On Analyzing Process Compliance in Skin Cancer Treatment: An Experience Report from the Evidence-Based Medical Compliance Cluster (EBMC^2)*

Michael Binder[1], Wolfgang Dorda[2], Georg Duftschmid[2], Reinhold Dunkl[3], Karl Anton Fröschl[3], Walter Gall[2], Wilfried Grossmann[3], Kaan Harmankaya[1], Milan Hronsky[2], Stefanie Rinderle-Ma[3], Christoph Rinner[2], and Stefanie Weber[1]

[1] Department of Dermatology, Medical University of Vienna, Austria
firstname.lastname@meduniwien.ac.at
[2] Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Austria
firstname.lastname@meduniwien.ac.at
[3] Faculty of Computer Science, University of Vienna, Austria
firstname.[midname.]lastname@univie.ac.at

Abstract. Process mining has proven itself as a promising analysis technique for processes in the health care domain. The goal of the EBMC^2 project is to analyze skin cancer treatment processes regarding their compliance with relevant guidelines. For this, first of all, the actual treatment processes have to be discovered from the available data sources. In general, the L* life cycle model has been suggested as structured methodology for process mining projects. In this experience paper, we describe the challenges and lessons learned when realizing the L* life cycle model in the EBMC^2 context. Specifically, we provide and discuss different approaches to empower data of low maturity levels, i.e., data that is not already available in temporally ordered event logs, including a prototype for structured data acquisition. Further, first results on how process mining techniques can be utilized for data screening are presented.

Keywords: Data Quality, Healthcare Processes, Process Modeling, Process Mining.

1 Introduction

The L* life cycle model – defined in the Process Mining Manifesto [2] – presents a structured modeling approach for monitoring and analyzing real world processes in general. Application of this proposal to the domain of health care is a challenging task due to the following reasons. First of all, because of quality

* The work presented in this paper was conducted in the context of EBMC^2 that is co-funded by the University of Vienna and the Medical University of Vienna.

J. Ralyté et al. (Eds.): CAiSE 2012, LNCS 7328, pp. 398–413, 2012.

concerns, event-logs in medical treatment do not fit easily into the hierarchy of maturity levels provided in the Manifesto. Rather, activities in the management of electronic health records, including data interchange, indicate that current practice still is a long way from transactional data management in, for example, business applications featuring already a fairly automatized and standardized tracking of business processes. Moreover, the specification of medical treatment process models is generally expressed in terms of high-level *medical guidelines* which is to say that, again, the accomplished level of standardization and formalization is fairly weak as compared to typical business applications. Many and even critical process details lack concise specification, much room is left for tacit, or uncodified, medical expert knowledge, and often important diagnostic or therapeutic decisions rest on patient conditions stated informally only. Since many treatment aspects and medical interventions are still questions of vivid research, this is no different in the particular case of skin cancer (melanoma) treatment focused upon in the subsequent discussion.

In spite of the various sources of imprecision and formal under-specification of treatment processes, this domain of consideration has been chosen deliberately because, on the one hand, there is a lot of current interest in assessing and improving clinical melanoma treatment routines while, on the other hand, growing incidence rates – combined with steeply rising costs of ever more efficient therapy options – suggest a significantly increasing relevance of effective as possible skin cancer treatment regimes. Thus, in addition to improvements in individual treatment episodes, treatment process analyses are also of particular interest in terms of population health and health economy, that is: resource allocation.

Methodologically, the issue obviously provides a fertile field of interdisciplinary research in medicine, data management, and epidemiology, seeking to improve the practice of medical decision making by use particularly of *evidence-based analysis* of data originating from empirical treatment processes. In this experience report, we describe first steps and results of an on-going effort to model medical care processes in the domain of skin cancer treatment as a prerequisite to the application of process mining techniques. Accordingly, Section 2 introduces the methodological set-up envisaged, trying to interrelate formal process analysis with relevant problem perspectives, notably evidence-based medicine and epidemiology. Next, Section 3 describes and evaluates data sources at hand. Section 4, then, presents a *temporal* data model in favor of integration of different data sources. On top of this, Section 5 provides preliminary evidence on how results of process mining may effectively support medical treatment practices. Finally, Section 6 summarizes the state of development, and concludes with an outlook on further activities of the project consortium.

2 The Treatment Compliance Analysis Framework

The ambition of analyzing empirical conformity of medical care processes to treatment standards defined by medical guidelines requires a linkage of expertise in fields of science as diverse as business process engineering, formal process modeling, medical information management, data analytics and statistical modeling,

and, last but not the least, medical knowledge and experience in skin cancer diagnosis and therapy. Apparently, such an integrated effort of analyzing evidence-based medical practice towards a valid interpretation according to statistically defined health determinants over whole patient populations calls for a cooperation of quite a range of disciplines. To this end, the Evidence Based Medicine Compliance Cluster (EBMC^2, `http://ebmc2.univie.ac.at/`) has been instigated as a co-funded collaborative project of the Medical University of Vienna together with the University of Vienna, where the former contributes expertise predominantly in medical informatics and dermatology while the latter is in charge of process management knowledge and statistical expertise.

