# TOWARDS STANDARDS BASED HEALTH DATA EXTRACTION FACILITATING PROCESS MINING

**Emmanuel Helm[(a)], Barbara Franz[(b)], Andreas Schuler[(c)], Oliver Krauss[(d)], Josef Küng[(e)]**

[(a), (c), (d),] Research Department of e-Health, Integrated Care,
University of Applied Sciences Upper Austria, 4232 Hagenberg, Austria
[(b)] Department of Software Engineering,
University of Applied Sciences Upper Austria, 4232 Hagenberg, Austria
[(e)] Institute for Applied Knowledge Processing
Johannes Kepler University, 4040 Linz, Austria

[(a), (b), (c), (d),] [emmanuel.helm | barbara.franz | andreas.schuler | oliver.krauss]@fh-hagenberg.at
[(e)]jkueng@faw.jku.at

## ABSTRACT

As an evidence based business process analysis method, process mining can be used to investigate variations in clinical practice and delivery of care. However, to enable cross-organizational comparative analysis, healthcare institutions need a common ground for the description and representation of health data. In this work, we analyze different approaches, to describe clinical and patient pathways. The Healthcare Reference Model represents a bottom-up approach, the HL7 v3 RIM as a generic health information model represents a top-down approach and HL7 FHIR, the newest standard of the HL7 family stands in-between. We highlight similarities and differences according to interoperability and process mining tasks. We conclude that a standards (RIM) based top-down approach, and the derived FHIR approach respectively, is able to provide similar insights and, on top of that, operational support for the ETL process on all interoperability levels.

Keywords: *Process Mining, Data Extraction, Semantic Interoperability, Evidence Based Medicine*

## 1. INTRODUCTION

Clinical pathways are management tools used to define the best process in a healthcare organization, using the best procedures and timing, to treat patients with specific diagnoses or conditions according to evidence-based medicine (Panella, Marchisio and Di Stanislao 2003).

For the development of clinical pathways and medical guidelines a comparative analysis of the existing approaches is useful. (Partington et al. 2015) propose the application of process mining as *an evidence-based business process analysis method* to investigate variations in clinical practice and delivery of care across different hospital settings.

However, existing approaches to use process mining for comparative analysis of healthcare processes are based on data sources within one organization. More precisely the formal representation and the semantics (including code systems and value sets) of the different data sources were basically the same (Partington et al. 2015; Mans et al. 2008). To gain insight into and enable comparative analysis of clinical practice and delivery of care across different organizations, a preceding step to identify and reach *common ground* is necessary.

In the recent book *Process Mining in Healthcare* the authors describe a *healthcare reference model* (HRM) that aims to help locating the needed data in healthcare information systems and thus facilitate the data extraction for process mining (Mans, van der Aalst and Vanwersch 2015). This data model can be seen as the *common ground*, and in their *Use Case 5: Healthcare Process Comparison* the authors also use it to compare processes of two different hospitals.

### 1.1. Prerequisites for Process Mining

Process mining algorithms work on event logs with a certain structure. Event logs must contain only data related to a single process and it must be ensured that all events in the log can be related to this process. Moreover, each event in the log must represent an activity and refer to a single process instance (case). To get the data out of the (distributed) data sources and to put them in this structure, preprocessing steps are necessary.

The Extract, Transform and Load (ETL) steps preceding the actual process mining tasks describe: (a) extraction data from outside sources, (b) transforming it to fit operational needs (dealing with syntactical and semantical issues while ensuring predefined quality levels), and (c) loading it into the target system, e.g. a data warehouse or relational database (Van der Aalst 2011).

