aalst@tue.nl (W.M.P. van der Aalst), rob.vanwersch@mumc.nl (Rob J.B. Vanwersch)

1. Introduction

People working in healthcare are in need of actionable information [1]. For example, physicians are checking medical files for patients, nurses use care systems to see the subsequent steps that need to be taken in order to care for patients, and radiologists check which examinations need to be done. In addition, the management of a hospital needs real-time information about the hospital's costs and services. In order to provide reliable and up-to-date information to all stakeholders and to achieve high-quality and efficient patient care, the aforemention types of information all need to be stored in the information systems of a hospital. The collective name of such a system is often referred to as the Hospital Information System (HIS).

As a result, a typical HIS contains a wealth of information. This also means that a lot of knowledge about the care processes that are performed and the steps within these care processes is available. The recorded process steps can be used as input for process mining, i.e. it is possible to obtain insights into how the processes are *really* executed. As a next step, processes can be analyzed and optimized.

Not surprisingly, there is an uptake of process mining in the healthcare domain. Up to now, we have discovered 35 publications in which a real-life application of process mining in healthcare is described (see Section 6 for an overview). For these applications often only data is taken from one or two systems in order to solve a particular problem. These applications seem not to exploit the full potential of process mining due to the fact that an overview is missing of all the data that exists within a HIS and that can be used for process mining. Furthermore, when analyzing literature on HISs and actual HIS implementations in hospitals (see also Section 6), this knowledge could not be obtained.

Based on the discussion above, an interesting and challenging question arises: *Which data exists within a HIS that can be used for process mining?* In order to answer this question, we developed a *healthcare reference model* outlining all the different classes of data that are potentially available for process mining. We also discuss the relationships between these classes. Subsequently, given this reference model it is possible to reason about application opportunities for process mining, e.g., we will discuss several kinds of analyses that can be performed. This enables us to answer the following question: *What are the possibilities of process mining within hospitals?* When reasoning about these possibilities, we initially assume that no data quality issues exist. However, as in reality this is never the case we also elaborate on data quality issues that may apply for the data part of the *healthcare reference model*. We provide a classification of data quality issues and evaluate and illustrate the classification using a *large dataset consisting of more than 60 tables that has been obtained from the Maastricht University Medical Center* (MUMC), a large academic hospital in the Netherlands[1].

The remainder of this paper is structured as follows. Section 2 introduces process mining. In Section 3, the *healthcare reference model* is introduced. In Section 4, based on the reference model, the possibilities of process mining

[1] http://www.mumc.nl/

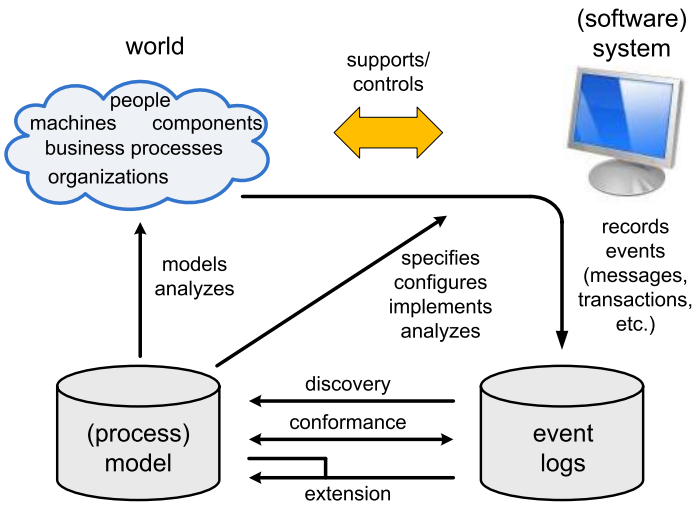


Figure 1: Three types of process mining: (1) Discovery, (2) Conformance, and (3) Extension.

within all disciplines of a typical hospital will be illustrated. In Section 5, data quality issues that may apply for data described in the reference model will be identified. Related work is listed in Section 6. In Section 7, we discuss our experiences associated to defining the reference model and identifying the opportunities for process mining. Section 8 concludes this paper.

2. Process Mining

Process mining is applicable to the event data of a wide range of systems. Examples of these systems are Business Process Management (BPM) systems (e.g. *BPM|One*, *Filenet*), ERP systems (e.g. *Microsoft Dynamics NAV*), Product Data Management (PDM) systems (e.g. *Windchill*), and Hospital Information Systems (e.g. *i.s.h.med*, *Chipsoft*, *iSOFT*, *McKesson*, and *Epic*). For the application of process mining, the only requirement is that the system produces *event logs*, thus recording (parts of) the actual behavior. For these event logs it is important that each event refers to a well-defined step in the process (e.g., a radiology examination) and is related to a particular case (e.g., a patient). Also, additional information such as the performer of the event (i.e. the physician performing the examination), the timestamp of the event, or data elements recorded along with the event (e.g. the weight of the patient) may be stored. Based on these event logs, the goal of process mining is to extract process knowledge (e.g. process models) in order to discover, monitor, and improve real processes [2]. As shown in Figure 1, three types of process mining can be distinguished [2].

- **Discovery:** inferring process models that are able to reproduce the observed behavior. The inferred model may be a Petri net [3], a BPMN model [4], a Declare model [5], an EPC [6], or any other type of process model. For example, the discovered model may describe the typical steps taken in order to diagnose a patient. Note that also models describing the organizational, performance, and data perspective may be discovered.

- **Conformance:** checking if observed behavior in the event log conforms to a given model [7, 8]. For example, it may be checked whether a medical guideline which states that always a CT-scan of the abdomen and a chest X-ray need to be performed is always followed.
- **Extension:** projection of the information extracted from the log onto the model. For example, performance information may be projected on a discovered healthcare process in order to see for which examinations a long waiting time exists [8].

The ProM framework and tool set has become the de facto standard for process mining [9]. ProM[2] is a "plug-able" environment for process mining using MXML, SA-MXML, or XES as input format.

3. Healthcare Reference Model

In this section, we discuss the *healthcare reference model* outlining all the different classes of data that can be available for process mining together with the associated relationships between these classes of data. In particular, we focus on the classes related to the *primary care* process (care which is directly related to or provided for patients). As such, classes related to supportive processes are not included in the reference model (e.g. the billing of services or the purchase of materials). The reference model is described in terms of a UML class diagram [10].

In order to build the *healthcare reference model* the following five-step approach has been used. First, we selected a HIS for which we could investigate an actual running implementation of the system within a hospital. More precisely, in the context of an existing collaboration between MUMC and TU/e we have investigated the *i.s.h.med* system of Siemens Healthcare that is in use at the hospital. In this system, detailed information was available concerning the tables that are present and the content of these tables, i.e. the names of the columns in the tables. Also, it was possible to learn about the usage of the system in practice by interviewing people and by inspecting the system itself. Moreover, data from practice could be obtained by extracting data from multiple tables (see Section 4). Secondly, based on the information obtained in the first step, the *healthcare reference model* has been developed. In the third step, we have investigated a database which came from an implementation of a HIS in another hospital and which is provided by another software supplier. This involved an database which came from the Chipsoft EZIS system[3] which is in use at the Catharina hospital in Eindhoven[4], the Netherlands. The database contained all data for patients treated in 2005 and also contained detailed information about the tables and their content. In the fourth step, based on the previous investigation, the reference model has been revised. Finally, the reference model was discussed with HIS professionals from multiple hospitals in order to identify missing data entities and to further revise the model.

[2] www.processmining.org
[3] http://www.chipsoft.com
[4] http://www.cze.nl

Taking the above mentioned approach into account, we believe that the developed *healthcare reference model* is representative for HISs.

In total, the reference model consists of 125 classes. We use the following grouping:

- **General Patient and Case Data:** General information about patients and the cases that are executed for them.
- **Processes and Process Steps:** Information about all steps that are performed for patients and the constitution of multiple steps into a (sub)process.
  - **Medication:** Medications that are given to patients.
  - **Patient Transport:** The transportation of patients within the hospital.
  - **Radiology:** Radiology examinations that are performed for patients.
- **Document Data:** The medical data that is saved in the context of steps that are performed for patients.
- **Organization and Buildings:** The organizational and building related structure within the hospital.
- **Nursing Plans:** Plans for the care that is given by nurses to patients.
- **Pathways:** The definition and execution of pathways.

Note that the first two groups ("general patient and case data" and "processes and process steps") relate to general information about patients and the process steps that have been performed for them. Subsequently, in the "patient transport", "medication", and "radiology" groups, we illustrate some more fine grained process information that can be found for the services related to patient transport, radiology, and medication. The "document data" and "organization and buildings" groups define additional data that can be found for the performed services and patients in general, i.e. medical documentation and resources that are involved. Finally, in the "nursing plans" and "pathways" groups we elaborate on process knowledge that is stored about nursing plans and pathways.

Below, each group is discussed in more detail. This is done by focussing on the classes that have been modeled for the group and the relationships between these classes. Since the time aspect is important for the application of process mining, we elaborate on timing related information. Moreover, several attributes are provided in order to illustrate the data in a class. However, we particularly focus on these attributes that are important for process mining, i.e. names of steps, timestamps, resources, and case identifiers. The attributes that together uniquely identify each record in a class, i.e. the primary key, are indicated by the "+" character in front of them. Similarly, an attribute which is not a private key is illustrated by a "-" character.

3.1. General Patient and Case Data

For a patient there are several kinds of general information that are important for the entire treatment trajectory in the hospital. The associated classes are shown in Figure 2. For a patient (class `patient`) general information exists such as `patient number`, `name`, `sex`, `birthdate`, `religion`, and `telephone`

number. Note that the patient number is a primary key in the patient class as indicated by the "+" character. Additionally, a patient may suffer from multiple health problems (class health problems for which the associated multiplicity is 0..*) and multiple risk factors may exist for a patient (class risk factors for which the associated multiplicity is 0..*). Also, if a patient is a Very Important Person (VIP) then all transactions that take place may need to be logged (class logging for VIP patients with associated multiplicity 0..*).

A patient may suffer from multiple illnesses. For each illness a separate case is created (class case for which the associated multiplicity is 0..*). Note that for a case it is saved on which day the case started and on which day it is completed. Also for a case, several kinds of general information can be saved. That is, information is saved about complications that have been observed (class complications), diagnoses that have been identified (class diagnoses), the assignment of persons to a case (class assignment of a case to a person), and the classification categories which belong to a case (class case classification). Finally, during treatment it may be desired to assign a case to another case (class assignment of a case to another case). This can be for example caused by a complication that occurred and that subsequently a different treatment is needed for the patient.

For several classes, different timing information is recorded. Regarding the complications and diagnoses, the exact date and time is recorded on which respectively the complication occurred and the diagnosis was made. Furthermore, for a classification category of a case and an assignment of a case to a person it is specifically saved for which period it is valid, i.e., for each record the start and end date of its validity is recorded.

3.2. Processes and Process Steps

For a case, several steps can be performed. In Figure 3 until Figure 5 it is visualized which information is stored about the process steps that have been performed. For these process steps we look at their timing and at which level of granularity they are saved. In total 26 different classes are defined. Three groups of classes can be distinguished: (1) the referral of the patient to the hospital; (2) the different kind of process steps that are recorded; and (3) the registration of orders and appointments.

3.2.1. Referral

In Figure 3 the classes regarding the referral of patients are shown. For a referred patient it is recorded on which day the patient is referred (class referrals). Furthermore, additional information about the referral is recorded such as the referring hospital and the kind of referral (e.g. via the general practitioner, emergency, or self-referral). Also, a referral number is created which is important in order to link the patient to one or more cases (class reference referral data with associated multiplicity 1..*).

3.2.2. Different Kind of Process Steps

Figure 4 shows all classes which are related to the recording of process steps during the diagnosis and treatment of patients. At the coarsest level of granularity there are the cases of illness from which a patient may suffer (class cases). A case may consist of multiple so-called movements (class movements

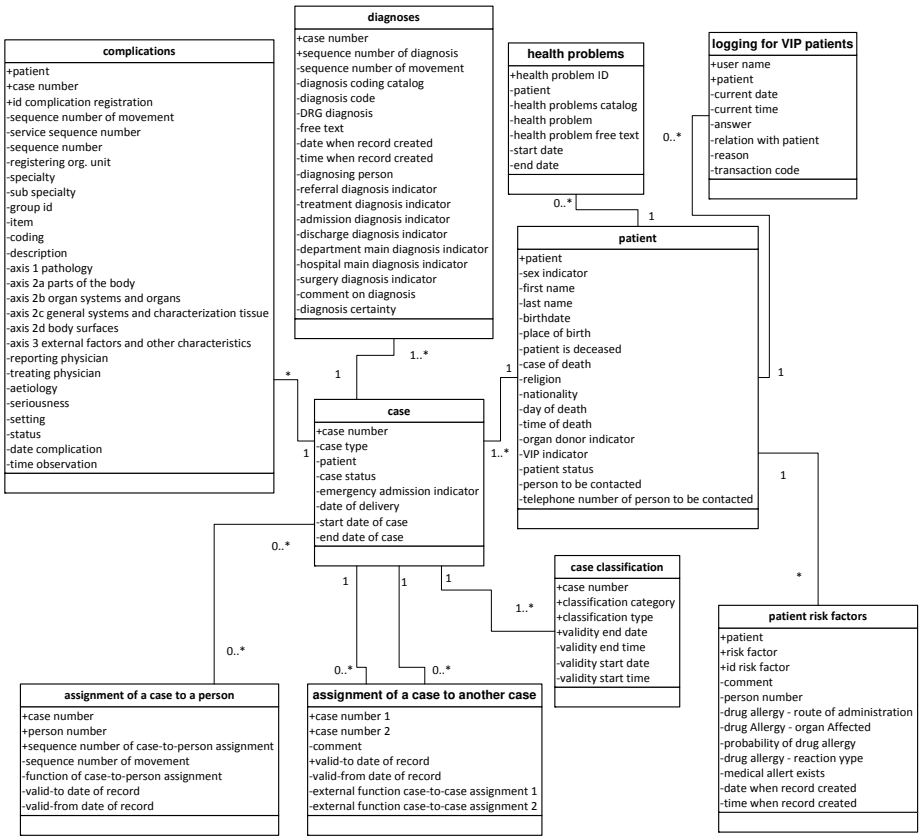


Figure 2: Classes describing general information for patients and cases. A patient (`patient` class) may suffer from multiple health problems (class `health problems`) and multiple risk factors may apply (class `patient risk factors`). For a patient multiple (ongoing) cases may be recorded (`case` class). Furthermore, for a case multiple kinds of information may be recorded such as the complications that occurred (`complications` class) or the diagnoses (`diagnoses`) that have been made.

`for case` with associated multiplicity `1..*`). A movement can be seen as an encounter between a patient and a healthcare provider which spans a period of time. Furthermore, the length and detail of a movement may vary according to local procedures, conventions, and data capturing standards. A movement can be seen as a step which is defined at a coarse level of granularity. Examples are a stay at the nursing ward, a visit to the outpatient clinic, or a surgical intervention.

As a part of a movement, multiple services may be delivered to patient (class `services performed` with associated multiplicity `0..*`). Similar to a movement, a service spans a period of time but its duration is typically shorter than for a movement. Also, it's level of detail is more fine-grained and refers to a concrete piece-of-work that is performed by a person. For example, during a visit to an outpatient clinic (which is a movement) a physician may perform multiple services such as an echography and an intake. As can be seen in Figure 4, a service can be further specialized into three other kinds of services. Medical services (class `medical service`) relate to diagnostic and therapeutic

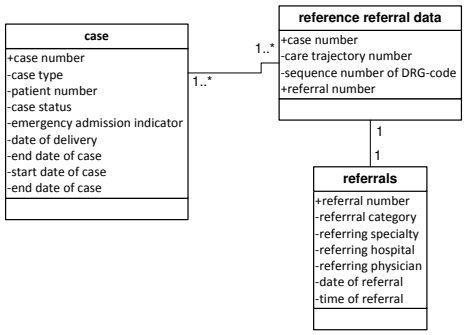


Figure 3: Classes describing the referral of patients.

services which are performed for patients (e.g. an X-ray or the removal of a tumor), allied health services (class `allied health service`) relate to services which are related to medicine (e.g. services performed by a physiotherapist or dietician), and non-medical services (class `non-medical service`) relate to the coordination of healthcare professionals and the support of a patient and his relatives (e.g. the planning of an appointment or informing the family about the status of a patient). For services, the involvement of each person is recorded including the period in which each person was involved (class `involved staff members`). Note that as a service refers to a concrete piece of a work, many additional attributes may exist for a specific service itself. This is illustrated by the attributes in the class `service catalog`. For example, for a service its type, costs, minimal duration, and by which discipline it may be performed may be given. Furthermore, for a service some specific context attributes may exist (class `context of service`) which may be different for medical discipline (classes `surgery`, `radiology`, and `cardiology`). For example, in the context of a surgery it is recorded which specific diagnoses are relevant (class `surgery diagnoses`) and which complications occurred (class `surgery complications`).

Finally, for a service or a movement even more fine-grained information may exist. This is illustrated by the `occurred events` class which refers to events that have occurred as part of a service or a movement. For example, regarding a surgery, event information is recorded for the request of transporting the patient to the surgical department, the arrival of the patient in the holding area, the arrival of the patient in the operating room, the start of the induction[5], the departure from the operation room, and the arrival in the recovery room.

Note that for each case, movement, and service information exists about the person who created, modified, and canceled a record together with its timing. Furthermore, for each movement and service it is recorded by which organizational unit it has been requested and for which organizational unit it has been performed.