To lay the foundation for further discussion, this section maps, in brief, how formal process modeling techniques may help in assessing medical treatment compliance by way of identifying and representing, respectively, different empirical variants of treatment processes – some compliant to defined guidelines, others possibly not – from meshed-up medical data of available sources.

2.1 The Basic Analytic Setting and Related Work

The basic setting of medical treatment compliance assessment rests on research in computer science, medicine, statistics, and the interplay between these areas. From a computer science viewpoint, the most important topics to be mentioned certainly are: process modelling, process mining, and information systems. In particular, the activities of the IEEE task force on process mining (notably, its Process Mining Manifesto [2]), recent developments in the area of standardization of event-log formats (cf. www.xes-standard.org) as well as the seminal process mining book [1] constitute basic theoretical pillars for project work in EBMC^2.

As to medicine, the project considers medical guidelines (cf. GRADE [11]) as its most important source of input. Medical guidelines integrate bits and pieces of medical evidence studies into coherent treatment processes and, in so doing, provide assistance to health professionals towards effective, good-quality medical practice. However, one has to keep in mind that besides such codified medical knowledge there exists a considerable amount of *informal*, and thus hard to come by, knowledge based on the long-term expertise of physicians. Through observation of actual treatment paths the inferred physicians' diagnostic and therapeutic choices are practically filling in the top-level advice of guidelines.

As a first use case in the project, the guidelines for treating patients with cutaneous melanoma have been selected. Skin malignancies are generally recognized as global major health problem among Caucasian populations, all the more so as incidence rates keep rising.

Currently, already a number of proposals combines ideas from workflow management with medical guidelines; for example, modeling languages such as GLIF, Asbru, EON/SAGE, PROforma, GUIDE, or PRODIGY, all of which are fairly close to traditional workflow languages [14]. From the field of business process engineering, BPMN (Business Process Modeling Notation) represents the state-of-the-art, and a direct comparison between BPMN and PROforma confirms

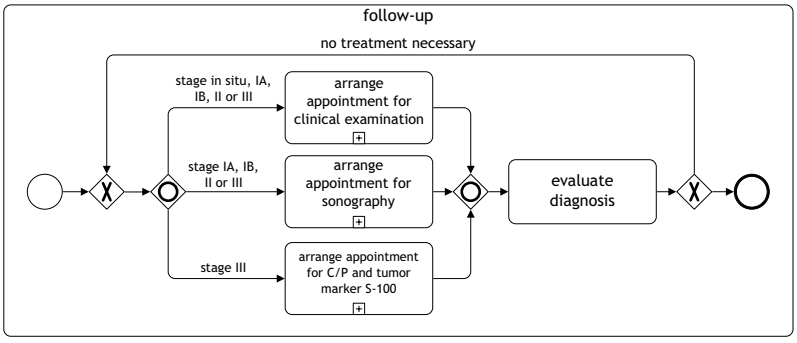

Fig. 1. Guideline Stage: Follow-up Subprocess

that BPMN can indeed cope quite well with requirements of the health care domain [9]. Hence, EBMC^2 decided to favor BPMN as modeling language for re-expressing formally the mostly verbal medical guideline.

For the sake of illustration, Figure 1 shows part of the European guideline on melanoma treatment [10] modeled in BPMN. The exhibit depicts abstractly the aftercare process highlighting follow-up examinations of melanoma patients, that is, patients are coerced to arrange appointments for regularly spaced follow-up examinations depending on their respective *stage* of melanoma status as defined in the cited guideline. In the example shown, stage I to III melanoma patients are covered, while stage IV patients – who already suffer from distant metastases – are not included in this scheme as, because of the severe progress of their medical condition, all due examinations are arranged individually in order to locate any further metastases as quickly as possible.

It is little surprise that empirical data about treatment processes are often structured simply according to the specific purpose at hand. Thus, many times the guiding principle of data organization involves barely more than a crude template for diagnostic findings. Certainly, those ad hoc data approaches badly cover any unprecedented later data usage, encouraging diverse efforts in medical informatics to remedy this obvious shortcoming: (i) a terminology-oriented approach, giving rise to medical ontologies like SNOMED [18]; (ii) a data exchange-oriented approach like the HL7 model [12]; and (iii) an archetype-oriented approach, most typical of which probably being the Open-EHR initiative [19]. To EBMC^2, all of these ideas are of utmost importance as a means of formally re-expressing (medical) information and (treatment) knowledge: based on such formal representations, models for the integration of patient data of different origin can be arranged to carry data transformations bearing data formats compatible with process mining tools. The latter provide the machinery for the identification of *typical* empirical realizations of treatment processes as defined theoretically by medical guidelines and, in doing so, deliver the main (statistical) input for all successive compliance analyses. In terms of methodology, statistics mainly contributes to the interpretation of empirical process structures, based on which the

re-design of process models commences with respect to a perspective of population health. In this regard, cancer registries and cause-of-death registries, respectively, provide the most valuable analytical contributions of official statistics as these sources of data allow the calculation of both incidence and prevalence rates (using established principles of epidemiology); however, this issue will not be pursued any further in what follows.