### 1.2. Structure of this Work

In the following sections, this paper presents a different approach to reach common ground based on established healthcare interoperability standards. In section 2,

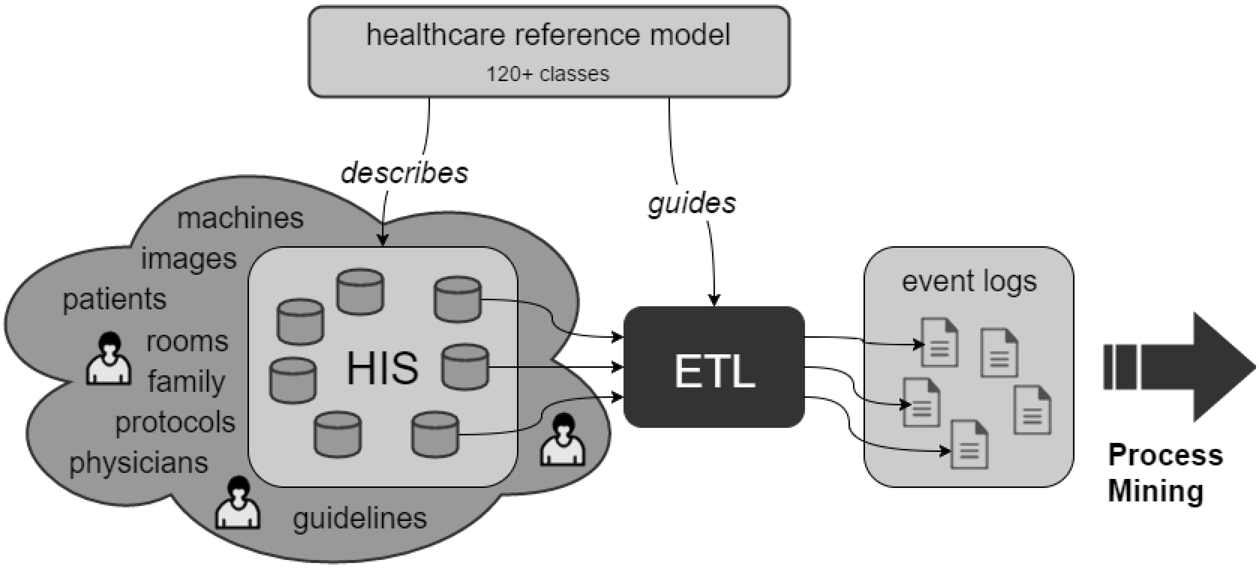

Figure 1: The healthcare reference model in context of process mining. It is used as a starting point for locating the data and extracting event logs, guiding the *Extract, Transform and Load* process (Mans, van der Aalst and Vanwersch 2015).

Background, the referenced standards and their applications in data modelling as well as the development approach of the HRM are described. Section 3, Methods, defines the analytical approach used in this research and explains the interoperability levels used to compare the various approaches. Section 4, Results and Discussion, describes the analyzation results of the various approaches with focus on the ability of the models to describe patient pathways. The results are compared results by their ability to integrate data from different sources. Section 5, Conclusion, summarizes important findings and emphasizes the additional options of a standards-based approach.

## 2. BACKGROUND

This section presents interoperability levels necessary for integrated healthcare and the modelling approaches HRM, RIM and FHIR, which are analyzed in this paper.

### 2.1. Levels of Interoperability

National as well as international initiatives for information integration in healthcare aim at the increase of interoperability of information systems and at minimizing integration efforts (Norgall 2003, Sunyaev et al. 2008). The term *interoperability* denotes the ability of systems to collaborate. Combining the definitions of (Heitmann and Gobrecht 2009) and (Serrano et al. 2015) we can defer three levels of interoperability, which incorporate, among others, the following characteristics:

- Exchange of meaningful, actionable information between two or more systems across organizational boundaries (*technical interoperability*)
- A shared understanding of the exchanged information (*semantic interoperability*)
- An agreed expectation for the response to the information exchange and requisite quality (reliability, fidelity, security) of services and processes (*process interoperability*).

Integrated care and the achievement of high quality healthcare over institutional borders require all three levels of interoperability between different healthcare providers.

### 2.2. Healthcare Reference Model

The HRM was specifically designed for the healthcare domain to guide the ETL steps and to help locating the data needed for process mining. Figure 1 shows the role of the HRM in the data preparation.

It was developed using a two-step approach as described in (Mans, van der Aalst and Vanwersch 2015). First, the data model of a running *i.s.h.med* hospital information system (HIS) was reverse engineered based on the actual database table structure, expert interviews and hands-on inspection. Secondly, the model was validated via interviews with HIS professionals of other hospitals.