[5]Induction is an anesthesiological term for the administration of a drug or combination of drugs at the beginning of an anesthetic that results in a state of general anesthesia

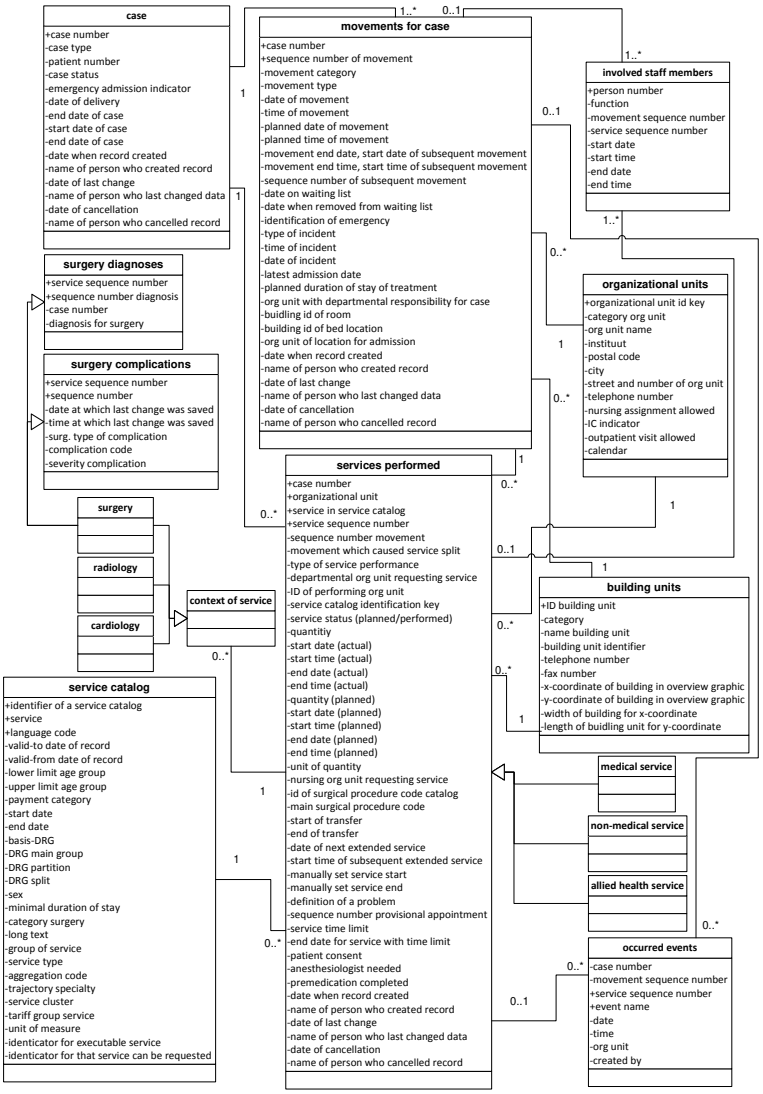


Figure 4: Classes describing the process steps that are performed and which may reside at different levels of granularity. That is, for a case (`case` class) multiple movements may be recorded (class `movements`). Similarly, for a movement one or more services may have been performed (`services performed`). As part of a service or movement multiple events may have occurred (`occurred events`). A service can be a medical service (class `medical service`), a non-medical service (`non-medical service` class), or an allied health service (`allied health service` class).

3.2.3. Orders and Appointments

Figure 5 relates to orders and appointments that are created for patients and the associated relationships. As already indicated earlier, a movement spans a period of time during which multiple services can be performed. As a result, an appointment can only be made for a movement (class `appointments` with associated multiplicity `1..1`). An appointment can be booked into the calendar of one or more resources (e.g. a physician (class `involved staff members`), a

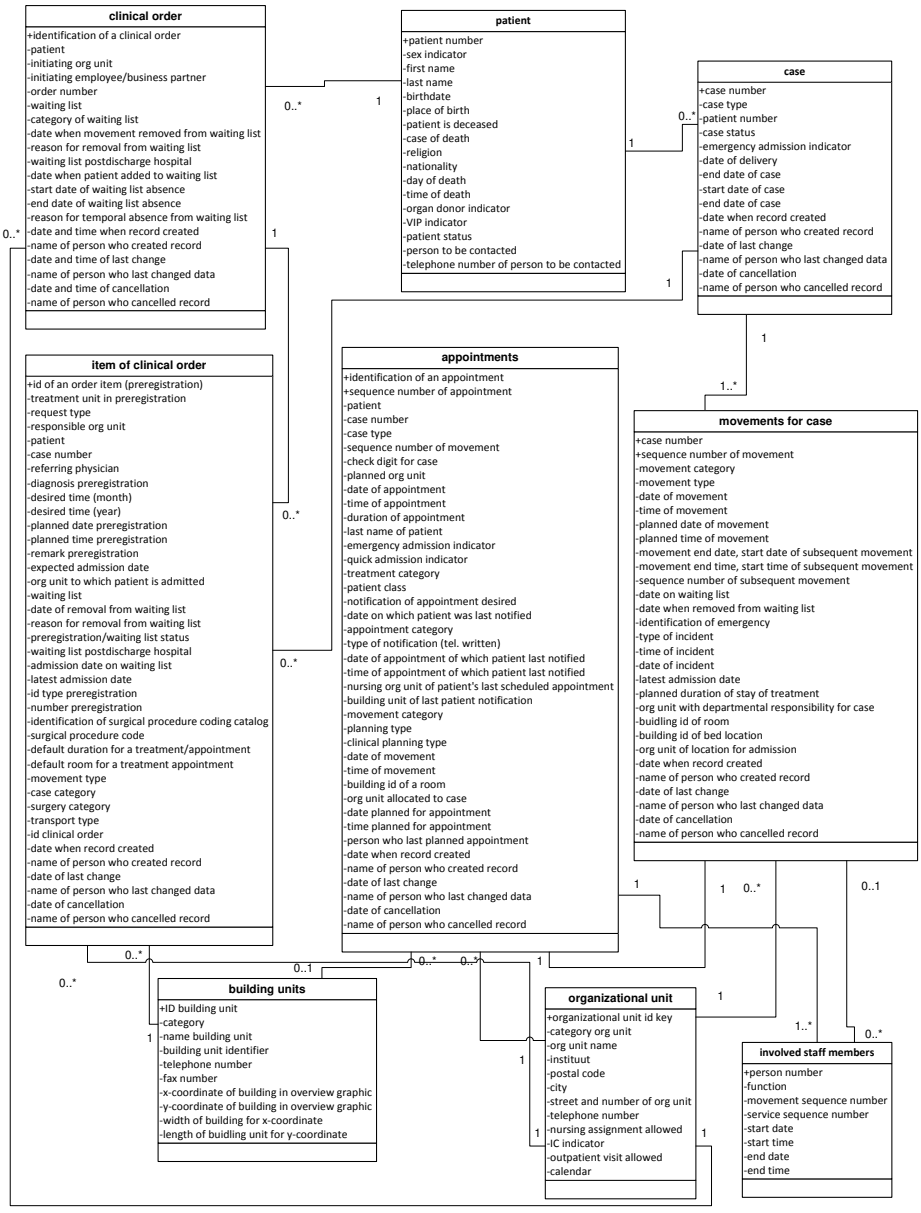


Figure 5: Classes related to orders and appointments that are made for patients. Only for a movement an appointment can be made (`appointments` class). An appointment may be booked as part of an item of a clinical order (`item of clinical order` class). Note that an order item is part of a clinical order (`clinical order` class).

room) and has a start and end date and time. Furthermore, information can be kept about the date when the patient was last notified and the date and time of the appointment at that time. Note that it is not required that for each movement an appointment exists.

The booking of one or more appointments for a patient can be triggered via a clinical order (class `clinical order` with associated multiplicity `0..*`).

This may involve putting the patient on a waiting list. As part of the latter, it is recorded at which time the patient is put on the waiting list. Also, timing information is recorded describing if and when a patient was temporarily absent from the waiting list. An order may consist of multiple order items (class `item of clinical order` with associated multiplicity `0..*`). For each order item an appointment may be created and linked to the corresponding movement (e.g. for a surgery it is needed to make an appointment for the preoperative assessment, the surgical interventional itself, and the admission to the nursing ward). As such, for an order item, in comparison to a clinical order, some aspects can be filled in which are important for the appointment to be made (e.g. the performing organizational unit, the referring physician, the default duration and room of the appointment, and the desired time of the appointment). Note that it is not required that for each appointment an order item exists (e.g. a lab order). Also, an order does not always need to refer to an appointment (e.g. an order may also be created for a medication that needs to be given). For appointments, clinical orders, and order items, information may be kept about the requesting and performing organizational unit and the involved building unit. Furthermore, also here, information exists about the person who created, modified, and canceled a record together with its timing.

3.3. Document Data

Related to the entire treatment trajectory of a patient many medical documents are created. The information contained in these documents may either be structured, unstructured, or both. As shown in Figure 6, for patients, cases of illness, movements, and services performed, a link with medical documentation is possible (class `link for assignments to documents`). For all three multiple links may exist as indicated by the multiplicity `0..*`. However, for each of them, medical documentation is not obligatory (multiplicity `0..1`).

In the entire hospital there are many kinds of documents each containing specific information. Nevertheless, as shown in Figure 7, still some commonalities exist for them. First of all, to each document link, multiple documents may be attached (class `header document` with associated multiplicity `1..*`). In this way, it can be indicated that a document consists of multiple lines. With regard to medical documentation, specific documents which are created by a nurse and medical documents which are created by a medical specialist are distinguished. The documents created by a nurse (class `nursing EPR`) are typically centered around the clinical admission of a patient[6]. Examples are the anamnesis done at the start of the clinical admission (class `nursing EPR: nursing anamnesis`) and the evaluation of pain where a patient suffers from (`nursing EPR: anamnesis pain evaluation`).

For documentation created by medical specialists, we make a distinction between a generic document (class `EPR: generic document`) and documents which are specific to a medical discipline or a process that is executed (class `process/discipline specific document`). Note that the generic document contains general medical data for a patient such as length, weight and blood pressure. To illustrate medical discipline specific documents, Figure 7 shows documents specific for gastro-enterology (classes `gastroenterology document`,

[5]Note that EPR is an abbreviation for Electronic Patient Record.

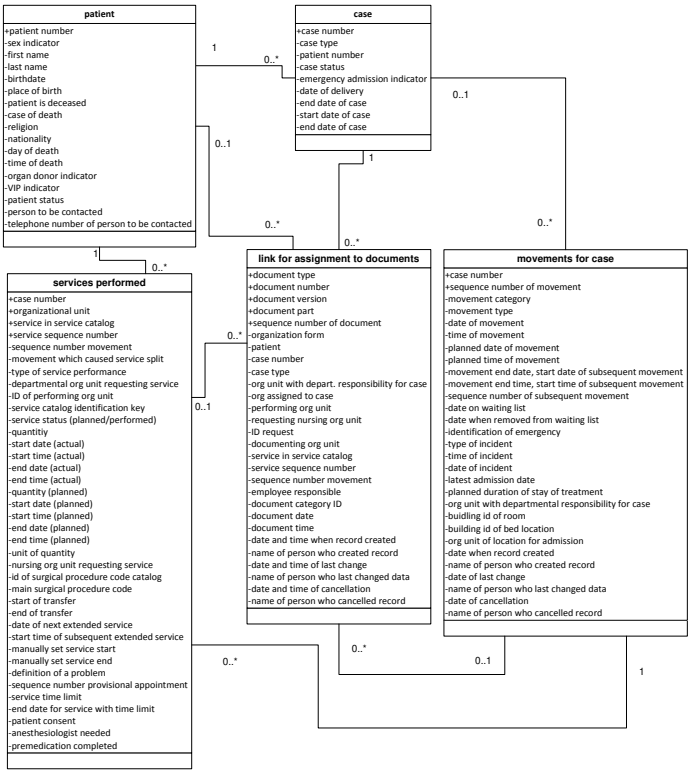


Figure 6: Linkage of medical documents to patients, cases, movements, and performed services. For each of them, multiple links may exist (`link for assignment to documents` class).

`gastro-enterology: colorectal carcinoma`, `gastro-enterology: examination of the body`, and `gastro-enterology: multidisciplinary meeting`) and cardiology (classes `cardiology`, `cardiology: MRT`, and `cardiology: TTE`). These documents are very specific and only a few examples are illustrated in Figure 7.

To illustrate process specific documents we list some documents that exist for the entire surgical process which starts from the moment that an order is created until the admission of the patient at the nursing ward. First, there are general surgery documents (class `surgery: general`) such as the `surgery: peroperative registration` document outlining general information around the entire surgical process. Second, there are documents for the preoperative phase. For example, the `surgery: anesthesia preoperative strategy` document contains information that is important before the surgery starts. Third, there are the documents related to the surgery itself (class `surgery: surgery phase`). This may involve the materials and devices that are used (classes `surgery: technical devices` and `surgery: dressings/tamponades`), the anesthesia procedure (class `surgery: anesthesia procedure`), and the recording of events during the surgery (class `surgery: times registration`). Finally, there are the documents around the postoperative phase containing information that is important for after the surgery (e.g. the medication that is given (class `surgery: postoperative medication`)).

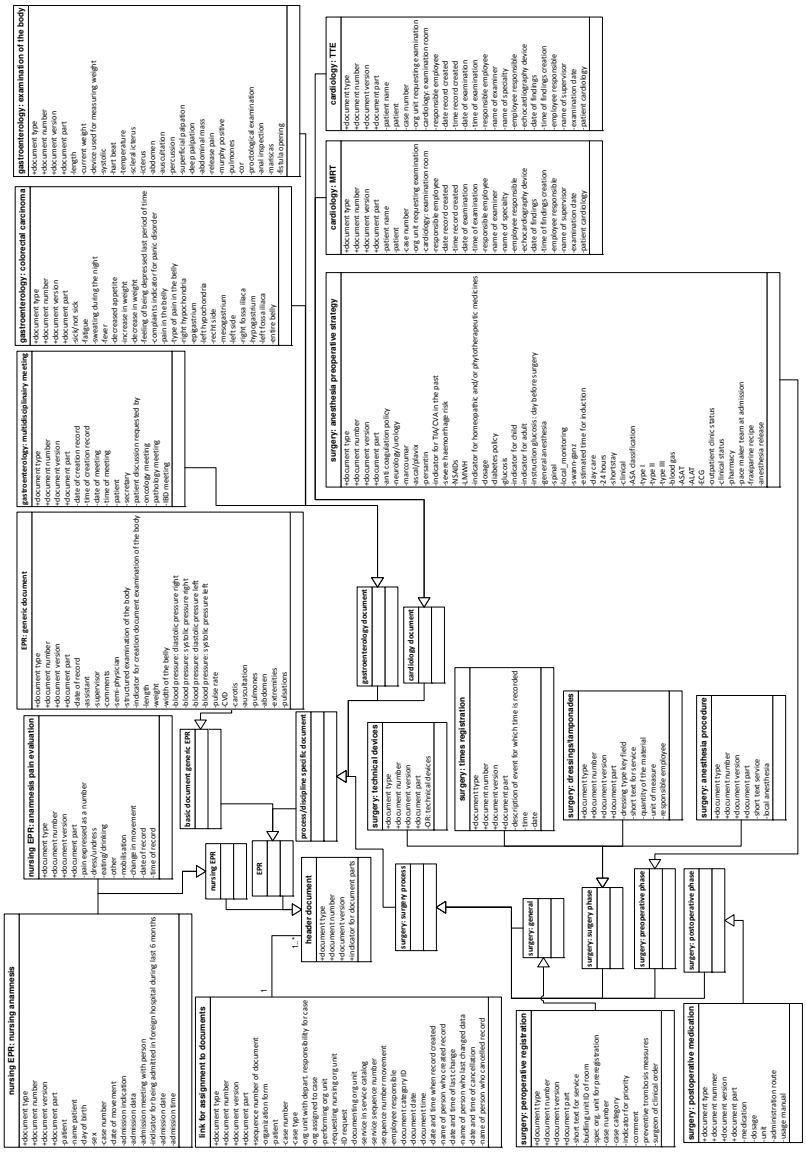


Figure 7: Different kinds of medical documents that can be found in a hospital. There are specific documents that are created by a nurse (`nursing EPR` class) and that are created by a medical specialist (`EPR` class). For medical specialists, either generic documents exist (`EPR: generic document` class) or documents which are specific to a medical discipline or a process that is executed (`process/discipline specific document` class). Note that TTE is an abbreviation for Transthoracic Echocardiogram and MRT for Magnetic Resonance Tomography.

Note that within the information system of MUMC there exist in total 550 different documents. From these documents there are 62 documents related to the surgery process. However, for each HIS the number of documents may be different. Above, we have indicated how these documents can be classified. Also, a few document types have been explicitly mentioned but there are many more.

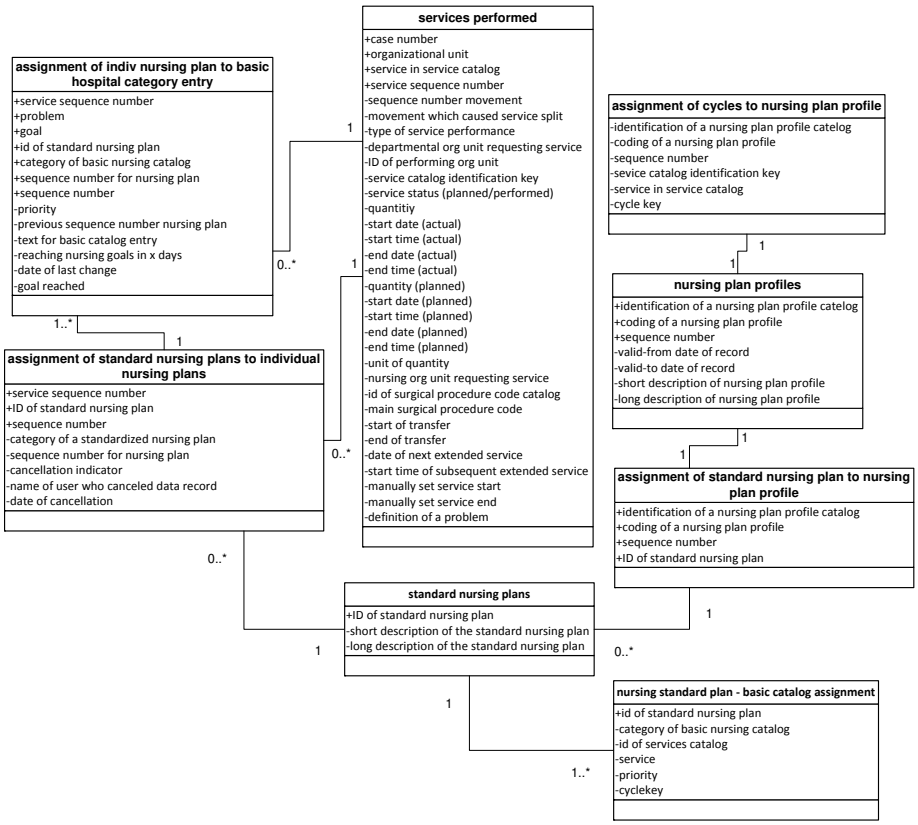


Figure 8: Classed related to nursing plans. A nursing plan (`nursing plan` class) is part of a nursing plan profiles (`nursing plan profiles` class). A nursing plan is individualized to the needs of a patient and multiple services may be executed for it (`assignment of standard nursing plans to individual nursing plans` class).

3.4. Nursing Plans

In Figure 8 the corresponding classes for nursing plans are shown. In case a patient is admitted, nursing plans define the nursing care that is provided to a patient and his relatives. A nursing plan consists of the services that will be provided by a nurse in order to handle problems that are identified by a nursing assessment. A nursing plan (class `standard nursing plans`) may be part of a nursing plan profile (classes `assignment of standard nursing plan to nursing plan profile` and `nursing plan profiles`). Furthermore, for a nursing plan it is indicated which services are part of it (class `nursing standard plan - basic catalog assignment`).