2.2 Comparing Guidelines with Empirical Data

Analyzing the relationship between the process model defined by guidelines on the one hand and empirical process data represented in formal event-logs on the other hand constitutes one of the main tasks in the stages 2 and 3 of the L* model. In EBMC^2 context, an overall methodology is being developed that admits the transformation of both, real world medical data and (mostly informal) medical guidelines, to a (formal) level of representation amenable to a rigorous analytical comparison. Furthermore, this transformation is supposed to allow compliance checking to feed back derived evidence-based process models to adaptations of the guidelines themselves (Fig. 2).

This model conceives the treatment process depending, in all medical applications, on two different types of attributes, viz. (i) patient specific attributes (denoted by X in Figure 2), and (ii) institutional attributes, θ, capturing a variety of peculiarities of health care units possibly affecting treatments in a relevant way; the latter attributes are termed "parameters" further on. Usually, medical guidelines do not include all patient attributes which may conceivably influence the treatment process. For example, the guideline for skin cancer treatment accounts for the staging of the tumor only, but omits any reference to personal characteristics such as age or gender. Empirically, of course, such attributes may account significantly for a treatment progression. Hence, Figure 2 splits the personal attributes, $X = (X_d, X_p)$, where X_d denotes the set of diagnostic attributes

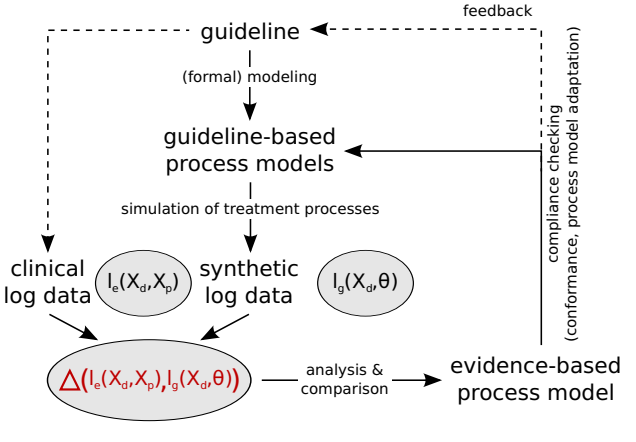

Fig. 2. Basic Methodology (cf. [5])

used explicitly in the guideline, while X_p refers to patient attributes not at all mentioned there.

The left-hand part of Figure 2 indicates the analytical path producing real world clinical data from applying treatments to (melanoma) patients supposedly implementing the pertinent medical guideline. Accordingly, the emanating clinical log data reflect both structure and terminology used by the guideline, albeit originating from different sources (e.g., hospitals, or clinical information systems etc.). Contrary to the medical processes bringing forth event-log data, the middle section of Figure 2 visualizes the path from guideline via formal process model (derived from the guideline) to synthetic log data, obtained from conducting simulated patient treatment processes. In other words, in a first step the guideline gets transformed into a guideline-based process model based on the verbal guideline as well as additional knowledge of domain experts, enriching the guideline-based process model. This enriched model, taking account of also diagnostic attributes as well as the institutional parameters (the statistical distributions of which chosen to reflect real patient demographics), feeds into a process simulation tool that produces synthetic log data as its output.

Both clinical and synthetic log data can now be analyzed with process mining tools to sift out empirical process representatives. The first step in analyzing log data is to generate a process model which allows further statistical techniques to be applied like clustering [1] or decision analysis [17]. In this respect, it is of course particularly interesting to inquire whether, and to which degree, there is consensus with medical experts about the actually mined treatment processes compared to their guideline-based reference process: process mining thus is used as a means of checking consistency of guideline specifications (cf. Section 5 for an example). A pivotal *delta* analysis [1] consists in comparing clinical and synthetic log data to assess actual concordance. Since, in general, treatment processes may depend on personal patient attributes, some diversity amongst – still compliant – treatment processes is to be expected (for a preliminary description cf. [5]). For data with irregular time structure methods well known in the area of longitudinal data analysis [7,8] can be applied.

Finally, the right-hand part of Figure 2 emphasizes the feedback of analysis results as condensed analytically in the derived evidence-based process model to the formal treatment process model (to better fit observed clinical practice), or – eventually – to the adaptation of the guideline itself. Feedback with respect to the population health as mentioned in the previous section will not be pursued any further in the present paper.

3 Available Data Sources: Problems and Challenges

For the time being, work in EBMC^2 draws upon two main sources of melanoma-related data, viz. (i) a detailed data collection of clinical *Cutaneous Melanoma* (CM) stage IV protocols (Stage IV Melanoma Database, S4MDB, for short), and (ii) administrative data of the Main Association of Austrian Social Security Institutions comprising a billing-oriented view of medical patient treatments

(GAP-DRG). In neither case, data sources conform well to the L* life cycle model (cf. [2]). This is true also of another relevant source of data, the Austrian Cancer Registry, which, however, is not used in EBMC^2 as yet.