This resulted in a model described in terms of a UML class diagram comprising 122 classes. The classes are grouped in several sub-models such as *General Patient and Case Data*, *Radiology* and *Document Data*.

The developers of the HRM do not claim completeness of the model since HISs of different vendors may contain data not present in the model. However the key elements needed for process mining should be included (Mans, van der Aalst and Vanwersch 2015).

### 2.3. Reference Information Model

HL7 standards have been specifically developed for the health sector. They define the exchange of messages, document based communications as well as cooperating services, their implementation and necessary infrastructural services (Benson and Grieves 2016).

Core of HL7 standards is the Reference Information Model (RIM), which is a generic healthcare specific information model. The base of this model are four core classes (Act, Entity, Role and Participation) and two additional classes (ActRelationship and RoleLink), as shown in Figure 2.

The goal is the development of a uniform understanding of objects and processes in the healthcare environment. The use of RIM provides specifications to structure, type, content as well as semantics, used vocabulary and underlying processes necessary for data transfer and interoperability, following a top-down approach. Well established standards like document-based exchange standard HL7 Clinical Document Architecture (CDA), are based on Refined Message Information Models (R-MIM), which are derived from the HL7 RIM.

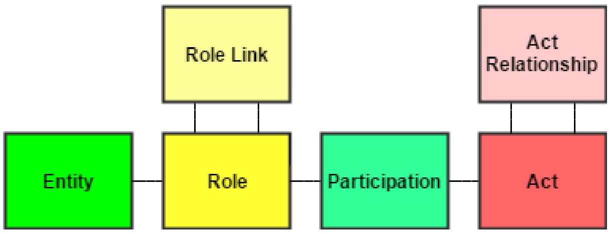

Figure 2: The backbone of the RIM with the main classes *Act*, *Role* and *Entity* and the association classes *Participation*, *Role Link* and *Act Relationship* (Benson and Grieves 2016).

For the representation of clinical and patient pathways HL7 proposes the care plan model, for which a layered modelling approach was applied to allow separation of business, information and interoperability requirements. This is achieved by separating the Care Plan model in three distinctive layers (HL7 2016):

#### 2.3.1. Care Plan Conceptual Model

The conceptual model identifies necessary concepts and the relationships between them. These are directly derived from high level business requirements and thus build the foundation for the consecutive layers (HL7 2016). More specifically the conceptual model consists of an abstract concept *Plan* that is associated with further concepts, e.g. *Care Giver*, *Patient*, *Provider*, *Activity*.

#### 2.3.2. Care Plan Organizing Framework for Coordination of Care Models

According to Health Level 7 (HL7 2016), the Care Plan Organizing Framework is a meta-model for coordination of care interaction and collaboration. Thus, the model defines relationships between a subset of the concepts defined in the conceptual model.

#### 2.3.3. Care Plan Logical Information Model

The final layer in the Care Plan Model adds data properties to the concepts defined in the predeceasing layers. This further allows to capture information relevant for dynamic coordination of care interactions and point in time data exchange (HL7 2016).

The resulting care plan structure is thus applicable in a wide range of scenarios and use cases, consisting of discipline- or treatment -specific plans as well as comprehensive multidisciplinary plans, e.g. in case of tumor board review meeting (HL7 2016).

An implementation of the Care Plan Model defined by HL7 is currently developed as part of HL7s standard Fast Healthcare Interoperability Resources and the *CarePlan* resource respectively.

### 2.4. Fast Healthcare Interoperability Resources

Fast Healthcare Interoperability Resources (*FHIR*) is a resource-based data exchange standard for healthcare information. FHIR Resources contain mapping information to the HL7 RIM, defining which fields of a resource correspond to which RIM concept.

FHIR is organized in different levels building upon each other:

- Level 1 – The basis of the standard, such as the API description, Data Types and Data Formats.
- Level 2 – Security and Implementer information as well as Terminology Bindings and Conformance Resources such as Structure Definitions describing FHIR Resources.
- Level 3 – Contains the basis for real world use-cases such as the Patient and Practitioners.
- Level 4 – Deals with data exchange in healthcare including clinical, diagnostic, medical, financial and workflow data.
- Level 5 – Describes clinical reasoning and contains resources enabling automation in that sector.