With regard to a nursing plan, it is important that it can be individualized according to the specific needs of the patient. So, for each patient a selection is made from the services that are part of a nursing plan. These selected services form together an individualized nursing plan (`assignment of a standard nursing plan to individual nursing plans`). Afterwards, the selected services are performed (class `services performed`).

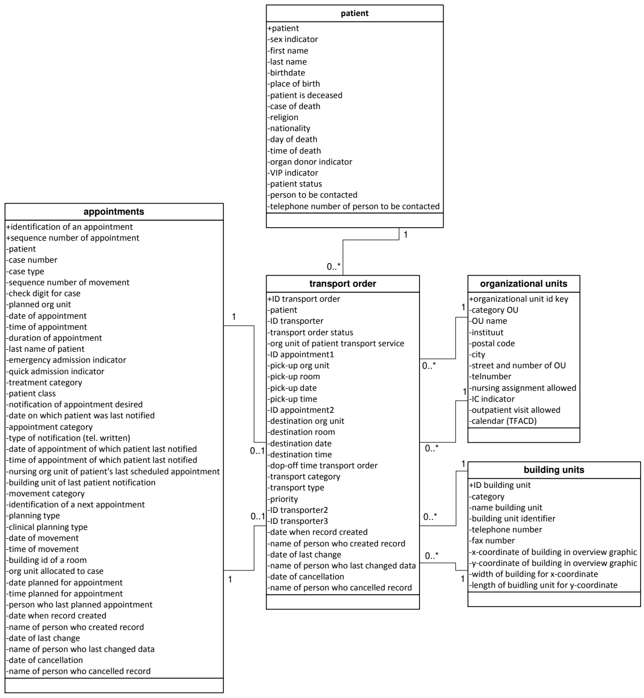


Figure 9: The transportation of patients within the hospital. Each transport order (`transport order` class) defines the appointment from which the patient needs to be picked-up and the appointment to which the patient needs to be transported.

3.5. Patient Transport

Some patients need to be transported within the hospital. Figure 9 shows the classes that are involved around the transport of patients. In order for a transport to take place, a transport order needs to be created (class `transport order`). A transport order may have various attributes, e.g., the transporter needs to be defined, the details of the appointment and the room from which the patient needs to be picked-up and the details of the appointment and the room to which the patient needs to be transported need to be given. Note that information about the person who created, modified, and canceled a record together with its timing are logged.

3.6. Medication

Many patients receive medication during their stay in the hospital. The classes that are related to the provision of medication are given in Figure 10. For the majority of medications that are given first an order is created (class `medication, drug order`). So, medications may also be given on an incidental basis if needed. Note that the drug order can either be linked to a case (the attached multiplicity is `0..*`) or a patient in case the medication is relevant for multiple cases (the attached multiplicity is `0..*`). Within an order the ordering organizational unit and responsible employee is indicated. Furthermore, the validity of the medication and its number of repeats can be indicated.

Once it is decided that a medication is needed, the medication itself needs to be administered (class `medication event`). As for one order, the medica-

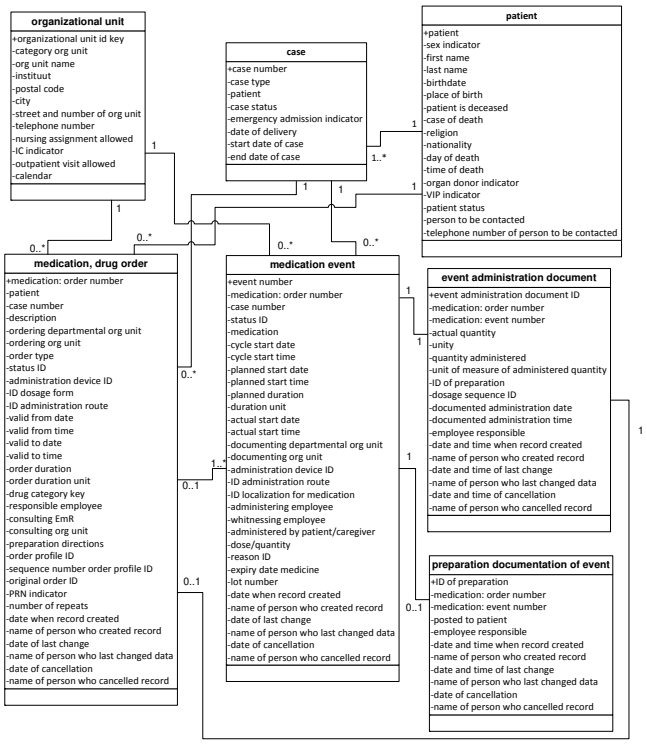


Figure 10: The provision of medication to patients. As a first step, a medication order needs to be created (`medication, drug order` class). Afterwards, the medication may be provided multiple times (`medication event` class). For each provided medication, its preparation is recorded (`preparation documentation of event` class) and the person who administered the medication (`event administration document` class).

tion may need to be administered multiple times the multiplicity attached to the `medication event` class is `1..*`. Amongst others, for the administration of medication, it is important that the planned start date and time, the cycle start date and time, the actual administration date and time, the planned dosage, and the person who administered the medication are given. Closely related to the administration of medication is that it may need to be prepared (class `preparation documentation of event` with associated multiplicity `0..1`). Here, it is saved who prepared the medication. Finally, with regard to the administration (class `event administration event with associated multiplicity 1`) the real administration date and time and the real quantity that is administered is saved.

3.7. Radiology

For radiological examinations, typically a well defined workflow exists. The associated classes are shown in Figure 11. In case a radiological examination is requested, a radiology service exists. To this service, medical documentation is attached. Therefore, the `link for assignment of documents` class of Figure 6 is repeated in Figure 11 in order to link medical documentation to the radiology service. In this documentation it is kept which procedure is requested and its priority (class `radiological service / examination`).

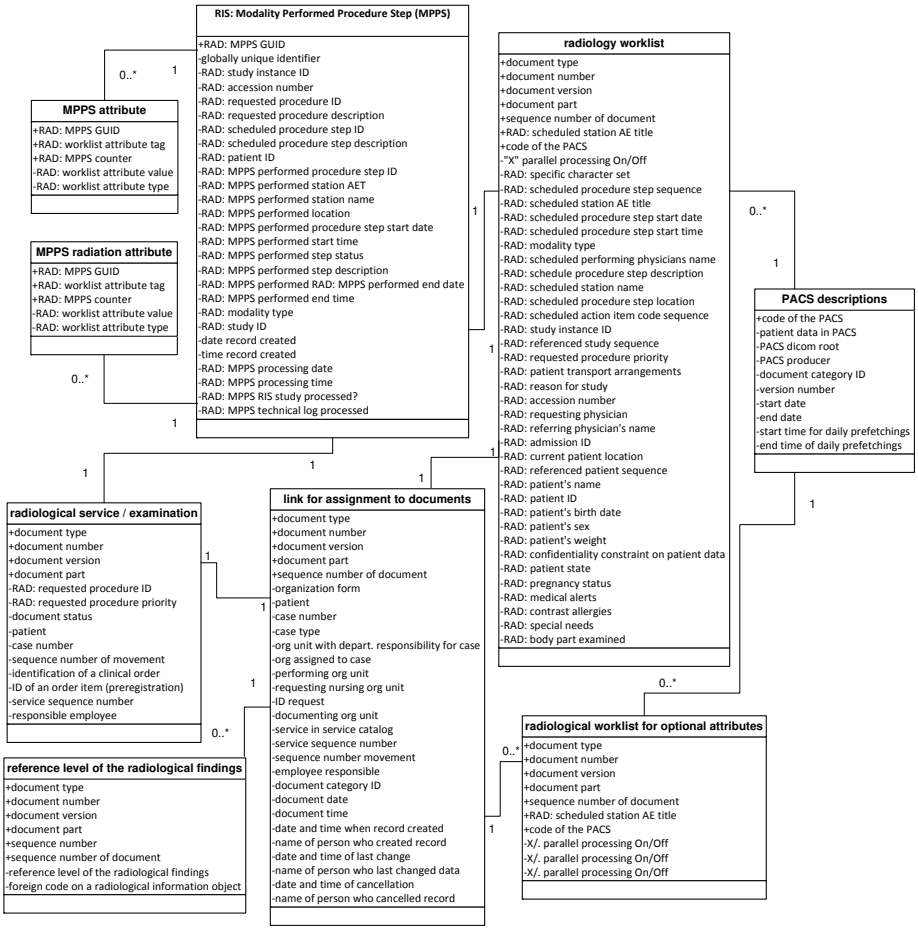


Figure 11: The radiological examinations for patients. As part of such an examination multiple procedures may take place (`radiological service / examination` class) which all appear in the specific worklist that is created for the patient (`radiology worklist` class). Moreover, each procedure is performed on a modality (`RIS: Modality Performed Procedure Step (MPPS)` class).

This is needed for creating the corresponding worklist for the patient (class `radiology worklist`) together with additional attributes (class `radiological worklist for optional attributes`). In this worklist, the entire planning of the required procedure can be found (e.g. the scheduled procedure step, the start date and time of the procedure step, the station at which the procedure is going to take place, and the scheduled physician who will perform the procedure) and the arrangement of the patient transport if needed. In order to be able to perform the procedure properly, general information about the patient is kept. Note that also an *accession number* is kept. This is a well known number in radiology. It represents a single patient encounter to a specific radiological procedure. Therefore, this number is also used for the procedure step that is performed on a modality (class `RIS: Modality Performed Procedure Step (MPPS)`). For each procedure, its actual performed procedure step, timing, processing time, and the station on which the procedure is performed is stored.

Additionally, specific attributes (class `MPPS attribute` with associated multiplicity `0..*`) or radiation related attributes may need to be saved (class `MPPS radiation attribute` with associated multiplicity `0..*`). Finally, the entire examination is completed by reporting the findings (class `reference level of the radiological findings`). This consists of the storage of the obtained images in the Picture Archiving and Communication System (PACS).

3.8. Organization and Buildings

In Figure 12, the classes corresponding to the organizational structure within the hospital and the structure of the buildings and rooms within the hospital can be found. First, the `organizational unit` class describes general information about an organizational unit such as its name, address, or whether it is an intensive care unit or outpatient clinic. In order to make some characteristics of an organizational unit clear it may be part of a category (e.g. whether the assignment of nurses is allowed, class `organizational unit category`). Furthermore, this category can be used for indicating to which case types it may be assigned (class `assignment org units to case types` with associated multiplicity `0..*`). Between multiple organizational units some relationships may exist (class `relations between org units`). Some examples of these relationships are the hierarchical structure that exists between organizational units (class `hierarchy of organizational units`) and the assignment of beds within an organizational unit by another organizational unit (class `inter-dept. bed asgmts in an org unit by another org unit`).

An organizational unit consists of multiple staff members. Some general characteristics of a staff member are described in the `staff` class (e.g. whether somebody is a nurse or a physician). As an organizational unit may consist of multiple staff members the multiplicity associated to the `staff` class is `1..*`. A staff member should be part of at least one organizational unit which is denoted by the `1..*` multiplicity attached to the `organization unit` class. For a staff member multiple attributes may exist (class `attributes staff` with associated multiplicity `0..*`) such as its functions (class `functions`) and its roles (class `roles`). Note that for a staff member, the start and end of its appointment is indicated. Also, the start of the validity of the rank is indicated.

In order for an organizational unit to perform its work, it may be allocated to multiple building units (class `assignment building units to org units`). Analogously to the organizational units, there are general characteristics for building units (class `building units`), additional attributes (class `attributes building units` with associated multiplicity `0..*`), and relationships between building units (class `relations between building units`).

3.9. Pathways

For many illnesses, guidelines exist in order to support in diagnosis and treating patients. The classes around the definition and execution of pathways can be found in Figure 13.

General information concerning a pathway is described in the `treatment pathway` class. Here, it is again interesting that the period for which the pathway is valid is given. A pathway consists of multiple items of which the type may be different (class `treatment pathway item` with associated multiplicity `1..*`). More precisely, the types of the items depend on the language that has been

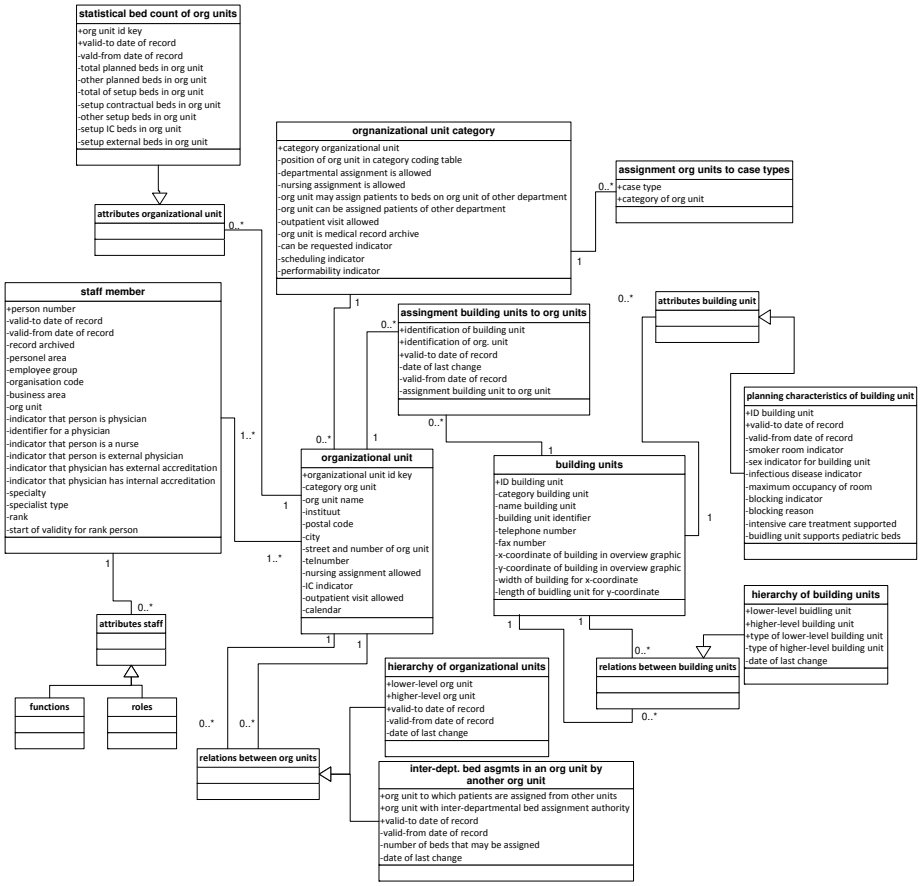


Figure 12: The organizational structure within the hospital and the structure of the buildings and rooms within the hospital. For an organizational unit (`organizational unit` class), multiple kinds of attributes may exist (`attributes organizational unit` class) and relations with other organizational units can be defined (`relations between org units` class). Similarly, for a building unit (`building units` class) different kinds of attributes may be defined (`attributes building units` class) and relations with other building units (`relation between building units` class). Finally, an organizational unit may be linked with building units (`assignment building units to org units` class) and multiple staff members may be part of it (`staff member` class).

chosen for modeling guidelines. For example, a certain type of item may refer to a specific service whereas another type may refer to a construct in order to split a process flow into multiple flows. An item may be connected with multiple other items but this is not required. This is indicated by the multiplicities `0..*` of the associations that are attached to the `connection` class.

A treatment pathway that is executed for a patient is called a *patient pathway* (class `patient pathway`). Note that multiple patient pathways may be executed for a patient and that they are linked to either a patient (the associated multiplicity is `0..*`) or a case (the associated multiplicity is `0..*`). For a patient pathway, specific timing information is recorded such as its assignment, its (planned) start, and its (planned) completion. Also, state information about the start of the next step is recorded. Specific information about the execution

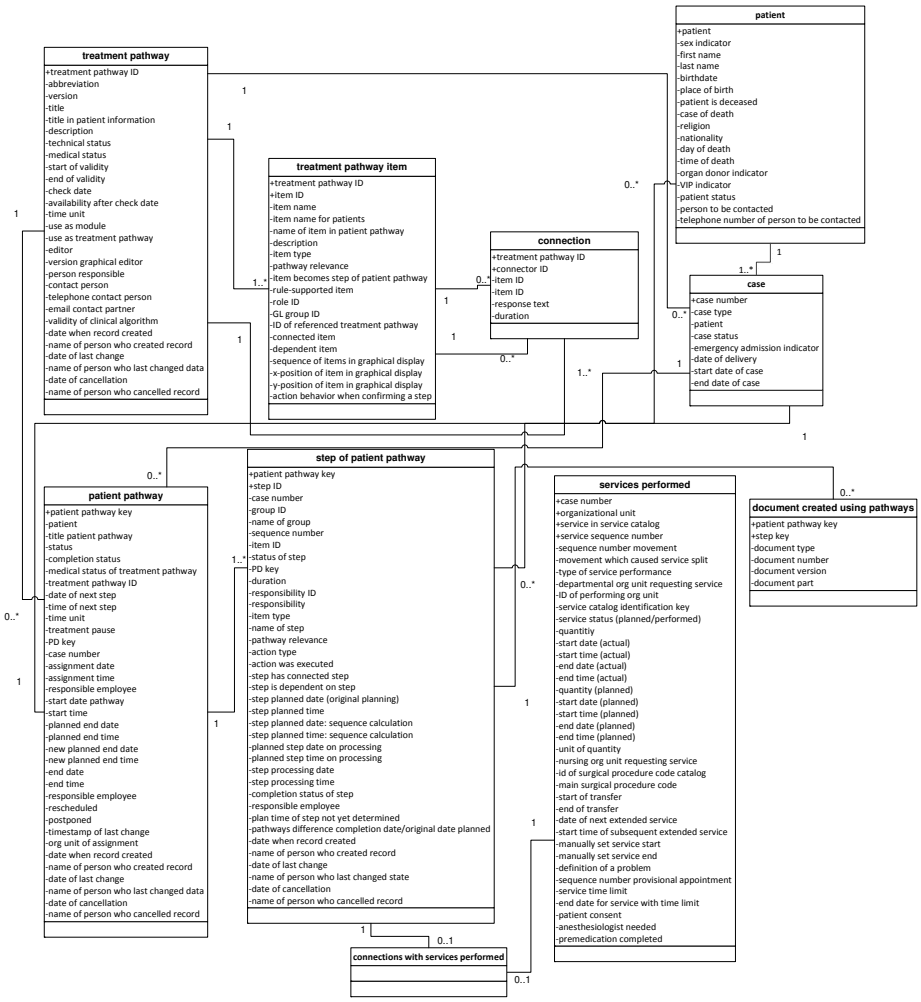


Figure 13: The definition and execution of treatment pathways. A treatment pathway (`treatment pathway` class) consists of multiple items (`treatment pathway item` class) which may be connected to each other (`connection` class). A treatment pathway is executed for a patient (`patient pathway` class) and information about each performed step is recorded (`step of patient pathway` class). Finally, each performed step may linked to a service that is executed for the patient (`services performed` class).

of a step is described by the `step of patient pathway` class. In particular, for the execution of a step it is important that it is known whether it is executed, and by which step it is preceded and succeeded. Furthermore, timing information is recorded about the (planned) start, and execution of the step. Moreover, it is possible to connect a step with a performable service (classes `connections with services performed` and `services performed`). Finally, a step may be linked with medical documentation (class `document created using pathways`). Note that the execution of pathways is related to a series of steps for which it is desired that they are executed. For a patient many more steps may be executed of which their execution is saved as part of one of the classes in the "processes and process steps" group.