3.1 The Stage IV Melanoma Database

Cutaneous Melanoma (CM) is a malignant tumor arising from melanocytic cells and primarily involving the skin. It is the deadliest form of skin cancer causing approximately 90 % of skin cancer mortality [10]. CMs are staged according to the classification of the American Joint Committee on Cancer (AJCC) [3]. In a retrospective data analysis, 389 patients diagnosed with *Metastatic Melanoma* (MM), who underwent therapeutic treatment between 2000 and 2010 at the Department of Dermatology, Medical University of Vienna, were included in an electronic database. Only patients with a histologically verified melanoma (AJCC stage IV) were included in this study. The data were retrospectively assessed by 2 diploma students under the supervision of an expert in dermato-oncology. The progress of patients with MM, the median survival rate and the influence of different treatment options on survival were the major objectives of this project. Furthermore, the median survival and long-term survival rate were evaluated with reference to different prognostic criteria. The study was approved by the local ethics committee. The data acquisition was based on 3 different sources: the hospital information system (HIS) as well as digital archived and physical patient records. The patient data were anonymized, using a unique patient-specific identification number. Relevant parameters included melanoma type, localization of the primary tumor, date of the primary excision, histological diagnosis, and the AJCC tumor stage at time of excision. For all metastatic events, diagnostic procedures to verify the metastatic tumor stage, localization of metastases, and treatment regimes, including the number of treatment cycles and the treatment duration, were evaluated.

In view of process mining purposes, evaluation of available stage IV data made it apparent quickly that data could not be used as is; rather, a structured data re-acquisition was due. While available data met the requirements for medical evaluation of treatment cycles, durations, and outcomes of patients with MM, they often typically lacked appropriate *temporal* references which, for obvious reasons, are of utmost importance to process mining. Additionally, causative relationships of different clinical pathways are of pivotal interest. Thus, a re-design of the original S4MDB structure became inevitable, including particularly time-related clinical pathways based on the analysis of individual patients.

3.2 Austrian Social Security Institution Data

The Main Association of Austrian Social Security Institutions maintain a huge data repository gathering patient and treatment related medical data, for accounting and billing purposes, of all (domestic as well as foreign) patients treated in Austria. A subset of pseudonymized data covering the period from January

2006 to December 2007 is available in an accessible GAP-DRG database including hospitalization data [4], patient treatments received from resident physicians, administered medications dispensed at pharmacies as well as sick certificates. GAP-DRG harmonizes some 15 data bodies of different Austrian Social Security Institutions, making use of established taxonomies such as (i) the ICD10 (International Statistical Classification of Diseases and Related Health Problems) to document primary and ancillary diagnoses, (ii) a MEL catalog [4] to describe medical treatments in hospitals, (iii) a national "Metahonorar" catalog coding medical treatments of resident physicians, and (iv) the ATC (Anatomical Therapeutic Chemical Classification System) codes to categorize the active agents of medications dispensed at pharmacies. GAP-DRG relates data entries to a rough temporal grid of billing periods determined by (legal) accounting requirements. Although entailing some effort, data entries can be linked to pseudonymized individual patients, so that the data subset of skin cancer (melanoma) patients can be selected effectively from the database.

3.3 Data Source Integration and Assessment

The joint analysis of CM patient treatment data of different provenance necessitates a preliminary data integration step. Accordingly, EBMC^2 explores and applies Extract-Transform-Load (ETL) methodologies [13] to wrap data sources, highlighting particularly

- *time-related data* ([1]: 113) indispensable for ordering medical activities and deriving / mining treatment processes, and
- *routing data* flowing into decision point analysis for expressing process branchings.

Identification and allocation of *time-related data* are acknowledged challenges in extracting event-logs from data sources, as are the different levels of granularity ([1]: 114). Compared to the level hierarchy established in the Process Mining Manifesto [2], the data sources introduced in Subsections 3.1 and 3.2, respectively, roughly fare as exhibited in Table 1.

First, and preliminary, process mining experience in analyzing patient data from both of the data sources considered can be summarized as follows:

- S4MDB data: throughout, a process comprising a sequence of three activities, appearing in identical order, was mined for all of the patients. Due to the specific purpose of the S4MDB, no system log was maintained for data recording, conveying only limited information to process mining. However, the database offers quite detailed information potentially useful as *routing data* for analyzing decision points within treatment processes.
- GAP-DRG data: After transformation, this dataset results in fairly meaningful process models due to the temporal indexing of the provided data. Conversely, it became apparent that *routing data* are lacking by and large, compromising decision point analysis severely.

Table 1. Evaluation of Data from Different Data Sources

	Level	Characteristics
Medical Records	⋆	• recorded by hand
S4MDB	⋆ / ⋆⋆	• recorded by hand • no systematic approach • trustworthy • correct
GAP-DRG	⋆⋆	• recorded by hand • recorded systematically • trustworthy • correct

– A *guided* post-collection of additional, or enriching, data is indispensable for obtaining satisfactory process mining results: since, in general, existing data sources have been designed quite unaware of later use in process mining, linking them to add-on data providing temporal references or further explanatory patient attributes, whenever possible, helps to increase their analytical value for process mining significantly.

Unfortunately, this is a recommendation hard to achieve most of the time because, as in the medical domain of interest, available data often is structured weakly only (e.g., in terms of digital scans of documents) and prepared ad hoc for particular analyses; furthermore, the diversity and locality of data sources tends to inhibit effective data integration, a problem additionally exacerbated by problems of measurement incommensurability and taxonomy mismatches enforcing, e.g., a laborious recoding or disadvantageous aggregation of data entries.

