A formal approach to build privacy-awareness into clinical workflows

Article · August 2019

DOI: 10.1007/s00450-019-00418-5

2 authors, including:

Saliha Irem Besik
Humboldt-Universität zu Berlin
2 PUBLICATIONS **0** CITATIONS

SEE PROFILE

SICS Software-Intensive Cyber-Physical Systems
https://doi.org/10.1007/s00450-019-00418-5

SPECIAL ISSUE PAPER

A formal approach to build privacy-awareness into clinical workflows

Saliha Irem Besik[1] · **Johann-Christoph Freytag**[1]

Abstract
Clinical workflows consist of sets of tasks involving patients and healthcare professionals. In such an environment, maintaining the privacy of patient data is a significant challenge. Healthcare providers have to consider both legislative compliances with tightening privacy regulations and growing privacy concerns of individuals. Unlike data security, which aims at preventing unauthorized access, privacy focuses on providing individuals the ability to control when, how, and to what extent their data is used with a particular purpose. In this paper, we present our first steps on transforming existing non-privacy-aware clinical workflows into privacy-aware ones through algorithms based on privacy policies and privacy preferences.

Keywords Data privacy · Business process modeling · General Data Protection Regulation (GDPR) · Compliance checking · Privacy policies · Privacy preferences

1 Introduction

The European Union General Data Protection Regulation (GDPR) has come into force very recently to protect the data privacy of all EU citizens [1]. "Data concerning health" has a special mention under the GDPR and it is subject to a higher standard of protection than personal data in general [GDPR, Article 9]. Data protection regulations define the principles to be met by organizations when processing personal data in order to guarantee privacy. The current privacy laws and regulations mandate that all healthcare providers specify privacy policies regarding their use and disclosure of personal health data. A privacy policy is a comprehensive and high-level description of a service provider's privacy practices [2]. Violation of compliance with the regulations and privacy policies could lead to heavy fines and loss of reputation. However, the complexity of the healthcare domain, the regulations, and privacy policies make it difficult to achieve compliance.

A survey regarding the views of EU citizens about issues surrounding data protection poses that around seven out of ten of them are concerned about their information being used for a different purpose from the one it was collected for [3]. In order to address these concerns, patients should be able to express preferences on sharing and/or processing their personal data.

Healthcare is often organized in a process-oriented way via clinical workflows. It is very important to capture the privacy requirements at the conceptual level; however clinical workflows generally do not support privacy constraints in an adequate way. Most approaches to workflow modeling and management in healthcare are founded on a *control-flow-centric* perspective, which focuses primarily on causal and temporal dependencies between tasks. Our approach is based on *data-centric* process modeling paradigm because data privacy in clinical workflows is a subject of how data objects are handled. Business Process Model and Notation (BPMN) is a de-facto standard for business process modeling, hence, we use BPMN for modeling clinical workflows.

Motivated by these above statements, *our aim is to transform existing non-privacy-aware clinical workflows into privacy-aware ones*. In our research, three sources contribute to privacy-awareness. Firstly, privacy-aware clinical workflows should be compliant with the privacy principles based on the GDPR. Secondly, privacy-aware clinical workflows should be compliant with the privacy policies provided by healthcare providers. Finally, privacy-aware clinical workflows should also consider the rights of patients to request restrictions or preferences on sharing or processing their personal medical data.

✉ Saliha Irem Besik
saliha.irem.besik@informatik.hu-berlin.de;
besiksal@informatik.hu-berlin.de

[1] Department of Computer Science, Humboldt-Universität zu Berlin, Berlin, Germany

Published online: 26 August 2019

We worked on Newborn Screening procedure applied in Germany as a running example in order to illustrate our methodology. Newborn Screening is an optional procedure in Germany which requires the explicit consent of at least one of the parents or guardians of the newborn babies. This procedure includes complex clinical tasks involving highly sensitive medical data like blood samples and several different healthcare providers such as pediatricians, nurses and laboratory assistants.