FHIR follows the Pareto Principle, better known as the 80/20 rule, and thus defines only the data exchange information 80% of identified use cases require. Specializations for the other 20% that may be required, can be modeled by implementers using *Extensions* and *Profiles* (Benson and Grieves 2016).

FHIR resources can be viewed in different formats, such as a Structural View, UML, XML, JSON and Turtle for understanding and rapid prototyping. All resources, which are currently defined in HL7 FHIR, are listed in the current STU 3 version (HL7 2017). FHIR is being updated regularly, with the next release planned in 2018.

## 3. METHODS

This section describes how the HRM and the partially RIM-based HL7 FHIR are used to model clinical and patient pathways. Furthermore, it describes how the various concepts are analyzed using the interoperability levels.

### 3.1. Using HRM to Model Pathways

For the HRM the definition and execution of pathways is described using 10 UML classes (see Figure 3). A pathway (*pathway*) consists of multiple items (*pathway item*) which may be connected to each other (*connection*). A pathway is executed for a patient (*patient pathway*) and information about each performed step is recorded (*step of patient pathway*). Finally, each performed step may be linked to a service that is executed for the patient (*services performed*) (Mans, van der Aalst and Vanwersch 2015).

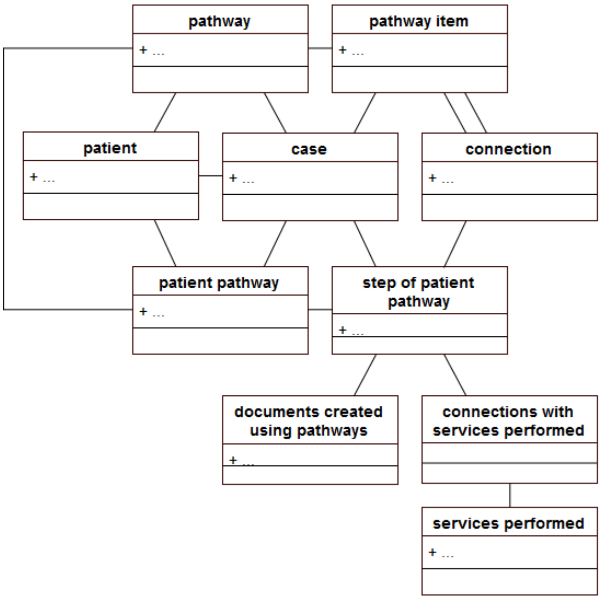

Figure 3: The UML classes used to describe clinical pathways in the HRM. The original figure in (Mans, van der Aalst and Vanwersch 2015, p49) also contained the ~150 attributes of those classes and provides cardinalities for the relations.

## 3.2. Using HL7 FHIR to Model Pathways

Using FHIR, the definition of an abstract clinical pathway is modelled with the *PlanDefinition* resource. This resources describes a goal that shall be achieved and defines a set of actions to explain what has to be done stepwise to achieve that goal. Actions have a timing when they can/should occur, as well as trigger definitions and conditions if and when they are applicable. Furthermore, each action has participants, explaining who is involved in the action.

Highly complex clinical pathways can also be described with the *PlanDefinition* as each action can have relations to other actions (HL7 2017). The *PlanDefinition* represents an abstract concept of what should happen in a medical pathway. It does not relate to a specific context such as actual patients or groups of persons.

### 3.2.1. Modeling Context-specific Pathways

To specify a *PlanDefinition* with a context, such as one real patient to be treated, the *CarePlan* resource can be used. The CarePlan is often based on a *PlanDefinition* (as seen in Figure 4), however it is allowed to modify the steps in the *PlanDefinition* as needed for treatment of the patient. In addition to representing a specific plan for a patient, or group of patients, it also, indirectly, documents what actually happens during execution of the *CarePlan* and accompanying treatment of the *CarePlan*. Similarily to the *PlanDefinition* a *CarePlan* contains a list of steps, here called activity, which describe activities to be performed. Each activity can have a template of what is going to happen, in the form of resource drafts, or alternatively contains the resource that documents what happened, such as a *MedicationRequest* documenting the administration of a medication, or an *Appointment* for the next treatment. Unlike the *PlanDefinition*, the steps in a *CarePlan* have no relation to each other (HL7 2017).