4 Empowering the Data Sources

Each data source presents its own challenges, rendering a *universal* approach of data management quite elusive; therefore, also in this process/data mining project an indispensable preparatory step concerns the consolidation of data sources [6]. Accordingly, the following subsections describe crucial steps of data preparation by data source. This experience report of data consolidation illustrates the importance of this central task to build a basis for the following process mining application.

4.1 Meta Data Enrichment of the Stage IV Melanoma Database

As already mentioned, the S4MDB lacks critical *time-related data*. While retaining the flat S4MDB structure, a set of meta data is introduced to add *structural* information. Figure 3 depicts an XML schema of this ensuing enrichment file. Now, a medical activity can either correspond to a *pointInTime* (with a single date) or a *period* (with both, a start and an end date). XML element attributes

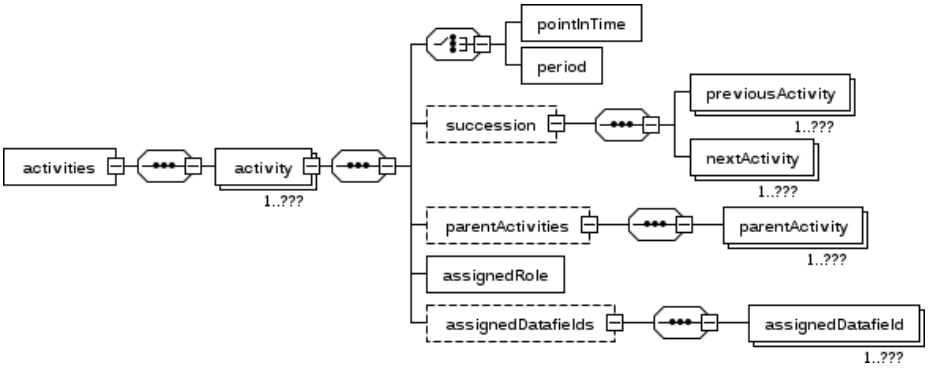

Fig. 3. XSD for Meta Data Enrichment of the Stage IV Melanoma Database

are used for referencing the S4MDB column headings comprising the respective data. If not yet existing, further columns are simply added to the S4MDB structure.

The *succession* element of the XML enrichment file enables the interpolation of calendar dates, whenever missing, for activity entries in S4MDB. To accomplish the interpolation, all *previousActivitiy* and *nextActivitiy* elements can be searched jointly in the meta XML file and S4MDB; detecting the latest calendar date of all previous activities and the earliest calendar date of all successive activities, respectively, allows derivation of the missing calendar date.

Likewise, the element *parentActivities* helps to resolve different data granularities (cf. Subsection 3.3). For instance, S4MDB, comprises a lot of subordinate follow up treatments not even mentioned in the guideline. Therefore, it is reasonable to group such treatments to the granularity level of the guideline in order to accommodate for comparisons without changing the original database.

The element *assignedRole* admits inclusion of organizational information about which institution performed the recorded activity. While not of specific interest for the time being, this information may become relevant later on. Additional data different from genuine activities to record – the bulk of S4MDB data, indeed – can be added using the *assignedDatafields* element; this information is of use in, e.g., decision point analysis.

The meta data enrichment XML file for the existing S4MDB, comprising a test sample of ten patients, was created by medical practitioners of EBMC^2 for evaluation purposes reported on in Section 5.

4.2 Restating the Data Model

Meta data enrichment relieves, to some degree, the rigidity of the flat S4MDB structure. To gain yet more versatility in representing treatment process data, the data model depicted on the left side of Figure 4 was proposed in favor of the requirements declared in Section 2. Similar to meta data enrichment, it defines activities (*Activity*) with assigned data fields (*Narrative_Parameter*).