The rest of this paper is organized as follows: Sect. 2 gives an overview of our proposed approach. Section 3 introduces three sources which contribute to privacy in our research: the GDPR, privacy policies and privacy preferences. Section 4 explains how we built an ontology through integrating privacy concepts and the BPMN elements. Section 5 states our formal definitions for privacy rules coming from privacy policies and privacy preferences. Section 6 gives details about how we formally define the compliance check. Section 7 presents our introductory steps towards transformation of non-privacy-aware workflows into privacy-aware ones. Section 8 discusses related work. Finally, Sect. 9 concludes this paper and discusses our future direction.

2 Approach

Figure 1 gives an overview of our research setting. Our proposed solution to transform existing non-privacy-aware clinical workflows into privacy-aware ones can be summarized as follows. We have non-privacy-aware clinical workflows as input. We define the privacy-awareness as the compliance of the clinical workflows with the privacy principles based on the GDPR, the privacy policies, and privacy preferences.

(i) *Integrate concepts and build ontology* We semantically represent privacy concepts which are related to the privacy principles, the privacy policies, and privacy preferences; and the BPMN elements in order to bridge the gap and share a common understanding between the domain of clinical workflows and the domain of privacy.

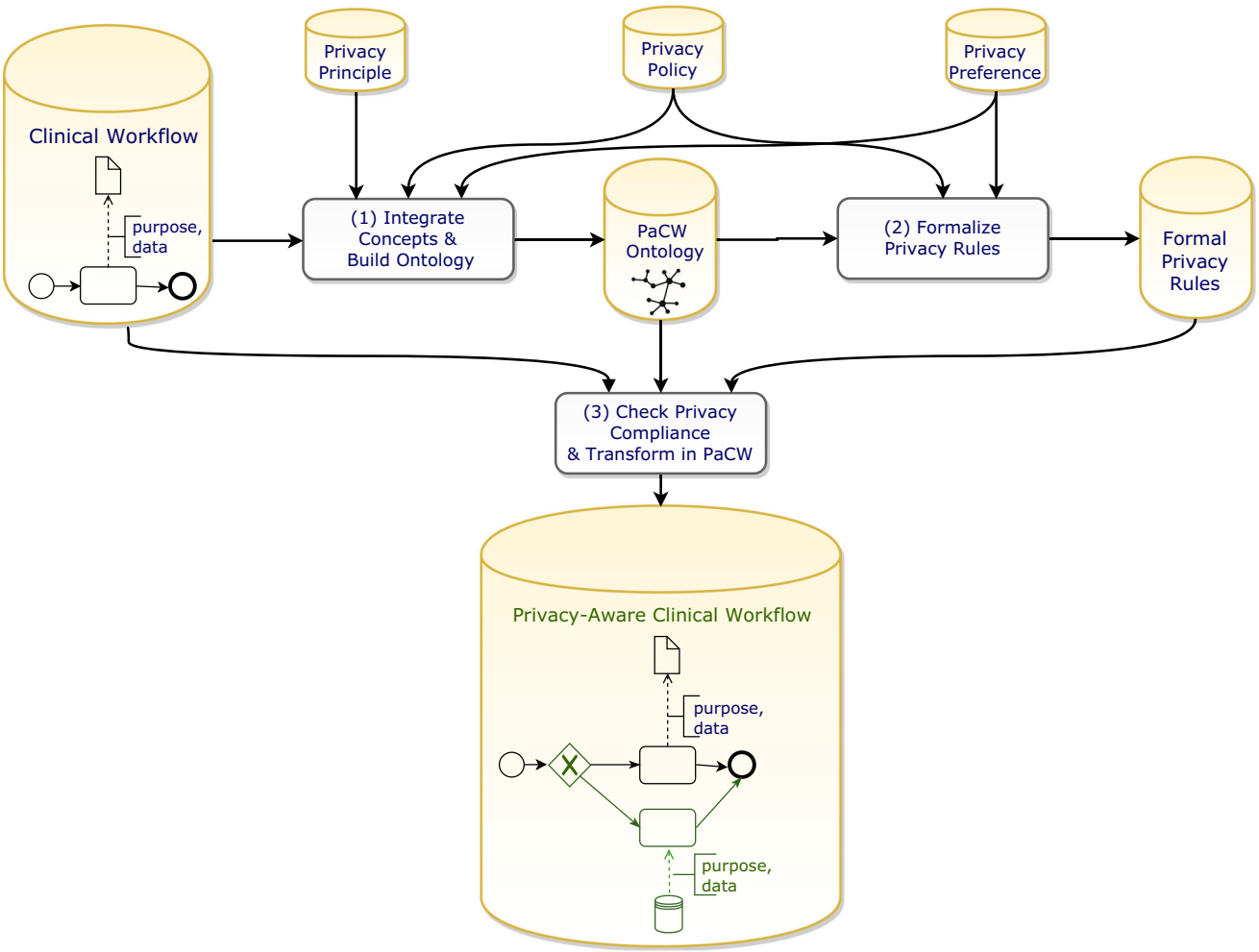

Fig. 1 Conceptual diagram

(ii) *Formalize privacy rules* We formally define privacy rules coming from privacy policies and privacy preferences.

(iii) *Check privacy compliance and transform into privacy-aware clinical workflow (PaCW)* We check privacy compliance of BPMN-based clinical workflows with formalized privacy rules and we provide a transformation algorithm to apply in case of privacy violation. Below the individual steps are further explained.

3 Privacy concepts

There are three sources which contribute to privacy in our research: the GDPR, privacy policies and privacy preferences. We analyzed them in order to understand and express the semantics of our privacy rules.

3.1 Privacy principles based on the GDPR

We introduced the founding privacy principles for clinical workflows on the ground of the GDPR:

- *Purpose specification* "Personal data shall be collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes [...]" [Article 5, §1(b)].
- *Data minimization* "Personal data shall be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed." [Article 5, §1(c)].