### 3.2.1. Using Security Details for Modeling

In addition to the *CarePlan*, which documents steps on a higher level, FHIR contains specifications for security auditing which can be used for a fine-grained view of a taken medical pathway. The *AuditEvent* resource documents every single access to a FHIR resource according to the Five Ws (Who, What, When, Where, Why). Additionally, the *Provenance* resource tracks similar information when a resource is being modified (HL7 2017).

## 3.3. Analyzation Based on Interoperability Levels

Using pre-defined use cases and based on the interoperability levels described in section 2.4, the HRM approach and the FHIR approach are analyzed how they add semantic information to the models and whether they support the ETL process. To achieve high quality healthcare over institutional borders, all three levels of interoperability defined in section 2.4, are needed between different healthcare providers.

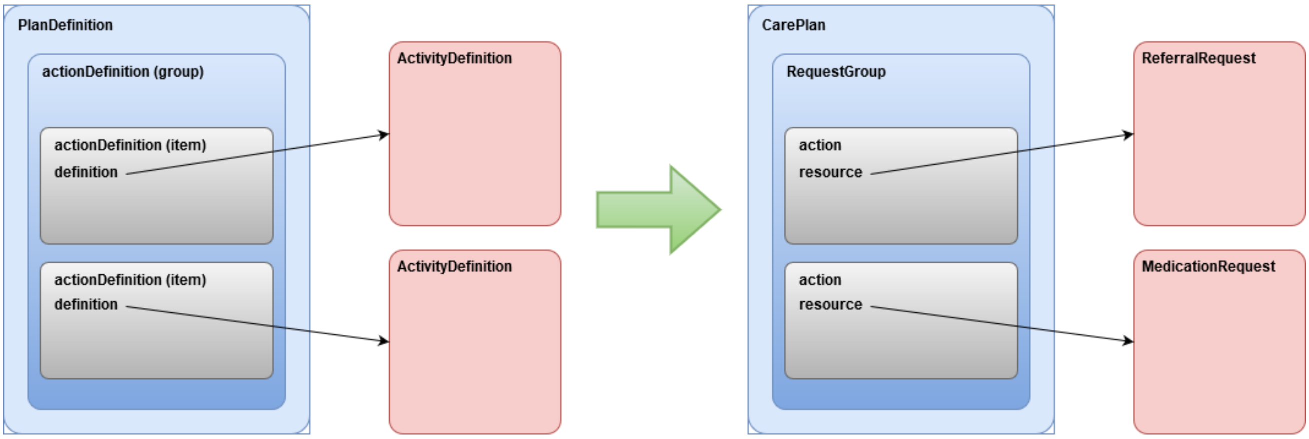

Figure 4: *PlanDefinition* groupings of actions and definitions of these actions in relation to the *CarePlan* groupings and actions. As can be seen, a *PlanDefinition* contains only abstract definitions of what should happen, while the *CarePlan* contains references to activities relating to a patient such as a *Referral* or *Medication* (HL7 2017).

Since technical interoperability has been the focus of standards organizations, alliances and consortia for many years, standards and implementations supporting this level of interoperability are generally available. Strategies for informational interoperability, however, which includes the whole area of semantic and process interoperability, are less mature (Serrano et al. 2015). Thus, we focus on these types of interoperability for the comparison of the various modeling approaches.

## 4. RESULTS AND DISCUSSION

Based on existing use cases and using the methods described in section 3, the HRM approach and the FHIR approach are analyzed concerning semantic interoperability, especially how they add semantic information to the models, and process interoperability with focus on whether they support the ETL process.

### 4.1. Analyzation of the HRM

The analyzation of the HRM is based on six use cases listed by (Mans, van der Aalst and Vanwersch 2015), where data described via the HRM was gathered from hospitals and used for process mining tasks like model discovery, conformance checking, bottleneck analysis and comparative analysis.