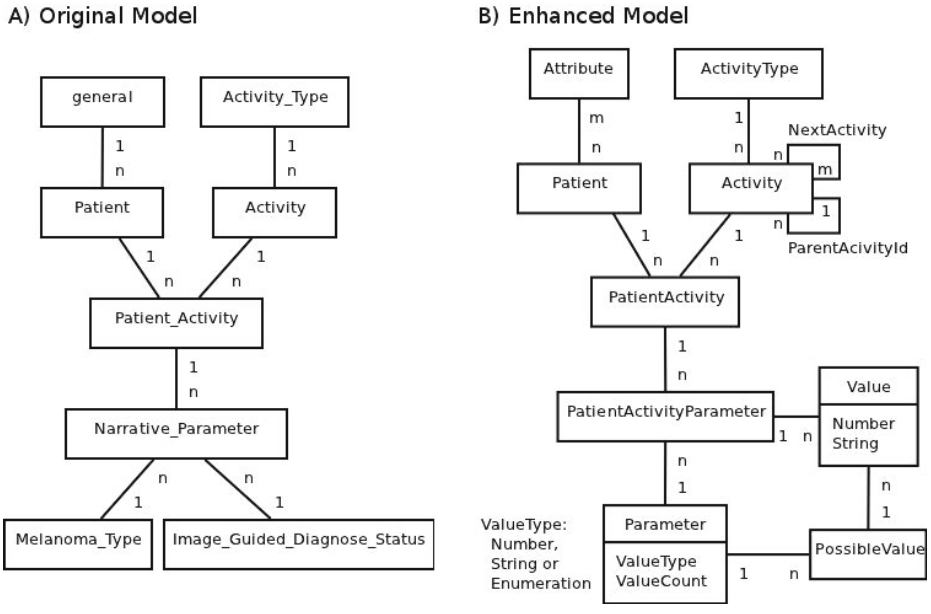

Fig. 4. Initial Melanoma Data Structure and Adapted Extensible Data Structure

This model does not yet comply with the Process Mining Manifesto [2], and thus does not apply generally to all conceivable clinical applications. Hence, a further generalization of the data model is depicted on the right side of Figure 4, allowing for the addition of – typed – data fields as separate entities, this way avoiding changes in the data structure while keeping the data model consistent and normalized. In turn, this more generic approach demands a higher discipline of data collectors as it allows the *ad hoc* definition of possible values and data types. The model also includes (i) information about the sequencing of activities (*NextActivity*), useful, e.g., in interpolating missing activity dates, and (ii) hierarchical information (*ParentActivity*) for handling varying value granularities of data entries.

The devised data structure has been implemented as a web application (named *PTDoc*, Patient Treatment Documentation), using the ZK Java Framework [16] on top of MySQL [15] as data store. The user interface, cf. Figure 5, splits into three panes, to be walked through from left to right:

- Pane 1: select either an existing patient, or add a new one
- Pane 2: assign either an existing activity or a new one to this patient
- Pane 3: add new parameters to that activity of this patient

The application has an *administration section* where new activities as well as their parameters and default values can be registered, modified or deleted. The PTDoc application is currently evaluated at the Department of Dermatology at the Medical University of Vienna.

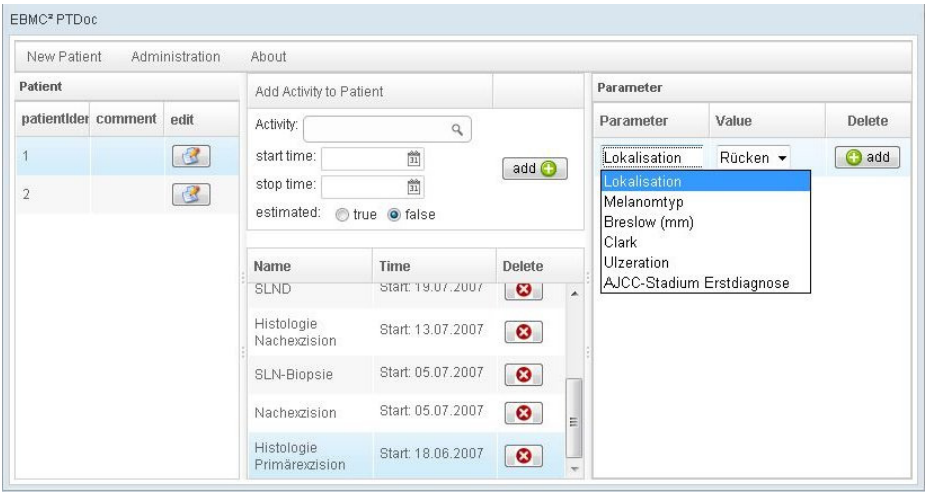

Fig. 5. Screenshot: Prototype - Patient Treatment Data

4.3 GAP-DRG Data Transformation

As described, this data source provides rather good time-related data. However, when it comes to routing data, little information is offered only. Hence, again, in spite of its analytical usefulness, GAP-DRG data needs a preprocessing.

Essentially, GAP-DRG data is organized by hospitalizations, each carrying calendar dates of patient admissions and dismissals but no direct reference to treatments (activities). Rather, diagnosis codes are attached to hospitalizations from which the relevance of the data for subsequent analysis can be deduced. Diagnosis codes also carry further activity-related information such as, for example, an ICD-10 code C43-1480 entry telling implicitly that the patient had a malign melanoma excised on the recorded day of hospitalization; additionally, the code also discloses the localization of the melanoma (in this case, the ear or the acoustic meatus). To achieve a canonical representation of this information, such super-code data entries need to be split into elementary PTDoc components.

Likewise, GAP-DRG data does not provide direct information about administered treatments as defined in the medical guideline; rather, medications picked up at the pharmacy are recorded, from which underlying treatments may be reconstructed.

5 Evaluation Based on Selected S4MDB Cases

In a pilot trial of process mining comparing actual clinical treatment processes with specific peer-reviewed medical guidelines, data of 10 patients selected from S4MDB (cf. Subsection 3.1) were entered in PTDoc. In this first attempt of deriving an evidence-based melanoma treatment process model, the heuristic miner (HM, for short) of the ProM framework [14] was used to generate a process

model using the provided data. The ProM framework hosts a set of modules implementing various process mining algorithms processing input log data. After transforming S4MDB data into the ProM input log file format, it was possible to generate a first process model. This model was used for evaluation with domain experts to verify its compliance with medical reality.

During discussion of the resulting process model, generated with the HM, with dermatological experts, the following three conspicuous features became apparent; cf. excerpts of this process model shown in Figures 6 and 7, respectively:

- starting activities are mis-identified as concurrent (Figure 6);
- an additional treatment commencing during another treatment already applied, properly reflecting reality (Figure 7 A);
- (two) treatments applied in succession, contradicting medical reality (Figure 7 B).