4. Possibilities with Process Mining

From the previous section, it can be concluded that a HIS includes an enormous amount of data. Based on the *healthcare reference model* it is possible to generate many logs each focusing on different (parts of) processes. Also, for these logs, the recorded process steps may reside at different levels of granularity. Some examples of possible processes that may be investigated are:

- All the medical services that are performed by gastro-enterology and surgery in order to treat rectal cancer patients.
- All the non-medical services that are performed in order to support the patient and his relatives.
- The surgical care process. This involves the process which starts at the moment the clinical order for the surgery is created until the patient is admitted at the ward again after the surgery. Note that this involves process steps which reside at different levels of aggregation. That is, it encompasses the pre-operative assessment but also the entire lower-level process starting from the moment that the patient is transported to the surgical department until the arrival in the recovery room.
- The different events that are registered for patients during a visit at the outpatient clinic. Examples of these states are: "arrival patient", "start visit to physician", "end visit to physician", and "departure patient".
- The steps performed by nurses in order to care for patients at a specific patient ward.
- The transportation of patients within the hospital.

In this section, the possibilities of process mining within all disciplines of a typical hospital are illustrated. To start, we first assume that the data described in the reference model is indeed available and that no data quality issues apply, i.e., data is assumed to be complete and correct. Given this, we give examples of process mining analyses that are possible. In particular, we want to focus on analyses that are particularly interesting for the healthcare domain. A large dataset that has been obtained based on the *healthcare reference model* is used for this purpose. To be more precise, from a concrete implementation of a HIS which is in use at the MUMC, a large university hospital in the Netherlands, data has been obtained.

First, for the dataset more details will be provided. Afterwards, we focus on the different kinds of analyses that are possible.

4.1. Dataset from the Maastricht University Medical Center

In this section, we provide more details about the data that has been obtained from the Maastricht University Medical Center (MUMC). The data contained information about a group of 296 gastro-enterology patients suffering from intestinal cancer and which have been treated in the hospital. The patients have been treated in the period of 2008 until 2012. In total, data has been extracted from 60 different tables. In this section, we indicate for which classes from the *healthcare reference model* data has been obtained. Moreover,

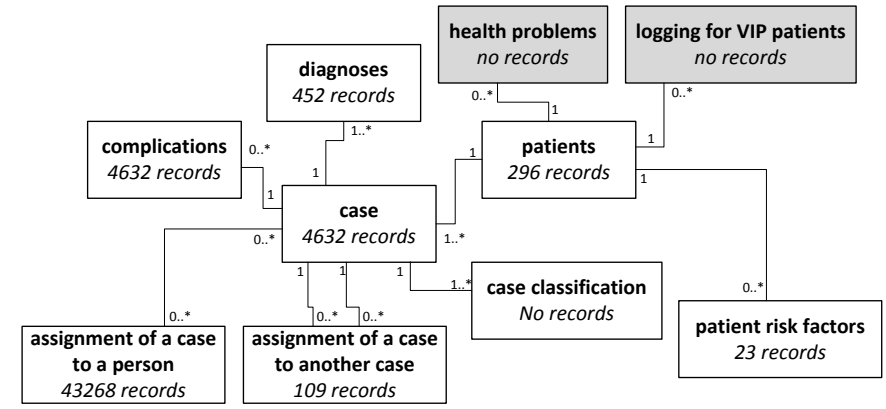
we provide a snippet of the real data for some classes in order to illustrate the raw data that is present in the *healthcare reference model*.

Figure 14 illustrates the data obtained for the "general patient data and case data" and "processes and process steps" groups of the *healthcare reference model*. More precisely, for each group the name of each class is given in a separate rectangle and the relationships between these rectangles are similar as for the classes in the respective group. Furthermore, for each class in Figure 14 it is indicated how many rows are present in the corresponding table in the dataset. If no data has been obtained for a particular class this is indicated by a grey colored rectangle and the "no records" text in it. Furthermore, for all classes which are a generalization of other classes, by definition no data has been obtained: concrete data is associated to the more specific classes. Note that these classes are visualized by a white rectangle in which the name is written in italics. For example, Figure 14a shows for the `case` and `complications` classes of the "general patient data and case data" group, 4632 and 468 records have been obtained respectively. For the `health problems` and `logging for VIP patients` classes, no data has been obtained. Figure 14b shows that for the `movements` and `services performed` classes of the "different kind of process steps" subgroup of the "processes and process steps" group, 41741 and 158918 records have been obtained respectively. For the `context of service` class no data has been obtained as it is a generalization of the `surgery`, `radiology`, and `cardiology` classes. Furthermore, also data for the "document data", "nursing plans", and "organization and buildings" groups of the *healthcare reference model* have been obtained. In the appendix more details are provided about the data that has been obtained for them.

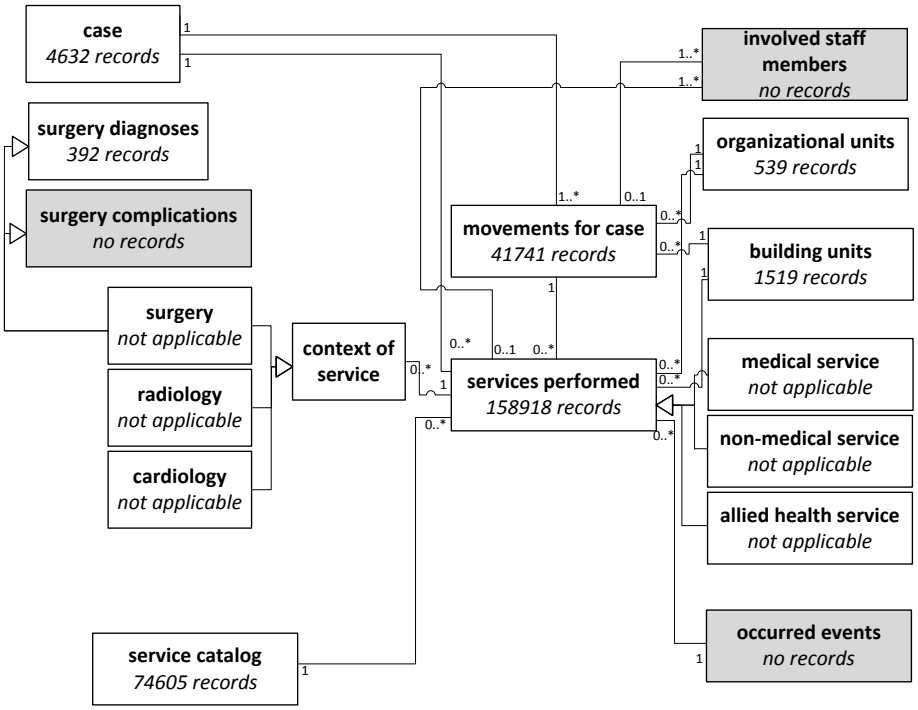
Figure 15 provides an insight into the data that is present in the *healthcare reference model*. More precisely, for the "general patient and case data" group and the "processes and process steps" group, two snippets have been provided of the data that has been obtained. Note that the data has been anonymized in order to maintain confidentiality. Regarding the data for the `patient` class of the "general patient and case data" group, we show for example that patient "p1" is Dutch and has been deceased. Moreover, the religion field shows value "02" which means that the patient is Roman Catholic. For the `services performed` class of the "processes and process steps" group it can be seen that service "v1" has both been started and completed at "30-10-2011". Moreover, as requesting medical discipline the value "INT" has been filled in which refers to the "internal medicine" discipline. Note that for several columns in both snippets only identifiers are given (e.g "religion", "sex indicator", "performing medical discipline", "requesting medical discipline", and "performing org unit"). For each of these columns a corresponding table exists in the *i.s.h.med* system in which more details are provided for the identifiers that are used in the column (e.g. a description).

4.2. Process Mining Analyses

In this section, we elaborate on the application opportunities of process mining given the *healthcare reference model*. In particular, we focus on the types of analyses that are most interesting for the healthcare domain and that give a broad overview of the possible types of process mining analyses. Note that the selection of analysis types is also based on our experiences obtained when analyzing data coming from healthcare processes.



a) data obtained for the 'general patient and case data' group (see Figure 2)



b) data obtained for the 'different kind of process steps' subgroup of the 'processes and process steps' group (see Figure 4)

Figure 14: For both the "general patient data and case data" and the "processes and process steps" groups of the *healthcare reference model* it is shown for which classes data has been obtained from the *i.s.h.med* system of the MUMC.

With regard to possible process mining analyses for the healthcare domain, in [11], an overview is given of the type of questions that are frequently posed by medical professionals in process mining projects. In total, four questions have been identified: (1) finding the most frequently followed (exceptional) paths; (2) investigating the differences in care paths followed by different patient groups; (3) analyzing the compliance with internal and external guidelines; and (4) identifying the bottlenecks in the process. Note that the first two questions

patient identifier	date when record created	date when record created	country of birth	cause of death	patient is deceased	marital status	etnical group	religion	sex indicator	nationality	langu age	coun try
Patiënt	Invoer datum	Datum laatste wijz	Gebland patiënt	Doods oorzaak	Tk pat overlede	Burgerlijke staat	Etnische groep	Religie	Geslacht	Nationaliteit	Taal	Land
p1	06/02/2009	21/07/2012			X			02	1			NL
p2	06/02/2009	07/10/2010							1			NL
p3	06/02/2009	13/04/2012						02	1			NL
p4	06/02/2009	03/10/2012						12	1			NL
p5	13/11/2009	03/12/2009							2	NL		NL
p6	12/10/2010	10/11/2010	NL						1	NL		NL
p7	06/02/2009	03/05/2012						02	1		NL	NL
p8	06/02/2009	22/06/2012			X			12	2			NL
p9	06/02/2009	31/12/2012			X			02	1			NL

a) snippet of the data that has been obtained for the *patient* class of the 'general patient and case data' group (Figure 2).

service sequence number	time service performanc e ends	partial delivery quantity of service	date when record created	created by employee	case	time service performanc e starts	date of last change	cancellat ion indicator	ZAT-code	performing medical discipline
Volgnr verrichting	Einddatum Verr	Deel hoeveelheid	Gecreëerd op	Gecreëerd door	Ziekte geval	Begindatum Verr	Gewijzigd op	Storno teken	ZAT-code	Int Uitvoerend Spec
v1	30/10/2011	0	07/11/2011	t1	c1	30/10/2011	09/06/2012		0000074896	KCH
v2	30/10/2011	0	07/11/2011	t1	c1	30/10/2011	09/06/2012		0000070419	KCH
v3	30/10/2011	0	07/11/2011	t1	c1	30/10/2011	09/06/2012		0000070116	KCH
v4	30/10/2011	0	07/11/2011	t1	c1	30/10/2011	09/06/2012		0000070402	KCH
v5	30/10/2011	0	07/11/2011	t1	c1	30/10/2011	09/06/2012		0000074110	KCH

requesting medical discipline	treating medical discipline	Indicator for dummy service	service payable	date of cancellation	record modified by employee	movement	performing org unit	requesting nursing org unit	requesting org unit
Int Aanvragend Spec	Int Behandelend Spec	Teken: Dummy Verrich	Kosten drverricht	Gestorneerd door	Gewijzigd door	Beweging	Uitvoerende OE	Aanvr verpl OE	Aanvr spec OE
INT	KCH		X		f1	6	LCHE	PEHU	SAIG
INT	KCH		X		f1	6	LCHE	PEHU	SAIG
INT	KCH		X		f1	6	LCHE	PEHU	SAIG
INT	KCH		X		f1	6	LCHE	PEHU	SAIG
INT	KCH		X		f1	6	LCHE	PEHU	SAIG

service	service in service catalog	patient	performing physician	anesthesiol ogist
Verrichting	Verricht catalogus	Patiënt	Uitvoerend Arts	Anesthesist
0000370423	RG	p1	a1	
0000370419	RG	p1	a1	
0000370403	RG	p1	a1	
0000370402	RG	p1	a1	
0000370401	RG	p1	a1	

b) snippet of the data that has been obtained for the *services performed* class of the 'processes and process steps' group (Figure 4).

Figure 15: Some snippets of the data that has been obtained from the *i.s.h.med* system in use at the MUMC. For each column the Dutch description has been provided.

relate to the discovery type of process mining, the third to conformance, and the last one to extension (see Figure 1).

Consequently, we propose the following five types of analysis: (1) the exploration of selections of events; (2) the construction of a precise model; (3) the identification and quantification of deviations; (4) the identification and quantification of bottlenecks; and (5) drilling down into the data. Note that the first two analyses relate to the first frequently posed question, the third analysis to the third question, and the fourth analysis to the fourth question. The fifth analysis relates to all questions. Below, each of them will be discussed in more detail. Moreover, each process mining analysis type will be illustrated using the earlier mentioned dataset. Note that for the various examples we will use data according to different classes of the reference model in order to illustrate some possible processes that can be investigated.

4.2.1. Exploring Selections of Events

For this type of analysis the focus is on exploring a selection of events. During

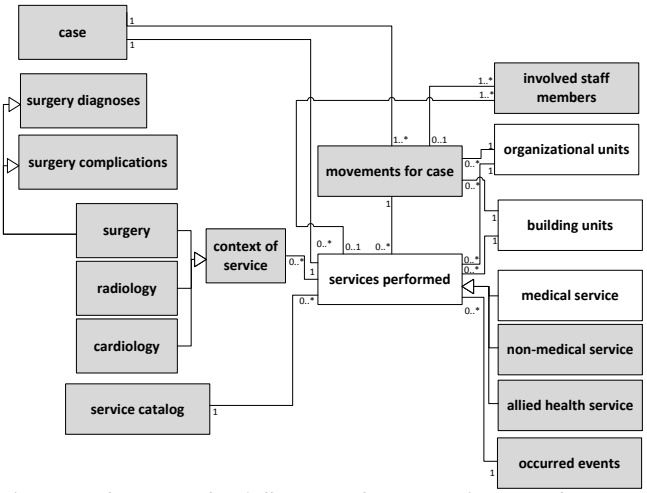
this analysis, it is important to decide on a specific scope, i.e. from which classes of the *healthcare reference model* events need to selected and which are the most important events for these classes. So, process mining analyses are executed which allow for getting a visual view on the characteristics of the data (e.g. the dotted chart [12], the Basic Performance Analysis plug-in) and for obtaining initial process related insights (e.g. the heuristics miner [2] and the fuzzy miner [13]).

Figure 16 and Figure 17 illustrate this type of analysis. Figure 16a shows that data has been obtained from the `services performed`, `organizational units`, `building units`, and `medical service` classes. Note, that also here for each group the name of each class is given in a separate rectangle and the relationships between these rectangles are similar as for the classes in the respective group. If data has been obtained for a class, the corresponding rectangle is given a white color whereas for a class for which no data has been obtained the associated rectangle is colored grey. For the group of gastro-enterology patients, we have selected the services that have been performed for the patients suffering from large intestine cancer until the surgical intervention. In total, this resulted in a log with 105 patients, 6225 events, and 516 event classes.

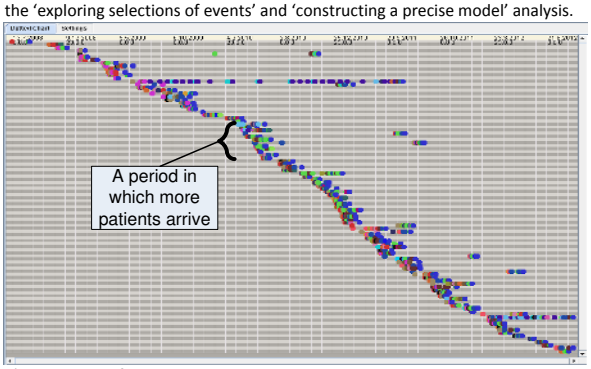
Figure 16b and Figure 16c both show a dotted chart which visualizes events as dots. As such, a "helicopter view" of the process is obtained and patterns can be discovered. On the vertical axis the different cases (i.e. patients) are shown and events are colored according to their activity names. In Figure 16b the process is shown using actual time, i.e. all cases are sorted based on the first event that took place. As can be seen from the chart, the arrival pattern of patients is pretty stable. However, there are periods in which only a few patients arrive (e.g. June/July 2010) or periods in which more patients arrive (e.g. March 2010 till June 2010). In Figure 16c the process is shown using relative time, i.e. all cases start at time zero. The chart shows that for a small group of patients the time until surgery is much longer. That is, for 26 out of the 105 patients the time until surgery is more than 60 days. This group of patients turned out to be complex cases for which an individualized treatment was necessary. For the majority of patients this is less than 60 days. Also, the variation in time is small as shown by the steep line in the chart.

For getting the first process related insights, it has been decided to only include the patients for which the time until surgery was less then 60 days. Also, we focused on the most frequent events. In this way, not an overly complex model will be obtained when applying a control-flow discovery algorithm. In the Figures 17a and b, the discovered models by respectively the fuzzy miner and the heuristics miner are shown. In both models, the services performed and the ordering between these services can be seen. Both the fuzzy miner and the heuristics miner have been chosen as they 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.