The HRM is instrumental in locating the data needed and facilitating the actual extraction. It supports the ETL process by giving the analysts an idea about what kind of process-related data can be found in a HIS.

The abstract HRM does not directly refer to the database structure and thus does not support the operation of *Extracting* of data (Mans, van der Aalst and Vanwersch 2015). Furthermore, the HRM model does not explicitly refer to code lists, structured value sets or nomenclatures. Attributes listed in the UML classes are identified by name, the semantics for those attributes are not provided in the textual description of the models (Mans, van der Aalst and Vanwersch 2015).

### 4.2. Analyzation of the HL7 FHIR

The FHIR approach was analyzed using a system for Multidisciplinary Team Meetings (MDTM) as described in (Krauss et al. 2017). MDTM are modeled in HL7 FHIR using the *PlanDefinition* resource. The models are then automatically transformed into the Business Process Model Notation (BPMN) and executed with a Workflow Engine. The process is documented using the *CarePlan* resources as well as FHIR *Auditing*. An ETL process, as described in the preceding section, could generate XES Event Logs, subsequently enabling Process Mining.

FHIR implicitly enables the entire ETL process. The *Extraction* from a data source is a given for any healthcare system implementing the FHIR standard, as the FHIR API allows reads and searches on resources.

The *Transformation* process can be executed directly in FHIR. In addition to the *ConceptMap* which allows semantic transformations, the *StructureMap* allows the mapping of any FHIR Resource to (or from) any other *Concept* defined by a *StructureDefinition*. This allows the implementer the definition of *StructureDefinitions* tailored to the exact requirements of the process-mining target. From there the *StructureMap* can describe how a regular FHIR resource can be transformed into the required structure – e.g. the eXtensible Event Stream (XES) used in recent process mining tools (Verbeek et al. 2010). Since the *StructureMap* contains machine-executable rules for transformation this mapping can be done fully-automated in a standardized way.

The *Loading* of the transformed resources can be achieved in several different ways. For one the FHIR API once again allows the access to the transformed resources similar to the *Extraction* process. Alternatively the data-mining application can use the FHIR *Subscription* service to receive push messages while the process is executed. In another push-based approach one can create a FHIR *Operation* that publishes the transformed resources directly to the data mining application.

Machine readable semantics of FHIR resources are handled through *Codings* that represent fields of resources, similar to *Codings* in the HL7 RIM. Each *Coding* consists of a *System* that defines the Code, the actual *Code* for machine-readability, and a *DisplayName* for human-readability. HL7 FHIR uses LOINC, SNOMED CT, HL7v3, ICD-10, and DICOM among others (HL7 2017).

FHIR enables the restriction of allowed values in a *Coding*, and often provides default sets, or enforces usage of a specific set. This is documented in the *ValueSet* resource which groups *Codings* into a set. To enable semantic interoperability between different *Codings* the *ConceptMap* Resource exists to define unidirectional mappings between *ValueSets*. In addition, FHIR defines a *Terminology Service* specification that uses these resources to enable usage and transformation between *ValueSets* (HL7 2017).

### 4.3. Comparison of the interoperability status

The described models allow the interaction of technical components and systems. Further, they enable a larger interconnected system capability that transcends the local perspective of each participating subsystem, which complies to the technical interoperability defined by (Serrano et al. 2015).

Since the HRM model does not explicitly refer to code lists and semantic details are neither provided in the model definition nor description, semantic interoperability can hardly be achieved in an automated way using the HRM.

Using FHIR, semantic interoperability is achieved through common information models and the terminology service, thus enabling process definitions in a certain domain as well as across various domains or communities. Besides technical standards, agreements are essential, how medical and domain specific terminologies are used, which have to be maintained and further developed over time. This way it is possible to relate pathways across various healthcare institutions, while preserving the intended meaning, which conforms to the definition of semantic interoperability in (Heitmann and Gobrecht 2009).