In both figures, each activity bears either a *start* or a *complete* label indicating the starting and ending times of activities, respectively (instantaneous activities record an ending time only). Within the figures, numbers represent (absolute) frequencies of activities or activity transitions.

Figure 6 depicts the case of activities *First Medical History* and *Excision Melanoma* identified erroneously as feasible starting activites pari passu while *First Medical History* – recorded in only 3 out of 10 observed cases – is supposed to precede *Excision Melanoma* throughout. Checking the data set for these three cases immediately explains why the HM assumes these activities as parallel: in only one case *First Medical History* actually preceded *Excision Melanoma*. In another case both activities recorded the same day, thus inhibiting the inference of any activity precedence by the algorithm, while in the third case the sequence of the activities had been recorded in reverse order – probably a previously unrecognized data glitch. Based on just three cases like these, the HM simply cannot detect the correct activity precedence relation.

Conversely, Figure 7 A) demonstrates a treatment modality for MM patients where, during treatment with the chemotherapeutic drug Fotemustine, also an IL2sc treatment – a proinflammatory cytokine – commenced. This treatment pathway is plausible and represents clinical reality.

In contrast to that, Figure 7 B) again demonstrates a treatment modality for MM patients where the application of Gamma Knife (used in radiosurgery of the head to treat brain metastases) follows the treatment with Thalidomide (a drug also used in oncology). Both treatments are applied for patients with late-stage melanoma. However, even though the therapeutic setting implies a thematic proximity of these treatments, this time the induced activity precedence conflicts clinical reality, representing rather an artifact probably caused by another data error.

Altogether, these examples of a first evaluation of the methodology show the strengths of the process view as a data-screening instrument. Clearly, a sample of just 10 observation cases falls short of admitting any far-reaching conclusions, but still can pinpoint salient data (and data quality) requirements as well as the advantages of a *process-centered* view of patient treatment data, shedding light on guideline adherence of medical practice.

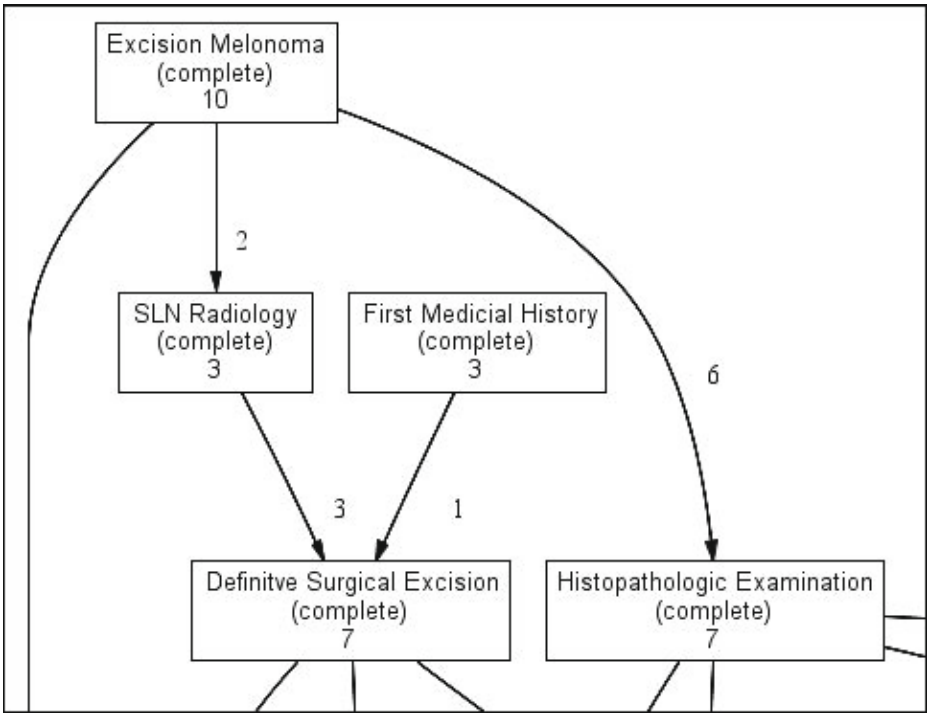

Fig. 6. Excerpt from the Process Model minded by the Heuristic Miner showing the identified starting activities

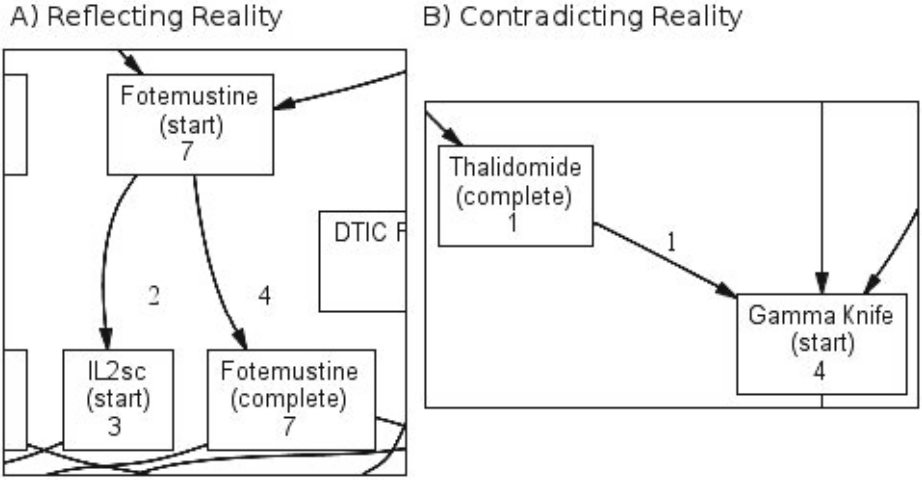

Fig. 7. Excerpt from the Process Model minded by the Heuristic Miner showing an activity sequence A) reflecting and B) contradicting reality