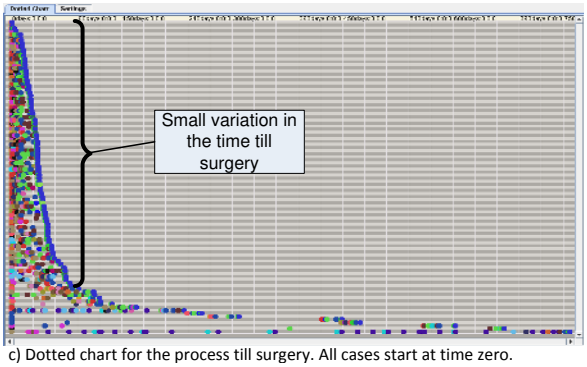
Both models reveal some clear causal relations between services. Note that for the fuzzy miner the significance of a relation between two services are indicated by the thickness of the corresponding edge between these services, i.e. more significant relations have a wider edge. So, as important causal relations we find both in the models discovered by the fuzzy miner and the heuristics miner that the "preoperative assessment" service is followed by the "admission hospital" service. After, the admission there is either a "hemicolectomy",



a) Overview of the classes of the 'different kind of process steps' subgroup of the 'processes and process steps' group for which data has been taken in order to perform the 'exploring selections of events' and 'constructing a precise model' analysis.



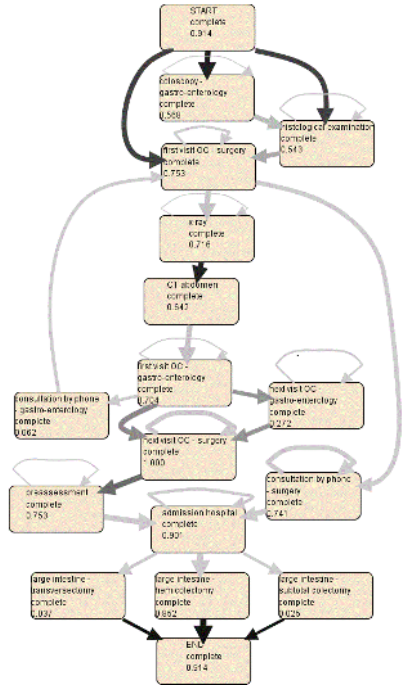
b) dotted chart for the process till surgery. All cases start at their real time.



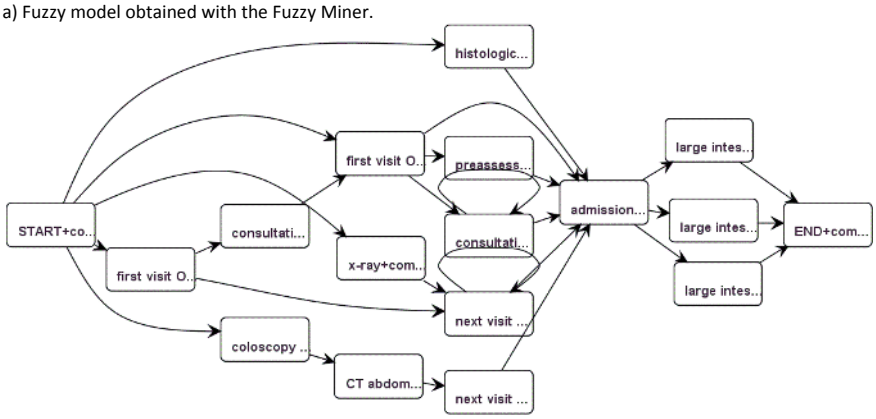
c) Dotted chart for the process till surgery. All cases start at time zero.

Figure 16: The data and some results that have been obtained during the "exploring selections of events" analysis.

"transversectomy", or "subtotal colectomy" surgery. Moreover, the fuzzy miner shows that after the "coloscopy" there is a "histological examination" and after the "histological examination" there is a first visit to the outpatient clinic of surgery ("first visit OC - surgery" service). Also, after the first visit to the outpatient clinic of surgery an "X-ray" occurs and after the "X-ray" a "CT-



a) Fuzzy model obtained with the Fuzzy Miner.



b) Heuristics Net obtained with the Heuristics Miner.

Figure 17: Two models showing the discovered control-flow for the group of 89 patients suffering from large intestine cancer.

abdomen" occurs.

Moreover in the fuzzy miner the brightness of edges between nodes emphasizes their correlation, i.e. more correlated relations are darker. So, for example, for the "X-ray" and "CT abdomen" services it is indicated in the model that they are highly correlated. In this case, dependent on the chosen settings for the fuzzy miner, it is indicated that the temporal proximity between these services is high.

As a result of the analysis, it turned out that several patients (with long overal throughput times) fall outside the scope and that some clear causal relationships exist between several activities. Using the above mentioned analyses

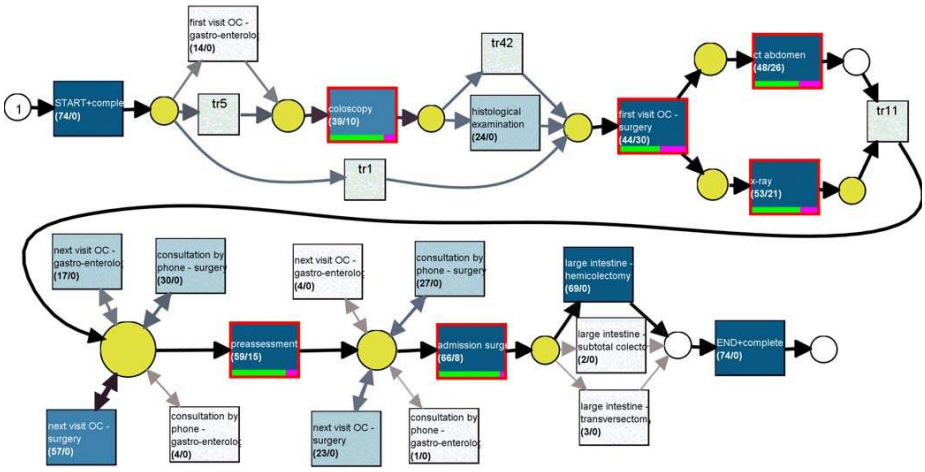


Figure 18: The Petri net model that has been obtained during the "constructing a precise model" analysis. Using the "conformance analysis" plug-in of ProM it was checked whether the adapted model reflected the behavior in the model well.

these things can obviously be identified.

4.2.2. Constructing a Precise Model

The models produced by a process mining algorithm give clear insights about the way the process was executed. However, depending on the goal of the analysis it may be needed to, (semi-)automatically tailor a discovered model to satisfy certain requirements. Examples of these requirements are that the obtained model is a good reflection of the behavior captured in the log (i.e. the model fits the log), the obtained model is simple, or the model allows only minimally more behavior than seen in the log.

For illustrating the construction of a precise model, we focus on the approach that has been used for coming to a precise model based on the control-flow models that have been obtained for the previous analysis (Figure 17). As mentioned in Section 4.2, medical professionals are often interested in learning about the performance of the process in order to identify the bottlenecks in the process. In order to reliably enrich a given (discovered) model with performance information it is important that the precise model is a good reflection of the behavior captured in the log. So, first the model obtained by the heuristics miner was converted into a Petri net. Based on the additional insights shown in the fuzzy model, the Petri net model was adapted by hand and it was checked in the process mining tool ProM how well it reflected the behavior in the log. This second step was repeated until the model was a good reflection of the behavior captured in the log. The resulting Petri net model is shown in Figure 18. Note that the Petri net is annotated with diagnostics generated by the conformance checker plug-in.

In brief, the process of Figure 18 is as follows. First, a coloscopy ("coloscopy") takes place followed by a histological examination ("histological examination"), if needed, or the patient immediately visits the outpatient clinic of surgery ("first visit OC - surgery"). After, the visit to the outpatient clinic of surgery, both a

CT abdomen ("CT abdomen") and X-ray ("X-ray") is performed followed by a next consultation at the outpatient clinic ("next visit OC - gastro-enterology" and "next visit OC - surgery") or a consultation by phone ("consultation by phone - gastro-enterology"). Subsequently, a preassessment ("preassessment") takes place and a next consultation, if needed, before the patient is admitted to the hospital ("admission surgery" service). During the admission, a surgery on the large intestine takes place ("large intestine - hemicolectomy", "large intestine - subtotal colectomy", and "large intestine - transversectomy"). As shown in the model, there are some parts in the process where the process does not conform with the log. For example, after the first visit to the outpatient clinic of surgery, not always both a CT abdomen or X-ray takes place.

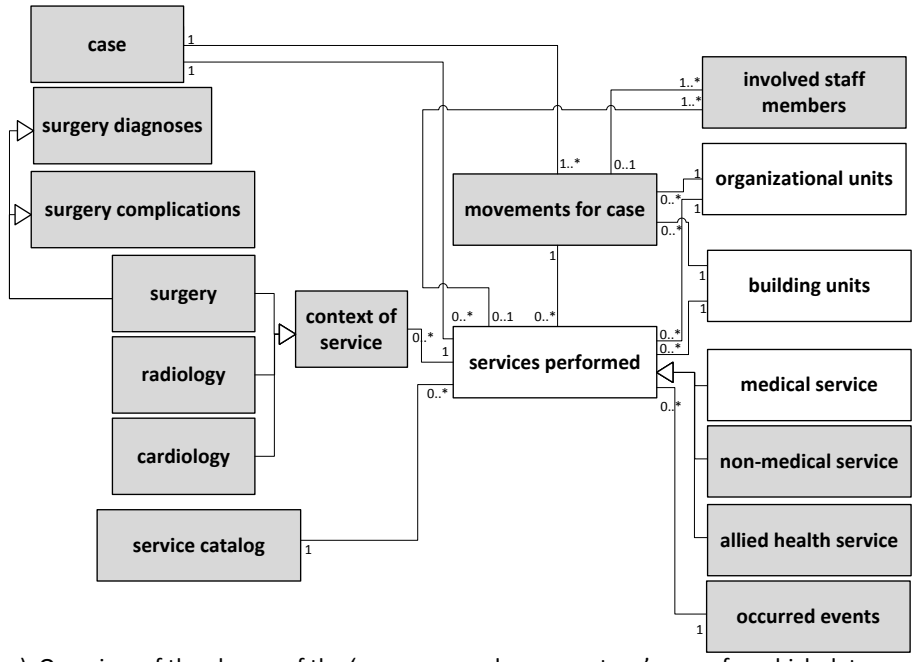
All together, this analysis demonstrated an approach for coming to a model which is a good reflection of the behavior captured in the log. In next analysis steps the model can be enriched with additional information (e.g. performance information).

4.2.3. Identifying and Quantifying Deviations

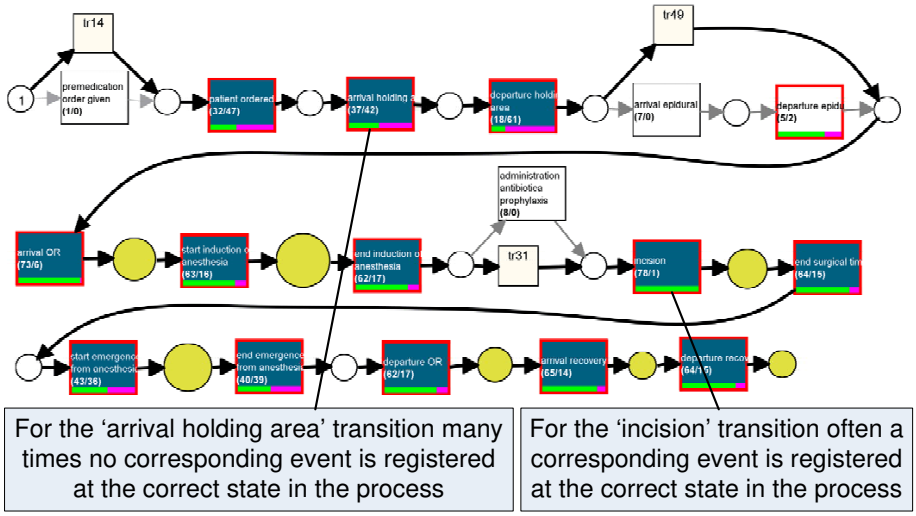
For this type of analysis, the aim is to investigate the conformance between a given process model and an event log. As such, tasks in the model are identified which should have occurred but did not occur in reality ("move on model") or tasks are identified which have occurred in reality although this was not anticipated in the model ("move on log"). Moreover, the aforementioned deviations are quantified.

In order to illustrate this type of analysis, consider Figure 19. Here, data has been obtained from the `occurred events` class. More precisely, for the group of 79 patients of the previous analysis, we have obtained all the events that are registered in the context of a surgery that is performed. These events involve the period starting from the transportation of the patient to the surgery department until the arrival in the recovery room. In total, this resulted in a log with 79 patients, 767 events, and 17 event classes. Note that the recording of events is done manually. Furthermore, for the events that need to be registered in the aforementioned period, a "normative" process model exists which describes the events that need to be recorded and the ordering of them.

The model is shown in Figure 19 together with diagnostic information about the deviations between the log and the model. First, a premedication order is given if needed ("premedication order given"). After that the patient is ordered, the patient arrives in the holding area in order to be prepared for surgery ("arrival holding area" and "departure holding area"). Next, the patient may need to be transported to a different room in order to allow for epidural anesthesia ("arrival epidural" and "departure epidural"). For the surgery, the patient is transported to the surgery room ("arrival OR") after which the anesthesia team starts the induction of the anesthesia ("start induction of anesthesia"). Subsequently, there is the option to provide an antibiotica prophylaxis treatment ("administration antibiotica prophylaxis"). The surgery starts with the incision ("incision") and finishes with the closure of the tissue with stitches ("end surgical time"). Next, the anesthesia team performs the emergence from the anesthesia ("start emergency from anesthesia"), which ends when the patient has regained consciousness ("end emergency from anesthesia"). Subsequently,



a) Overview of the classes of the 'processes and process steps' group for which data has been taken in order to perform the 'identifying and qualifying deviations' analysis.



b) Obtained Petri net annotated with diagnostics generated by the conformance checker plugin.

Figure 19: For a group of 79 patients conformance information is collected for the events that need to registered in the context of a surgery that takes place.

the patient leaves the OR ("departure OR") and is transported to the recovery room ("arrival recovery"). Once the patient is stable enough, the patient is transported from the recovery room to the nursing ward ("departure recovery").

As the events in the model are recorded manually, there are several deviations between the model and the log. In order to identify and quantify these deviations

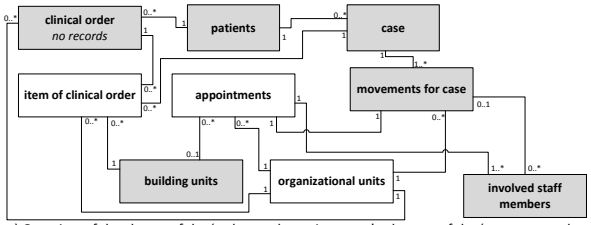
the model is annotated with diagnostic information by the conformance checker plug-in. More precisely, for each transition the ratio is given between the cases for which there existed a corresponding event during replay of the transition (the green colored bar) and the cases for which there was no matching event during replay of the transition (the purple colored bar). The size of the yellow colored places indicates for the respective state in the model the number of cases in which for the event to be replayed no matching transition could be found. For example, it can be seen that the "incision" event most often has been registered correctly whereas for the "departure holding area" this was exactly the opposite.

Next to that, the conformance checker allows for checking for each patient which parts of the model could be successfully replayed and which not. For example, for 5 cases it was found that the "patient ordered", "arrival holding area", and departure "holding area" events have not been recorded although this is mandatory. Also, via the plug-in it was discovered that the average fitness was only 0.77 (note that the minimal value is "0" and the maximal value is "1") and that there were no cases which could be successfully replayed. Overall, during the analysis it becomes clear that due to the manual registration many events are not registered or registered in the wrong order. Using conformance analysis deviations between a process model and an event can clearly be identified and quantified.

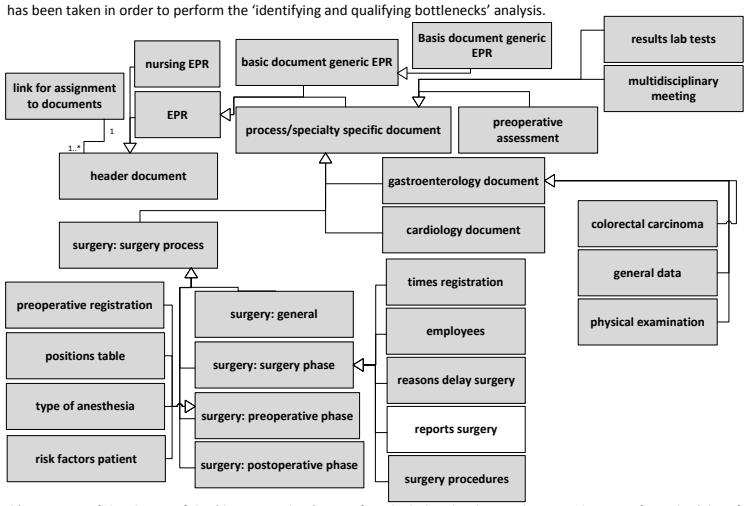
4.2.4. Identifying and Quantifying Bottlenecks

In healthcare there is a lot of attention for preventing unnecessary waiting times. As such it is important that performance information can be obtained for a healthcare process. For this type of analysis the focus is on the identification and quantification of bottlenecks within the process.

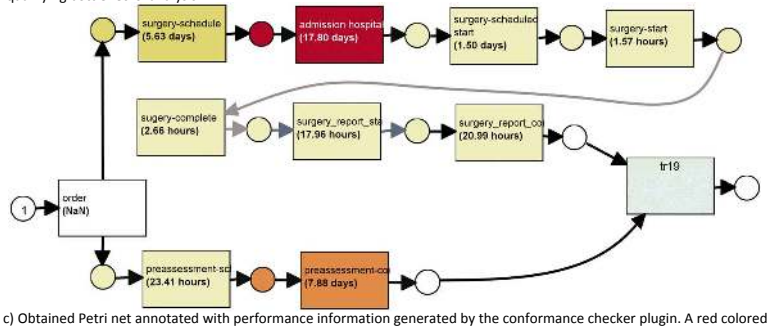
Figure 20 illustrates this type of analysis. The constructed event log contains data from the `items of clinical order`, `appointments`, and `organizational units` classes. More precisely, for the group of 79 patients of the "exploring selections of events" analysis type we have now collected information about the steps that are performed in the context of ordering and organizing all appointments that are needed for a surgical intervention and the surgical report that is produced afterwards. The process is depicted in Figure 20b. First, an order is created for the surgery ("order") which involves the scheduling and execution of several steps. One step relates to the scheduling ("preassessment schedule") and the occurrence of an appointment for the preassessment ("preassessment complete"). Another step involves the scheduling ("surgery schedule") of an appointment for the surgery, the admission to the hospital ("admission hospital"), and the execution of the surgery itself. Note that for the surgery a distinction has been made between the scheduled start of the surgery ("surgery scheduled start"), the actual start of the surgery ("surgery start"), and the actual completion of the surgery ("surgery complete"). Finally, after the surgery, a surgical report is created ("surgery report start") and finalized afterwards ("surgery report complete"). Note that obtaining execution data for this process involved collecting data from several classes of the *healthcare reference model*. More specifically, execution data for the "order" task were collected from the *item of clinical order* class, execution data for the "preassessment start", "preassessment complete", "surgery schedule", and "surgery scheduled start" tasks was collected from the *appointments* class, and execution data for the "surgery



a) Overview of the classes of the 'orders and appointments' subgroup of the 'processes and process steps' group for which data has been taken in order to perform the 'identifying and qualifying bottlenecks' analysis.



b) Overview of the classes of the 'document data' group for which data has been taken in order to perform the 'identifying and qualifying bottlenecks' analysis.



c) Obtained Petri net annotated with performance information generated by the conformance checker plugin. A red colored transition indicates a high average waiting time for the respective event whereas on the other end of the spectrum a light yellow colored transition indicates a low average waiting time for the respective event.