The FHIR resources described in section 3.2 allow the definition (*PlanDefinition*, *CarePlan*) as well as the documentation (coarse granularity: *CarePlan*; fine granularity: *AuditEvent*, *Provenance*) of a medical pathway. This also allows a comparison between what should happen versus what actually did happen, which can be used to further process interoperability as well as check, compare and evaluate clinical pathways in and across institutions.

## 5. CONCLUSION

We compared the HRM and FHIR approaches on a conceptual basis and analyzed their ability to describe clinical and patient pathways. Furthermore, we analyzed how the models and their design approaches support the ETL process to prepare data from different sources for subsequent Process Mining. By analyzing the different approaches and their implications regarding the support of the ETL process, the use cases for the models become apparent. In the Epilogue of (Mans, van der Aalst and Vanwersch 2015) the authors highlight the importance of the HRM to *reason about questions that may or may not be answered using process mining*. Moreover the HRM produces *awareness* of the data present in a hospital.

We conclude that a standards (RIM) based top-down approach, and the derived FHIR approach respectively, is able to provide similar insights and, on top of that, operational support for the ETL process on all interoperability levels. As described above, the main reason for that is the ability to (automatically) integrate data from different sources by taking their semantic properties into account.

Further research on the implementation of a FHIR-based ETL process is necessary. The authors plan a case study with data from the MDTM system described in (Krauss et al. 2017).

## REFERENCES

Benson, T. and Grieve, G., 2016. Principles of Health Interoperability: SNOMED CT, HL7 and FHIR, Springer International Publishing, pp. 243-264

Chelsom, J. et al., 2015. Document-Driven Care Pathways Using HL7 CDA. In: eTELEMED 2015: The Seventh International Conference on eHeatlh, Telemedicine, and Social Medicine, Lisbon, Portugal.

Heitmann, K. U. and Gobrecht K., 2009. HL7 Communincation standards for the healthcare system (in German). Köln: LUP AG Lithographie & Printproduktion.

HL7, 2017. FHIR Release 3 (STU) Resource Index. Available from: http://hl7.org/fhir/resourcelist.html [accessed 14 April 2017]

HL7, 2016. HL7 Version 3 Domain Analysis Model: Care Plan, Release 1, HL7 Informative Document

Krauss O., Holzer K., Schuler A., Egelkraut R., and Franz B., 2017. Challenges and approaches to make multidisciplinary team meetings interoperable – the KIMBo project. Stud Health Technol. Inform. 2017, 236, pp. 63-69.

Mans, R., van der Aalst, W. and Vanwersch, R. J., 2015. Process Mining in Healthcare: evaluating and exploiting operational healthcare processes. Heidelberg Springer.

Mans, R., Schonenberg, H., Leonardi, G., Panzarasa, S., Cavallini, A., Quaglini, S. and van der Aalst, W. 2008. Process mining techniques: an application to stroke care. Studies in health technology and informatics, 136, 573.

Norgall, T., 2003. Communication standard for health telematics: status and outlook (in German), in Proceedings e-Health 2003.

Panella, M., Marchisio, S. and Di Stanislao, F., 2003. Reducing clinical variations with clinical pathways: do pathways work? In: International Journal for Quality in Health Care, 15(6), 509-521.

Partington, A., Wynn, M., Suriadi, S. and Ouyang, C., 2015. Process mining for clinical processes: a comparative analysis of four Australian hospitals. In: ACM Transactions on Management Information Systems (TMIS), 5(4), 19

Serrano, M. et al., 2015. Semantic Interoperability Release 2.0, AIOTI WG03 - IoT Standardisation, Alliance for Internet of Things Innovation.

Sunyaev, A., Schweiger, A., Leimeister, L.M. and Krcmar, H., 2008. Software Agents for integration of information systems in healthcare (in German), in Proceedings of e-Health 2008, pp. 1455-1466.

Van der Aalst, W. M., 2011. Getting the Data. In: Process Mining, Springer Berlin Heidelberg, 95-123.

Verbeek, H. M. W., Buijs, J. C., Van Dongen, B. F. and van Der Aalst, W. M., 2010. XES, XESame, and ProM 6. In: CAiSE Forum Vol. 72, pp. 60-75