6 Summary and Outlook

In this report, we presented first experiences with the process-driven analysis of skin cancer treatment processes in the EMBC^2 project, specifically analysis of their guideline compliance. Focus was put on the transformation and integration of the available data sources, i.e., clinical Cutaneous Melanoma stage IV protocols as well as billing data of the Main Association of Austrian Social Security Institutions. The challenge was to extract and integrate the data in a process-oriented way in order to apply process mining techniques in the sequel. Due to the characteristics of the different data, techniques such as meta data enrichment and extended data models are proposed. A glimpse into process mining results in the form of data screening are presented. In future work, we will extend and enrich the available data in order to analyze the differences between the actual treatment processes and the corresponding guidelines and will use melanoma-relevant excerpts of the Austrian Cancer Registry for interpretation.

References

1. van der Aalst, W.: Process Mining: Discovery, Conformance and Enhancement of Business Processes. Springer, Heidelberg (2011)
2. van der Aalst, W., et al.: Process Mining Manifesto. In: Daniel, F., Barkaoui, K., Dustdar, S. (eds.) BPM Workshops 2011, Part I. LNBIP, vol. 99, pp. 169–194. Springer, Heidelberg (2012)
3. Balch, C., et al.: Final version of 2009 ajcc melanoma staging and classification. Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology 27(36), 199–206 (2009)
4. Bundesministerium für Gesundheit, Leistungsorientierte Krankenanstaltenfinanzierung - L K F - Medizinische Dokumentation, `http://www.bmg.gv.at/cms/home/attachments/0/4/1/CH1166/CMS1128332936305/medizinische_dokumentation_2012.pdf` (accessed November 20, 2011)
5. Dunkl, R., Fröschl, K.A., Grossmann, W., Rinderle-Ma, S.: Assessing Medical Treatment Compliance Based on Formal Process Modeling. In: Holzinger, A., Simonic, K.-M. (eds.) USAB 2011. LNCS, vol. 7058, pp. 533–546. Springer, Heidelberg (2011)
6. Džeroski, S.: Towards a General Framework for Data Mining. In: Džeroski, S., Struyf, J. (eds.) KDID 2006. LNCS, vol. 4747, pp. 259–300. Springer, Heidelberg (2007)
7. Everitt, B., Hothorn, T.: A Handbook of Statistical Analyses Using R. Chapman & Hall-CRC Press, Boca Raton (2006)
8. Fitzmaurice, G., et al.: Longitudinal Data Analysis (Chapman & Hall/CRC Handbooks of Modern Statistical Methods). Chapman & Hall-CRC Press, Boca Raton (2009)
9. Fox, J., Black, E., Chronakis, I., Dunlop, R., Petkar, V., South, M., Thomson, R.: From guidelines to careflows: Modelling and supporting complex clinical processes. In: Computer-based Medical Guidelines and Protocols: a Primer and Current Trends, pp. 44–61. IOS Press, Netherlands (2008)

10. Garbe, C., Peris, K., Hauschild, A., Saiag, P., Middleton, M., Spatz, A., Grob, J., Malvehy, J., Newton-Bishop, J., Stratigos, A., Pehamberger, H., Eggermont, A.: Diagnosis and treatment of melanoma: European consensus-based interdisciplinary guideline. European Journal of Cancer 46(2), 270–283 (2010)
11. Guyatt, G., Oxman, A., Schünemann, H., Tugwell, P., Knottnerus, A.: Grade guidelines: A new series of articles in the journal of clinical epidemiology. Journal of Clinical Epidemiology 64(4), 380–382 (2011)
12. Health Level Seven International, http://www.hl7.org/ (accessed November 29, 2011)
13. Jarke, M., Lenzerini, M., Vassiliou, Y., Vassiliadis, P.: Fundamentals of Data Warehouses. Springer, Heidelberg (2010)
14. Mans, R.: Workflow Support for the Healthcare Domain. Proefschriftmaken.nl, Netherlands (2011)
15. Oracle Corporation, http://www.mysql.com/ (accessed November 28, 2011)
16. Potix Corporation, http://www.zkoss.org/ (accessed November 28, 2011)
17. Rozinat, A., van der Aalst, W.M.P.: Decision Mining in ProM. In: Dustdar, S., Fiadeiro, J.L., Sheth, A.P. (eds.) BPM 2006. LNCS, vol. 4102, pp. 420–425. Springer, Heidelberg (2006)
18. The International Health Terminology Standards Development Organisation, http://www.ihtsdo.org/ (accessed November 29, 2011)
19. The openEHR Foundation, http://www.openehr.org/ (accessed November 28, 2011)