Figure 20: For a group of 79 patients performance information is collected for the process of ordering and organizing the steps that need to be done for a surgical intervention.

start", "surgery complete", "surgery report start", and "surgery report complete" tasks were obtained from the *reports surgery* class. In total, this resulted in a log with 79 patients, 639 events, and 11 event classes. Clearly, the *healthcare reference model* was of great help for identifying the existence of the aforementioned kinds of information and for the subsequent creation of the event log.

For the process of Figure 20c, performance information has been projected on it by coloring the places and the transitions. A red colored transition or place indicates a high average waiting time whereas on the other end of the spectrum a white colored transition or place indicates a low average waiting time. When inspecting the figure, several important insights can be obtained. For example, once scheduling a surgery, still it takes on average 17.80 days (standard deviation: 7.66 days) before the patient is admitted to the hospital. Also, on average there are 5.63 days (standard deviation: 7.07 days) between the creation of the order and the scheduling of an appointment for the surgery. Finally, there are on average 17.96 hours (standard deviation: 1.46 days) between the completion of the surgery and the creation of the surgery report.

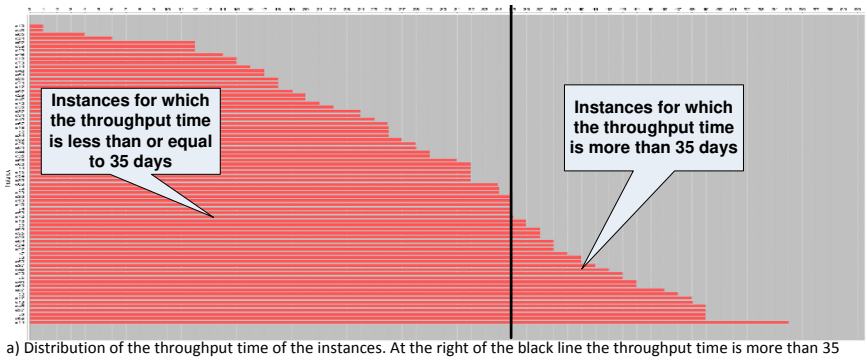
As a result of the analysis, some obvious performance bottlenecks are identified. This kind of information is crucial for determining measures for improving the efficiency of the process.

4.2.5. Drilling Down

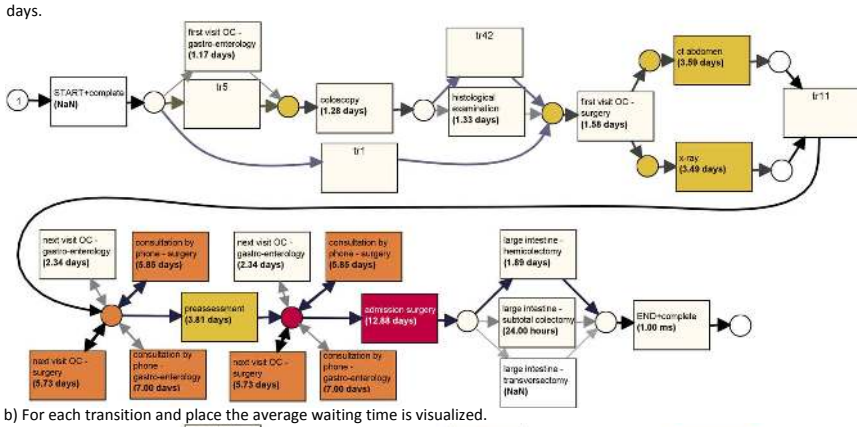
Given the outcome of a certain analysis it may be desired to further drill down into the data in order to investigate a (part of a) process in more detail. This drilling down may be done according to different dimensions. One dimension may be the *case type* dimension in which cases are selected that satisfy a certain property (e.g. the cases that are late). Another dimension may be the *event class* dimension which involves the selection of events satisfying certain properties (e.g. the name or resource of the event). Also, the *time* dimension may be used for focusing on a certain time window (e.g. a particular week or the activities performed in the first half year).

Examples of process mining techniques for further drilling down into the data are clustering techniques (e.g. the guide tree miner [14]) or decision point analysis techniques (e.g. the decision point analysis plug-in [15]). Moreover, within ProM several log filtering techniques can be used to drill down (e.g. the LTL checker [16]).

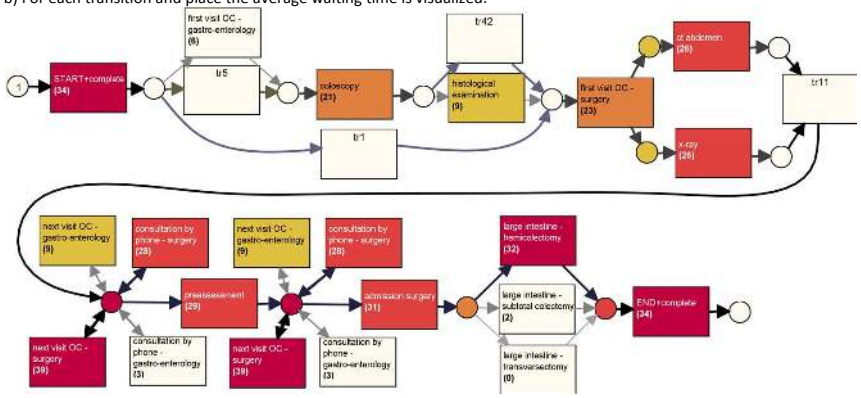
In order to illustrate this type of analysis, consider Figure 21. Here, we further drill down into the analysis that is performed for the group of 79 patients suffering from large intestine cancer and for which the associated process is shown in Figure 18. That is, in Figure 21a for all patients the distribution of the time until surgery is shown. In the Netherlands, the guideline of the Dutch cancer association, regarding an acceptable time between the start of the trajectory until surgery, states that this should be at most 35 days. In the figure, these 35 days are indicated by a vertical black line. As becomes clear, for a quite a large group of patients this guideline is violated (34 patients). In order to identify possible reasons for this violation, the process followed by this group of patients is investigated in further detail. Therefore, in Figure 21b performance information has been projected on the model in a similar way as done during the "identifying and quantifying bottlenecks" analysis type. Additionally, in Figure 21c for the transitions and the places the associated frequency is depicted. A red color indicates a high frequency whereas on the other end of the spectrum a white color indicates a low frequency. Both models make clear that there is a high waiting time for the admission to the hospital ("admission hospital" transition, average: 12.88 days, standard deviation: 7.55 days) and that this transition is also often executed (31 times). Also, for the next visit to the



a) Distribution of the throughput time of the instances. At the right of the black line the throughput time is more than 35 days.



b) For each transition and place the average waiting time is visualized.



c) For each transition and place the frequency is visualized.

Figure 21: At the top for the group of 79 patients (see Figure 18) the distribution of the time until surgery is given. Subsequently, for these patients for which the time till surgery is more than 35 days, in the middle, the process model is annotated with information about the average waiting time for both transitions and places. At the bottom, the process model is annotated with information about the frequency of both transitions and places.

outpatient clinic of surgery ("next visit OC - surgery" transition, average: 5.73 days, standard deviation: 2.50 days, occurrence: 39) and a consultation by phone of the surgery department ("consultation by phone - surgery" transition, average: 5.85 days, standard deviation: 4.44 days, occurrence: 28) there are reasonable high waiting times and the occurrence of these activities is also high. So, in order to improve the average throughput time of the entire process it

needs to be investigated whether the average waiting time of these three previous activities can be reduced. All together, this kind of analysis obviously revealed this kind of knowledge for the group of patients for which the time until surgery is more than 35 days.

5. Data Quality Issues

So far, we have introduced our *healthcare reference model* and demonstrated some of the many types of process mining it enables. While doing this, we assumed that there were no data quality issues, i.e., the data is available and correct. However, in reality we need to copy with many data quality issues [17]. In order to obtain reliable and trustful process mining results it is important that the analyst is aware of the kind of data quality issues that may manifest in a HIS and that may negatively impact the process mining analysis. Therefore, in this section, we investigate which kind of data quality issues may exist for the data that is part of the *healthcare reference model*. In order to identify these data issues, the following approach is taken. In [17], 27 data quality issues are presented which may apply for an event log. For the data present in the HIS of the MUMC, we investigate which of the 27 quality issues apply.

In Section 5.1 we briefly introduce the data quality issues that have been identified in [17]. In Section 5.2, the classification of such issues is used for evaluating the data that is present in the HIS of the MUMC.

5.1. Classification of Event Log Quality Issues

In [17], in total 27 event data related quality issues that may occur within an event log have been identified. For these issues the following four classes of broad problems have been identified. Note that the issues defined in [17] relate specifically to an event log. However, it is trivial to generalize these issues from a single event log to the event data stored in a HIS.

- **Missing Data:** Here different kinds of process mining information are *missing* although the information is mandatory. For example, an event, a process instance, or an attribute/value of an event may be missing.
- **Incorrect Data:** Although process mining data may be provided, it may be the case that the provided information is logged *incorrectly*. For example, for the timestamp of an event an impossible value has been provided.
- **Imprecise Data:** Here the information that is logged is too coarse leading to a *loss of precision*. As result, certain kinds of analysis are not possible anymore as a more precise value is needed in order to obtain reliable results. For example, the timestamps that are logged for certain events may be too coarse (e.g. in the order of a day) thereby making the order of entries unreliable.
- **Irrelevant Data:** Here the logged information may be irrelevant as it is for analysis but another relevant entity may have to be derived/obtained (e.g., through filtering/aggregation) from the logged entities. For example, for a certain analysis, the performance of a lab test may be sufficient instead of considering all the individual lab tests that have been performed.

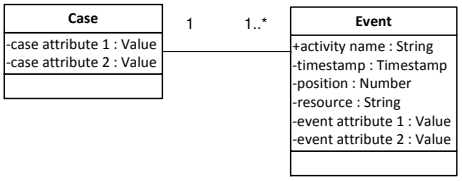


Figure 22: Cases and events have attributes and each event can be associated to a single case.

The four classes of problems hold for different entities within an event log. In Figure 22 these entities are depicted by means of a UML class diagram. That is, an event log captures the execution of a process. As part of this, for the process, multiple *cases* are stored. For each case, multiple attributes may be stored such as case id, etc. Each case consists of an ordered list of *events*. As an event can only belong to one case only it is explicitly stored to which case an event belongs. This is defined by the *relationship* relation. For an event, it is mandatory that it refers to an activity or task. Note that in our case, an event may for example refer to a service, a movement, or an event that occurred (the `movements for case`, `services performed`, and `occurred events` classes in Figure 4). Therefore, always a value needs to be provided for the *activity name* attribute of an event. Next to that, an event may have optional attributes such as a *timestamp*, *position*, *resource*, or other attributes. Note that it is important that events within a case are ordered. This ordering is either determined by the timestamp attribute or the position attribute which specifies the index of an event in a case.

Table 1 shows the manifestation of the four classes of problems across the different entities of an event log. In this way, 27 different quality issues have been identified. For each issue an unique number has been given with prefix "I''. For example, the "missing data" quality issue for the "relationship" entity of the log (I3) corresponds to the scenario where the association between events and case are missing. Furthermore, the "imprecise data" quality issue for the "timestamp" entity (I23) corresponds to the scenario where timestamps are imprecise since a coarse level of abstraction is used for the timestamps of (some of the) events. As a result, the ordering of events within a case may be unreliable.

Table 1: For each entity of an event log it is indicated whether the "missing", "incorrect", "imprecise", and "irrelevant" problem type is possible. For example, "I23" refers to imprecise timestamps, e.g., events have a date but no precise timestamp.

	case	**event**	**relationship**	**c_attribute**	**position**	**activity name**	**timestamp**	**resource**	**e_attribute**
Missing Data	I1	I2	I3	I4	I5	I6	I7	I8	I9
Incorrect Data	I10	I11	I12	I13	I14	I15	I16	I17	I18
Imprecise Data			I19	I20	I21	I22	I23	I24	I25
Irrelevant Data	I26	I27							

Table 2: Evaluation of data quality issues for the *i.s.h.med* system in use at the MUMC. In case a certain quality issue is not applicable, the associated box is colored black. The value "N" indicates that the issue does not occur. Furthermore, an "L" indicates that a quality issue occurs relatively infrequently whereas an "H" indicates that an issue occurs more frequently in comparison with other issues.

	case	**event**	**relationship**	**c_attribute**	**position**	**activity name**	**timestamp**	**resource**	**e_attribute**
Missing Data	N	H	L	L	N	L	N	N	L
Incorrect Data	N	L	L	L	N	L	L	N	L
Imprecise Data			N	N	N	N	H	H	N
Irrelevant Data									

For a detailed discussion of all the identified quality issues, we refer to [17].

5.2. Evaluation of Data Quality Issues

In this section, the 27 issues event log quality issues listed in Table 1 are used for evaluating the data that is present in the HIS of the MUMC. In this way, we can illustrate the kind of process mining related data quality issues that may occur for the entire data present in a hospital.

In order to identify the data quality issues that occur for the *i.s.h.med* system in use at the MUMC, we interviewed people within the hospital which have knowledge about the raw data that is present in the system. Moreover, we have inspected tables in the database ourselves in order to identify the most promising quality issues. In this way, a subjective evaluation is performed. Note that due to the large amount of data present in the entire system we consider it infeasible that an objective evaluation is performed. Moreover, for some quality issues it is difficult to obtain objective measures (e.g. the amount of events that are missing).

More specifically for the evaluation, we have focussed on identifying quality issues whose occurrence is more frequent than the occurrence of other data quality issues. In this way, we can identify the issues which are the most prominent. Therefore, as shown in Table 2, for each quality issue we distinguish four different values. The character "N" indicates that the quality issue does not occur. An "L" means that the quality issue may be present but does not occur frequently. A value "H" indicates that a quality occurs more frequently, i.e. is more prominent compared to other issues. Finally, the box associated to a quality issue is colored black if the quality issue does not apply. For example, as each entry in the HIS of the MUMC is relevant for a certain patient, the irrelevant data quality issues do not apply. Below, the quality issues, for which either an "L" or an "H" have been given, are briefly discussed.

- **Missing Events (I2)**: Many events need to be entered manually into the system. As a result, during our discussions we identified that one of the most prominent problems in the HIS is that people forget to enter services although they have been performed in reality. Therefore, as value for the evaluation an "H" is given.

- **Missing Relationship (I3)**: In principle, every event needs to be linked with a case. However, in the system there are events which are not linked to a case but it is expected that its frequency is low. For example, for the table associated to the services that have been performed for patients (class `services performed` in Figure 4) there are in total 67,295,460 events registered. For 17,568 of them (approximately 0.03%) there is no associated case stored. Consequently, as evaluation value an "L" is given.

- **Missing Case Attributes (I4)**: Also for the case attributes it holds that they need to be entered manually into the system. However, here it is expected that they are filled in in a better way than the events. Therefore, as evaluation an "L" is given.

- **Missing Activity Name (I6)**: Although for each event an activity name needs to be provided it may occur that this does not happen in reality. As a result of our discussions with data experts of the MUMC it is expected that its occurrence is low. For example, regarding the services that have been performed for patients (class `services performed` in Figure 4), there are only 84 events for which no activity name has been provided. Consequently, as value for the evaluation an "L" is given.

- **Missing Timestamp (I7)**: Also regarding timestamps it may happen that no value is registered. However, in many cases a timestamp is recorded automatically by the system itself. Therefore, together with the experts of the MUMC it has been decided to provide a "L" as value for the evaluation.

- **Missing Event Attribute (I9)**: Analogously to the case attributes, as value for the evaluation an "L" is given.

- **Incorrect Event (I11)**: By incident an event may be recorded for a patient which did not happen in reality. Here, it is anticipated in the hospital that its occurrence is low. Therefore, as evaluation the value "L" is given.

- **Incorrect Case Attribute (I13)**: For a case attribute it may always be the case that a wrong value has been given. However, during our discussions it was expected that this should be quite exceptional. Therefore, the value "L" is provided as evaluation.

- **Incorrect Activity Name (I15)**: In the system, events are found which are unlikely regarding the illness for which a patient is treated. For example, instead of a CT abdomen it is recorded that a CT-scan of the foot has been made. As for this kind of quality issue it is not expected that it occurs frequently, an "L" has been given as evaluation.

- **Incorrect Timestamp (I16)**: The entire system contains a wealth of timestamp information. As already indicated before, many of the timestamps are automatically saved as part of a record that is saved in the system. However, there are also records for which the associated timestamp is saved by hand. For example, regarding the services that have been performed for patients (class `services performed` in Figure 4), there are

124 of the 67,295,460 events (0.0001%) which have "31.12.2020" as associated timestamp. As for this quality issue it is not expected that it occurs frequently, as evaluation value an "L" has been given.

- **Incorrect Event Attribute (I18)**: For an event attribute it can always be the case that a wrong value has been given. Therefore, in a similar fashion as for the "incorrect case attribute" data quality issue the value "L" is provided as evaluation.

- **Imprecise Timestamp (I23)**: Within the hospital there are several medical disciplines which use their own dedicated system for the medical services they perform. In order for the hospital to be reimbursed for these services, they are imported afterwards into the *i.s.h.med* system. During this import only the day is saved on which the services have been performed. As this is the case for a considerable group of events, as evaluation value an "H" is given. For example, for only the radiology department and pathology departments there are 636363 and 664675 services stored respectively for which only a day timestamp is given. Note that in the own systems of these departments more precise timestamp information is available.

- **Imprecise Resource (I24)**: In the system for each action it is saved which resource recorded it. However, in some cases the saved resource may not refer a specific person. For example, regarding the services that have been performed for patients (class `services performed` in Figure 4), there are 234,378 of the 67,295,460 events (0.3%) for which the recorded resource refers to a specific operating room. As this issue tends to occur more frequently compared to other issues, as evaluation value an "H" is given.

The discussion above makes clear that the HIS of the MUMC suffers from several data quality issues identified in [17]. Moreover, these issues are also very likely to occur for HIS implementations. Given the existence of these issues, it is important that they are detected and properly handled (e.g. by applying repair techniques in order to alleviate these issues). Some process mining techniques exist which are able to deal with some quality issues (e.g. the fuzzy miner [13] and the heuristics miner are able to handle missing events [2]). Also, in [18] several methods are given for detecting time related quality issues.

Despite data quality issues that may apply, process mining is still possible. For example, events or cases for which issues may exist can easily be filtered. As a result, many analyses are still possible and many interesting insights can be obtained.

6. Related Work

A HIS plays an important role in a hospital as many people rely on it. To this end, it is important that HISs are systematically managed and operated [1]. Due to the complexity of such a system, particular requirements exist on methods for modeling such systems [19]. A well known method is the 3LGM^2 meta model [20, 21] which offers three layers, i.e., the domain layer consists of

enterprise functions and entity types, the logical tool layer focuses on application components and the physical tool layer describes physical data processing components. Another modeling method is MOSAIK-M [22], a method and tool to support modeling, simulation, and animation of information and communication systems in medicine. Although 3LGM^2 and MOSAIK-M provide a lot of information about the information that is stored in a HIS, none of them provides a method for showing the real event data that are stored in the system and infer meaningful and surprising process insights from them.

Next to this, several publications exist about the functionality that is provided by a particular HIS and its realization within a hospital. That is, in [23] the functionalities and the architecture of the Soarian system are described together with details of how it has been realized in a German hospital, in [24, 25] the functionalities and the components of the *i.s.h.med* system are elaborated upon and its realization in two hospitals (in Austria and Germany), and finally in [26] the advantages and disadvantages of an interfaced approach for a clinical information systems architecture are described based on a HIS implementation in the USA. For all publications, only high level information is given about the functionality that is provided and the data that is available in the system.

When considering the literature about the application of process mining in the healthcare domain it can be seen that the application of process mining in the healthcare domain is increasing. That is, in total, 40 scholarly publications have been identified in which a real life application of process mining in healthcare is reported. Here, a clear distinction can be made between papers that focus on the discovery of a healthcare process (e.g. [27–29]) and the works that focus on checking the conformance of a healthcare process (e.g. [30–32]).

Regarding the discovery of healthcare processes, 33 scholarly publications have been identified. That is, in [27, 33–37] the gynaecological oncology healthcare process within an university hospital has been analyzed; in [28, 38] several processes within an emergency department have been investigated; in [39] all Computer Tomography (CT), Magnetic Resonance Imaging (MRI), ultrasound, and X-ray appointments within a radiology workflow have been analyzed; in [40] the focus is upon the process that is followed by patients suffering from rheumatoid arthritis; in [41], the a cardiology process and a process within an emergency unit has been discovered; in [42] the treatment of patients within an intensive care unit has been investigated; in [42–44] process mining has been applied to different datasets for stroke patients; in [29] the activities that are performed for patients during hospitalization for breast cancer treatment are investigated; in [45] the journey through multiple wards has been discovered for inpatients; in [46] the processes for mamma care patients and diabetes foot patients has been investigated; in [47], the workflow of a laparascopic surgery has been analyzed; in [48] the anesthesia procedure during Endoscopic Retrograde Cholangiopancreatography (ERCP) has been studied; in [11] the process for colorectal cancer patients requiring surgery has been analyzed; in [49], process mining has been applied for understanding the process for patients requiring an implant-borne single crown restoration; in [50] the process corresponding to the guideline for treatment of ear infection has been discovered; in [51] the involvement of multiple medical disciplines for treating patients has been investigated; in [52] the workflow for congestive heart failure patients who underwent a radiological procedure has been discovered; in [53] the focus is on the patients following the bronchial lung cancer pathway; in [54] process mining has been used for

discovering the process for patients diagnosed with Late Onset Neonatal Sepsis (LONS); in [55] the focus has been on identifying the surgical model for cataract interventions; in [56] the pre-operative care process has been discovered; in [57] the activities that are performed for patients during radiotherapeutic treatment have been investigated; in [58] process mining has been applied to the event logs of more than thousand X-ray machines all over the world; in [59] some details are provided about a process that has been discovered for an Austrian hospital; in [60] the focus is on the discovery of a diabetes process; and finally in [61] the ordering of the activities that are performed by nurses within a hospital are investigated.

With regard to checking the conformance of healthcare processes, only 7 scholarly publications have been identified. In [62] the conformance of a surgery process has been investigated; in [30] the adherence to a cardiovascular pathway has been measured; in [31, 63] a declarative specification has been used for checking the conformance of a chronic cough guideline; in [32, 64] conformance checking has been applied to the Cutaneous Melanoma treatment process; and finally in [65] the focus has been on the identification of outliers for the cholecystectomy surgery process.

All the above mentioned publications demonstrate that process mining can be successfully applied in the healthcare domain. Also, event data may originate from various data sources in a hospital. For example, the data used in [27, 33, 40, 46] originated from an administrative system within the hospital. Furthermore, data may also come from an intensive care unit [42], neurology department [66], and radiology department [67]. Finally, process mining can also be applied to data of medical devices [47, 58]. However, for all earlier mentioned publications, only data from one or two sources has been used. However, these applications seem not to exploit the full potential of process mining due to the absence of an overview of all the data within a HIS that can be used for process mining and the characteristics of this data.

In contrast to existing literature we provide a comprehensive overview of all the data in a hospital that can be used for process mining and the characteristics of such data.

7. Discussion

In this paper, we present a *healthcare reference model* outlining for a hospital all the different classes of data that are potentially available for process mining together with the relationships between these classes. Subsequently, based on the reference model we have indicated the possibilities that exist for applying process mining in the entire hospital. That is, we have provided some examples of analyses that are possible due to the reference model and existing process mining techniques. As shown such analyses can be very useful for hospitals that want to improve their processes.

In order to measure the performance of a process, typically Key Performance Indicators (KPIs) are used. Examples of these KPIs are the total throughput time, utilization, and number of surgeries performed. However, whereas these KPIs can only indicate that there is a problem, process mining *looks inside* the process by discovering the typical ordering of process steps, i.e. control flow, bottlenecks, and deviations. So, based on factual data an objective view is obtained on how processes are really executed. Subsequently, the processes

can be improved. In [11] an overview is given of the type of questions that are frequently posed by medical professionals in process mining projects. These questions relate to identifying the process that is commonly followed, differences in care paths for different groups of patients, compliance to internal and external guidelines, and the identification of bottlenecks within the process. As has been indicated in Section 4, a wide variety of event logs covering different processes can be generated based on the *healthcare reference model*. For all these processes, the same or comparable questions can be answered using process mining.

When focusing on the questions that can be answered, several advantages can be imagined with regard to the *healthcare reference model* as well. One is that the reference model shows the kind of data that may be assumed to be present within a hospital. Based on this knowledge, the following approach can be used for the application of process mining. As a first step, based on the reference model, medical professionals can come up with questions that they like to be answered with process mining. As a second step, targeted data extraction efforts can be made for collecting the required data. Depending on the IT landscape of the hospital, data may be taken from one or more systems. Another advantage is that the reference model creates an *awareness* of all the data that is present within the hospital. When we were busy with developing the reference model, several employees of the MUMC were pleasantly surprised to see much more data than expected (e.g. referral information of patients and all information saved in the context of a surgery). Furthermore, at the IT department we saw that many people were specialized in a certain part of the system. However, only few people have a good overview of all the data that was present in the entire system.

Despite the aforementioned advantages, several challenges still remain. These are discussed below.

In order to perform a certain analysis, data from different sources may need to be collected. Merging data extracted from different sources may be difficult as there might be no shared identifier. In order to deal with the problem of correlating events belonging to the same case, it might be necessary to use some heuristics. Also, within the hospital it can be assumed that already a lot of knowledge exists about "connecting data together".

Furthermore, in Section 4 we have shown how drilling down into the data allowed for quickly arriving at the cause of a particular problem. Within the process mining field, we are not aware of a similar functionality. Therefore, future work is needed in order to develop process mining approaches which allow for easily drilling down into the data and "slicing and dicing" of the data.

A limitation of our approach is that for developing the reference model we mainly focussed on the documentation that existed for the *i.s.h.med* system and the implementation of this system within a single organization. Nonetheless, in order to increase the general applicability of the reference model, we have also received feedback from HIS professionals of two other Dutch hospitals, and to a lesser extent we have investigated the HIS within another hospital and which was from another software supplier (a database which came from the Chipsoft EZIS system which is in use at the Catharina hospital in Eindhoven). For example, based on the latter system we have made a distinction in the reference model (see Figure 4) between medical (class `medical service`), non-medical (class `non-medical service`), and allied health services (class `allied`

`health service`). Also, for services performed (see Figure 4) and appointments (see Figure 5), we have introduced a specific class for recording the involved persons (class `involved staff members`). In principle, HISs that are provided by other software suppliers may contain data and functionalities that are not present within our reference model. However, we feel that by following the above mentioned approach, the developed *healthcare reference model* may considered to be representative for other HIS implementations.

Another important issue is that medical professionals are typically not aware about process mining and its possibilities. Moreover, performing a good and reliable process mining analysis is not trivial. Therefore, in order to increase the uptake of process mining within the hospital one ore more people specifically need to be educated to this end. So, based on important results that are obtained by them, the technique can become part of the standard analysis toolset of the MUMC hospital.

8. Conclusion

In this paper, we presented a *healthcare reference model* which exhaustively lists the typical data that exists within a HIS and that can be used for process mining. Based on this reference model, many different event logs can be created each focussing on a different (part of a) process. Moreover, among healthcare professionals, the model aids in providing awareness of all the process related data that exist and that can be used for analysis.

In this context, we have given several examples of process mining analyses exploiting the information that is present within the reference model and the relationships that exist for the data. These examples made clear that the following kind of analyses (and many more) can be performed.

1. Exploring selections of events such that the scoping of the analysis becomes clear.
2. The construction of a precise model which satisfies certain model requirements.
3. The identification and quantification of deviations in order to investigate deviations between an event log and the model that describes how the process needs to be executed.
4. The identification and quantification of bottlenecks in order to find performance related problems within the process.
5. Drilling down into the data according to different dimension in order to analyse a (part of a) process in more detail.

In healthcare the above mentioned analyses contribute in investigating healthcare processes in great detail. The obtained insights can serve as basis for improving or governing these healthcare processes such that these are performed more efficiently.

Acknowledgements

This research is supported by the Dutch Technology Foundation STW, applied science division of NWO and the Technology Program of the Ministry of Economic Affairs.

Appendix

A The Gastro-Enterology Database

The database that has been obtained contains records related to a group of 296 gastro-enterology patients suffering from intestinal cancer contains 60 tables. Note that the data has been obtained from the implementation of the *i.s.h.med* system in use at the MUMC.

In Section 4.1 details have been provided concerning the classes of the "general patient and case data" group and the classes of the "different kind of process steps" subgroup of the "processes and process steps" group for which data has been obtained. In this section, details will be provided about the data that has been obtained for the remaining subgroups of the "processes and process steps" group and the data that has been obtained for the "document data", "nursing plans", and "organization and buildings" groups. Note that for the "patient transport", "medication", "radiology", and "pathways" groups no data has been obtained.

In a similar fashion as Section 4.1, for each group a separate figure is given illustrating the data that has been obtained. In a figure, a single rectangle is shown for each class in the group and the relationships between these rectangles are similar as for the classes in the respective group. In case for a class data has been obtained, it is indicated how many records are present in the corresponding table. If no data has been acquired for a class, the associated rectangle is colored grey and has as text "no records" in it. For all classes which are a generalization of other classes, the associated rectangle is colored white and the name is written in italics as concrete data for them can be found in more specific classes. So, in figures 23 until 26 it is shown for the "processes and process steps", "document data", "nursing plans", and "organization and buildings" groups respectively for which classes data has been obtained. For example, in Figure 23 it is shown that for the `reference referral data` and the `appointments` class 793 records and 10604 records have been obtained respectively.

[1] A. Winter, R. Haux, E. Ammenwerth, B. Brigl, N. Hellrung, F. Jahn, Health Information Systems: Architectures and Strategies (2nd edition), Springer Verlag Berlin-Heidelberg, 2011.

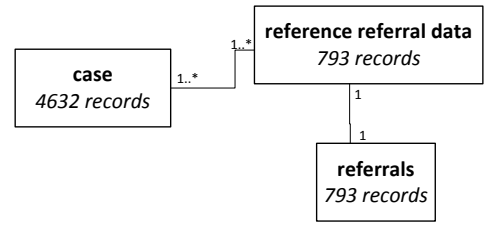
[2] W. van der Aalst, Process Mining: Discovery, Conformance and Enhancement of Business Processes, Springer-Verlag, Berlin, 2011.

[3] K. Jensen, L. Kristensen, Coloured Petri Nets: Modelling and Validation of Concurrent Systems, Springer, 2009.

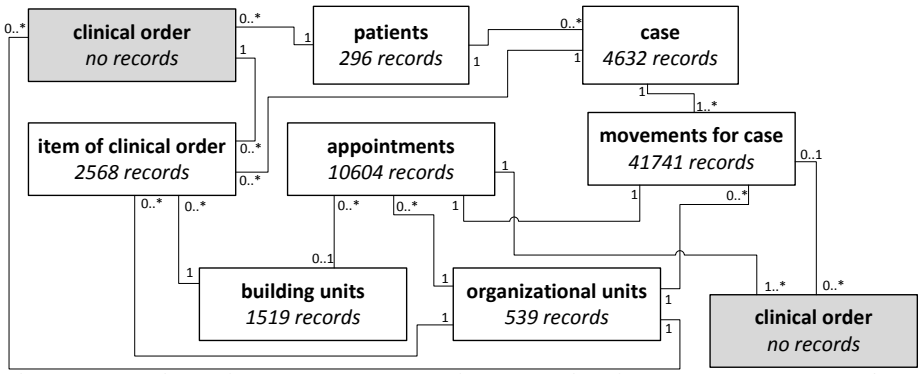
[4] S. White, Introduction to BPMN, BPTrends.

[5] W. M. P. van der Aalst, M. Pesic, M. Schonenberg, Declarative workflows: Balancing between flexibility and support, Computer Science - Research and Development 23 (2) (2009) 99–113. doi:10.1007/s00450-009-0057-9.

[6] G. Keller, M. Nüttgens, A. Scheer, Semantische Processmodellierung auf der Grundlage Ereignisgesteuerter Processketten (EPK), Veröffentlichungen des Instituts für Wirtschaftsinformatik, Heft 89 (in German), University of Saarland, Saarbrücken (1992).



a) data obtained for the 'referral' subgroup of the 'processes and process steps' group (see Figure 3)



b) data obtained for the 'orders and appointments' subgroup of the 'processes and process steps' group (see Figure 5)

Figure 23: For the 'referral' and the 'orders and appointments' subgroups of the 'processes and process steps' group of the *healthcare reference model* it is shown for which classes data has been obtained from the *i.s.h.med* system of the MUMC.

[7] A. Rozinat, W. van der Aalst, Conformance Checking of Processes Based on Monitoring Real Behavior, Information Systems 33 (2008) 64–95.

[8] W. M. P. van der Aalst, A. Adriansyah, B. van Dongen, Replaying history on process models for conformance checking and performance analysis, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2 (2) (2012) 182–192. doi:10.1002/widm.1045.

[9] H. Verbeek, J. Buijs, B. van Dongen, W. van der Aalst, XES , XESame , and ProM 6, in: Information Systems Evolution, Vol. 72 of Lecture Notes in Computer Science, Springer Verlag Berlin-Heidelberg, 2011, pp. 60–75.

[10] O. M. Group, OMG Unified Modeling Language, Version 2.0, Object Management Group, 2005.
URL `http://www.omg.org/spec/UML/2.0/`

[11] R. Mans, W. van der Aalst, R. Vanwersch, A. Moleman, Process Mining in Healthcare: Data Challenges When Answering Frequently Posed Questions, in: R. Lenz, S. Miksch, M. Peleg, M. Reichtert, D. Riano, A. ten Teije (Eds.), Proceedings of the BPM 2013 Workshops, Vol. 7738 of Lecture Notes in Computer Science, Springer Verlag Berlin-Heidelberg, 2013, pp. 140–153.

[12] M. Song, W. van der Aalst, Supporting Proces Mining by Showing Events at a Glance, in: K. Chari, A. Kumar (Eds.), Proceedings of the Seventeenth

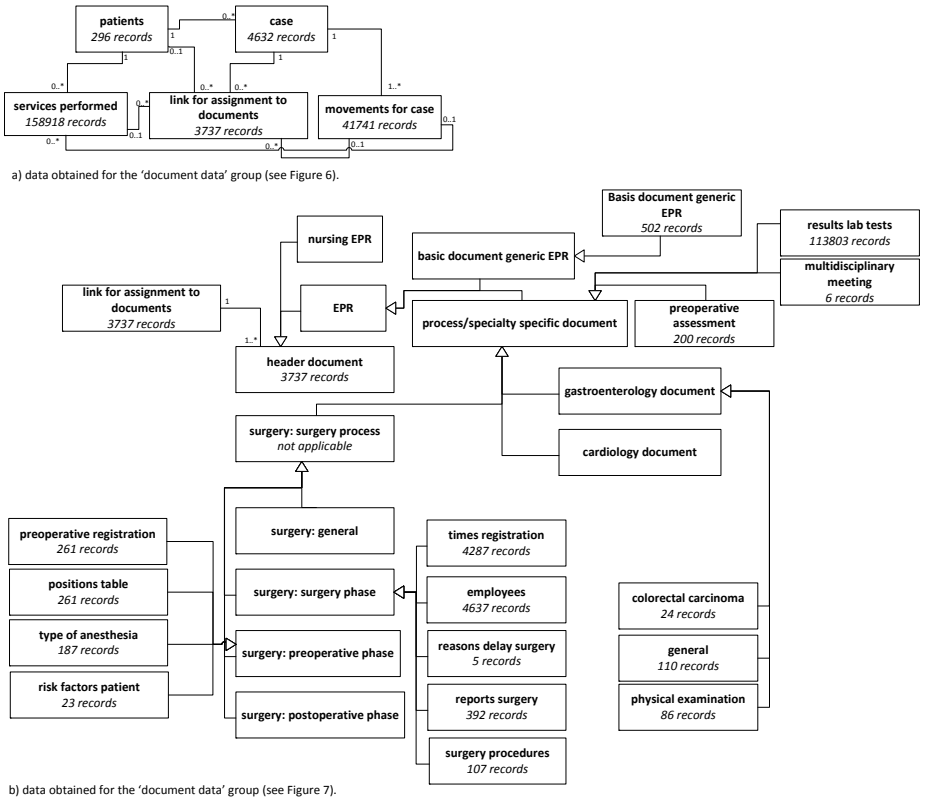


Figure 24: For the 'document data' group of the *healthcare reference model* it is shown for which classes data has been obtained from the *i.s.h.med* system of the MUMC.

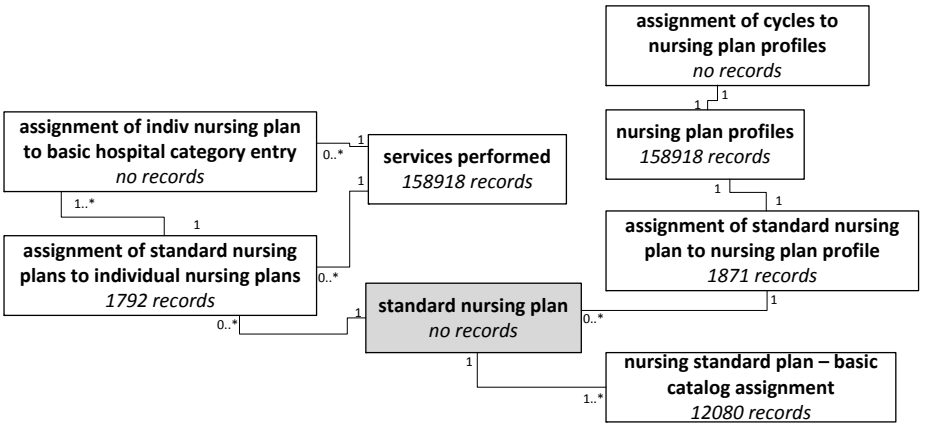


Figure 25: For the 'nursing plans' group of the *healthcare reference model* it is shown for which classes data has been obtained from the *i.s.h.med* system of the MUMC.

Annual Workshop on Information Technologies and Systems (WITS 2007), 2007, pp. 139–145.

[13] C. Günther, W. van der Aalst, Fuzzy Mining: Adaptive Process Simplifica-

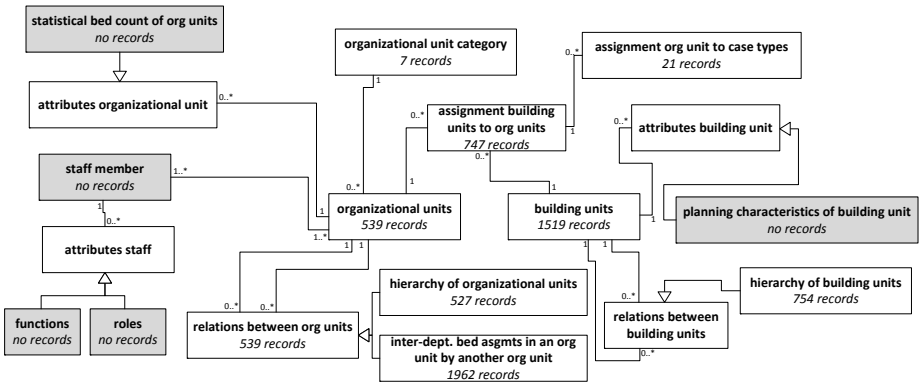


Figure 26: For the 'organization and buildings' group of the *healthcare reference model* it is shown for which classes data has been obtained from the *i.s.h.med* system of the MUMC.

tion Based on Multi-perspective Metrics, in: International Conference on Business Process Management (BPM 2007), Vol. 4714 of Lecture Notes in Computer Science, Springer-Verlag, Berlin, 2007, pp. 328–343.

[14] R. J. C. Bose, W. van der Aalst, Process Diagnostics using Trace Alignment: Opportunities, Issues, and Challenges, Information Systems 37 (2) (2012) 117–141.

[15] A. Rozinat, W. van der Aalst, Decision Mining in ProM, in: S. Dustdar, J. L. Fiadeiro, A. Sheth (Eds.), BPM 2006, Vol. 4102 of Lecture Notes in Computer Science, Springer-Verlag, Berlin, 2006, pp. 420–425.

[16] W. van der Aalst, H. de Beer, B. van Dongen, Process Mining and Verification of Properties: An Approach based on Temporal Logic, in: R. Meersman, Z. T. Et al. (Eds.), On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2005, Vol. 3760 of Lecture Notes in Computer Science, Springer-Verlag, Berlin, 2005, pp. 130–147.

[17] R. P. J. C. Bose, R. Mans, W. van der Aalst, Wanna Improve Process Mining Results? – It's High Time We Consider Data Quality Issues Seriously, BPM Center Report BPM-13-02, BPMcenter.org (2013).

[18] M. van Eck, Timestamps Within Healthcare Process Mining Logs, Master's thesis, Eindhoven University of Technology, Eindhoven (2013).

[19] O. Bott, A. Terstappen, D. Pretschner, Modeling ,Simulating, and Visualizing Information System Artifacts in Healthcare, http://www.umi.cs.tu-bs.de/full/research/mis/mosaikm/Mosaik-M_MI_New_mb.pdf.

[20] T. Wendt, A. Häber, B. Brigl, A. Winter, Modeling Hospital Information Systems (Part 2): Using the 3LGM2 Tool for Modeling Patient Record Management, Methods of Information in Medicine 43 (3) (2004) 256–67.

[21] A. Winter, B. Brigl, T. Wendt, Modeling Hospital Information Systems (Part 1): The Revised Three-Layer Graph-based Meta Model 3LGM2, Methods of Information in Medicine 42 (5) (2003) 544–51.

[22] O. Bott, J. Bergmann, I. Hoffmann, T. Vering, E. Gomez, M. Hernando, D. Pretschner, Analysis and Specification of Telemedical Systems Using Modelling and Simulation: the MOSAIK-M Approach, in: Connecting Medical Informatics and Bio-Informatics: Proceedings of MIE2005, Vol. 116 of Studies in Health Technology and Informatics, 2005, pp. 503–508.

[23] R. Haux, C. Seggewies, W. Baldauf-Sobez, P. Kullmann, H. Reichert, L. Luedecke, H. Seibold, Soarian–Workflow Management Applied for Health Care, Methods of Information in Medicine 42 (1) (2003) 25–36.

[24] P. Gell, P. Schmücker, M. Pedevilla, H. Leitner, H. Naumann, H. Fuchs, H. Pitz, W. Köle, SAP and Partners : IS-H and IS-H* MED, Methods of Information in Medicine 42 (1) (2003) 16–24.

[25] G. Gell, T. Gitter, Hospital Information System/Electronic Health Record (HIS/HER) and Clinical Research, in: P. Welfens, E. Walther-Klaus (Eds.), Digital Excellence, Springer Verlag Berlin-Heidelberg, 2008, pp. 137–146.

[26] P. Clayton, S. Narus, S. Huff, T. Pryor, P. Haug, T. Larkin, S. Matney, R. Evans, B. Rocha, W. Bowes, F. Holston, M. Gundersen, Building a Comprehensive Clinical Information System from Components. The Approach at Intermountain Health Care, Methods of Information in Medicine 42 (1) (2003) 1–7.

[27] R. Mans, M. Schonenberg, M. Song, W. van der Aalst, P. Bakker, Application of Process Mining in Healthcare : a Case Study in a Dutch Hospital, in: A. Fred, J. Filipe, H. Gamboa (Eds.), Biomedical engineering systems and technologies (International Joint Conference, BIOSTEC 2008, Funchal, Madeira, Portugal, January 28-31, 2008, Revised Selected Papers), Vol. 25 of Communications in Computer and Information Science, Springer-Verlag, Berlin, 2009, pp. 425–438.

[28] A. Rebuge, D. Ferreira, Business Process Analysis in Healthcare Environments: A Methodology Based on Process Mining, Information Systems 37 (2).

[29] J. Poelmans, G. Dedene, G. Verheyden, H. van der Mussele, S. Viaene, E. Peters, Combining Business Process and Data Discovery Techniques for Analyzing and Improving Integrated Care Pathways, in: P. Perner (Ed.), Advances In Data Mining Applications And Theoretical Aspects, Vol. 6171 of Lecture Notes in Computer Science, Springer-Verlag, Berlin, 2010, pp. 505–517.

[30] J. van de Klundert, P. Gorissen, S. Zeemering, Measuring Clinical Pathway Adherence, Journal of Biomedical Informatics 43 (6) (2010) 861–72.

[31] M. Grando, W. van der Aalst, R. Mans, Reusing a Declarative Specification to Check the Conformance of Different CIGs, in: Proceedings of the BPM 2011 Workshops, Vol. 100 of Lecture Notes in Business Information Processing, Springer Verlag Berlin-Heidelberg, 2012, pp. 188–199.

[32] R. Dunkl, K. Fröschl, W. Grossman, S. Rinderle-Ma, Assessing Medical Treatment Compliance Based on Formal Process Modeling, in:

A. Holzinger, K.-M. Simonic (Eds.), Proceedings of USAB 2011, Lecture Notes in Computer Science, Springer Verlag Berlin-Heidelberg, 2011, pp. 533–546.

[33] L. T. Ramos, Healthcare Process Analysis : Validation and Improvements of a Data-based Method using Process Mining and Visual Analytics, Master's thesis, Eindhoven University of Technology, Eindhoven (2009).

[34] R. Mans, Workflow Support for the Healthcare Domain, Ph.D. thesis, Eindhoven University of Technology (2011).

[35] J. de Weerdt, F. Caron, J. Vanthienen, B. Baesens, Getting a Grasp on Clinical Pathway Data : An Approach Based on Process Mining, in: T. Washio, J. Luo (Eds.), Emerging Trends in Knowledge Discovery and Data Mining, Vol. 7769 of Lecture Notes in Computer Science, Springer Verlag Berlin-Heidelberg, 2013, pp. 22–35.

[36] R. J. C. Bose, W. van der Aalst, Analysis of Patient Treatment Procedures, in: Proceedings of the BPM 2012 Workshops, Vol. 99 of Lecture Notes in Business Information Processing, Springer Verlag Berlin-Heidelberg, 2012, pp. 165–166.

[37] M. Song, C. Günther, W. van der Aalst, Trace Clustering in Process Mining, in: D. Ardagna, M. Mecella, J. Yang (Eds.), Business Process Management Workshops (BPM 2008), Vol. 17 of Lecture Notes in Business Information Processing, Springer-Verlag, Berlin, 2009, pp. 109–120.

[38] D. Ferreira, C. Alves, Discovering User Communities in Large Event Logs, in: F. Daniel (Ed.), Proceedings of the 2011 BPM Workshops, Vol. 99 of Lecture Notes in Business Information Processing, Springer Verlag Berlin-Heidelberg, 2012, pp. 123–134.

[39] M. Lang, T. Bürkle, S. Laumann, H.-U. Prokosch, Process Mining for Clinical Workflows: Challenges and Current Limitations, in: S. K. A. Et al. (Ed.), Proceedings of MIE 2008, Vol. 136 of Studies in Health Technology and Informatics, IOS Press, 2008, pp. 229–234.

[40] J. Zhou, Process mining : Acquiring Objective Process Information for Healthcare Process Management with the CRISP-DM Framework, Master's thesis, Eindhoven University of Technology, Eindhoven (2009).

[41] A. Manninen, Applying the Principles of Process Mining to Finnish Healthcare, Ph.D. thesis, Aalto University (2010).

[42] S. Gupta, Workflow and Process Mining in Healthcare, Master's thesis, Eindhoven University of Technology, Eindhoven (2007).

[43] R. Mans, M. Schonenberg, G. Leonardi, S. Panzarasa, S. Quaglini, van der Aalst, Process Mining Techniques : An Application to Stroke Care, in: S. K. A. Et al. (Ed.), eHealth Beyond the Horizon - Get IT There (Proceedings 21st International Congress of the European Federation for Medical Informatics, MIE 2008, Vol. 136 of Studies in Health Technology and Informatics, IOS Press, 2008, pp. 573–578.

[44] S. Quaglini, Process Mining in Healthcare : A Contribution to Change the Culture of Blame, in: D. Ardagna, M. Mecella, J. Yang (Eds.), Proceedings of the BPM 2009 workshops, Vol. 17 of Lecture Notes in Business Information Processing, Springer Verlag Berlin-Heidelberg, 2009, pp. 308–311.

[45] L. Perimal-Lewis, S. Qin, C. Thompson, P. Hakendorf, Gaining Insight from Patient Journey Data using a Process-Oriented Analysis Approach, in: K. Butler-Henderson, K. Gray (Eds.), HIKM 2012, Vol. 129 of Conferences in Research and Practice in Information Technology, Australian Computer Society, Inc., 2012, pp. 59–66.

[46] P. Riemers, Process Improvement in Healthcare : a Data-based Method using a Combination of Process Mining and Visual Analytics, Master's thesis, Eindhoven University of Technology, Eindhoven (2009).

[47] T. Blum, N. Padoy, H. Feuner, N. Navab, Workflow Mining for Visualization and Analysis of Surgeries, International Journal of Computer Assisted Radiology and Surgery 3 (2008) 379–386.

[48] U. Kaymak, R. Mans, T. V. D. Steeg, M. Dierks, On Process Mining in Health Care, in: 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE Computer Society Press, 2012, pp. 1859–1864. doi:10.1109/ICSMC.2012.6378009.

[49] R. Mans, H. Reijers, M. van Genuchten, D. Wismeijer, Mining Processes in Dentistry, in: Proceedings of the 2nd ACM SIGHIT symposium on International health informatics - IHI '12, ACM Press, New York, New York, USA, 2012, p. 379.

[50] M. Peleg, P. Soffer, J. Ghattas, Mining Process Execution and Outcomes Position Paper, in: A. ter Hofstede, B. Benatallah, H.-Y. Paik (Eds.), Proceedings of the BPM 2008 workshops, Vol. 4928 of Lecture Notes in Computer Science, Springer Verlag Berlin-Heidelberg, 2008, pp. 395–400.

[51] L. Maruster, R. Jorna, From Data to Knowledge: a Method for Modeling Hospital Logistic Processes, IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society 9 (2) (2005) 248–55.

[52] P. Helmering, P. Harrison, V. Iyer, A. Kabra, J. Slette, Process Mining of Clinical Workflows for Quality and Process Improvement, in: HIMMS2012, 2012, pp. 1–7.

[53] Z. Huang, X. Lu, H. Duan, On Mining Clinical Pathway Patterns from Medical Behaviors, Artificial Intelligence in Medicine 56 (1) (2012) 35–50. doi:10.1016/j.artmed.2012.06.002.

[54] C. McGregor, C. Catley, A. James, A Process Mining Driven Framework for Clinical Guideline Improvement in Critical Care, in: P. P. Rodrigues, M. Pechenizkiy, M. Gaber, J. Gama (Eds.), Proceedings of the Workshop on Learning from Medical Data Streams, Vol. 765 of CEUR Workshop Proceedings, 2011.

[55] T. Neumuth, P. Jannin, J. Schlomberg, J. Meixensberger, P. Wiedemann, O. Burgert, Analysis of Surgical Intervention Populations using Generic Surgical Process Models, International Journal of Computer Assisted Radiology and Surgery 6 (1) (2011) 59–71.

[56] H. Fei, N. Meskens, Discovering Patient Care Process Models From Event Logs, in: 8th International Conference of Modeling and Simulation - MOSIM10, 2010, pp. 10–12.

[57] J. Staal, Using Process and Data Improving Techniques to Define and Improve Standardization in a Healthcare Workflow Environment, Ph.D. thesis, Eindhoven University of Technology (2010).

[58] C. Günther, A. Rozinat, W. van der Aalst, K. van Uden, Monitoring Deployed Application Usage with Process Mining, BPM Center Report BPM-08-11, BPMcenter.org (2008).

[59] R. Pérez-Castillo, B. Weber, J. Pinggera, S. Zugal, I.-R. de Guzmán, M. Piattini, Generating Event Logs from Non-Process-aware Systems Enabling Business Process Mining, Enterprise Information Systems 5 (3) (2011) 301–335.

[60] B. Han, L. Jiang, H. Cai, Abnormal Process Instances Identification Method in Healthcare Environment, in: 2011IEEE 10th International Conference on Trust, Security and Privacy in Computing and Communications, IEEE, 2011, pp. 1387–1392.

[61] M. Kuo, Y. Chen, A Method to Identify the Difference between Two Process Models, Journal of Computers 7 (4) (2012) 998–1005.

[62] K. Kirchner, N. Herzberg, A. Rogge-solti, M. Weske, Embedding Conformance Checking in a Process Intelligence System in Hospital Environments, in: R. Lenz, S. Miksch, M. Peleg, M. Reichtert, D. Riano, A. ten Teije (Eds.), Proceedings of the BPM 2012 Workshops, Vol. 7738 of Lecture Notes in Computer Science, Springer Verlag Berlin-Heidelberg, 2013, pp. 126–139.

[63] M. Grando, M. Schonenberg, W. van der Aalst, Semantic-Based Conformance Checking of Computer Interpretable Medical Guidelines, in: F. Daniel, K. Barkaoui, S. Dustdar (Eds.), Proceedings of the BPM 2012 Workshops, Vol. 273 of Communications in Computer and Information Science, Springer Verlag Berlin-Heidelberg, 2013, pp. 285–300.

[64] M. Binder, W. Dorda, G. Duftschmid, R. Dunkl, K. Fröschl, M. Hronsky, S. Rinderle-Ma, On Analyzing Process Compliance in Skin Cancer Treatment : An Experience Report from the Evidence-Based Medical Compliance Cluster (EBMC2), in: J. Ralyté, X. Franch, S. Brinkkemper, S. Wrycza (Eds.), Proceedings of CAiSE 2012, Vol. 7328 of Lecture Notes in Computer Science, Springer Verlag Berlin-Heidelberg, 2012, pp. 398–413.

[65] L. Bouarfa, J. Dankelman, Workflow Mining and Outlier Detection from Clinical Activity Logs, Journal of Biomedical Informatics 45 (6) (2012) 1185–90.

[66] R. Mans, M. Schonenberg, G. Leonardi, S. Panzarasa, S. Quaglini, W. van der Aalst, Process Mining Techniques : An Application to Stroke Care, in: Proceedings of MIE 2008, Vol. 136 of Studies in Health Technology and Informatics, IOS Press, 2008, pp. 573–578.

[67] M. Lang, T. Bürkle, S. Laumann, H.-U. Prokosch, Process Mining for Clinical Workflows: Challenges and Current Limitations, in: Proceedings of MIE 2008, Vol. 136 of Studies in Health Technology and Informatics, IOS Press, 2008, pp. 229–234.