- *Consent check* "Processing shall be lawful if the data subject has given consent to the processing of his or her personal data for one or more specific purposes." [Article 6, §1(a)].
- *Limited retention period* "Personal data shall be kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed [...]" [Article 5, §1(e)].

The GDPR presents some other privacy principles which must be obeyed, such as data transparency and data accuracy [Article 5, §1(a) and 1(d)]. However, we initially focus the above principles of purpose limitation, consent check, limited retention period, and data minimization.

3.2 Privacy policies

Privacy policies primarily contain information regarding what data is collected, for what purpose the data will be processed, who are the data recipients, and how long the data will be retained. They might also describe the categories and typologies of the data; and the modality of data processing, whether it is obligatory or voluntary. Privacy policies include privacy policy statements and we entitle each of these statements as 'privacy policy rules'. Some of the example privacy policy rules for our newborn screening scenario are shown in Table 1.

Table 1 Privacy policy rules

P1. An explicit consent is required for newborn hearing screening
P2. A hospital can save results for the purpose of hearing screening with a retention limit by 3 years
P3. A nurse on duty can collect blood of newborn babies within three days after birth
P4. A nurse can transfer blood only to the assigned Newborn Screening Laboratories
P5. A laboratory assistant can request repetitive blood collection only if the examination is unsatisfactory
P6. A pediatrician can access the result of the lab examination only if the result is abnormal

Table 2 Privacy preference rules

R1. Alice does not give consent for her newborn baby's screening data to be shared with health insurance companies on March 1, 2019
R2. Alice prefers that only the pediatrician Bob can access her newborn baby's screening results for 6 months on March 1, 2019
R3. Alice does not give consent for her newborn baby's screening data to be used for the purpose of research on March 1, 2019
R4. Alice gives consent that medical staff can access her newborn baby's medical data without any restrictions for emergency cases on March 1, 2019

3.3 Privacy preferences

Patients might express their preferences on who can access their personal data and for what purposes via informed consents. They can also define the duration of their consent. Some of the example privacy preference rules for our newborn screening scenario are shown in Table 2.

4 Integrating concepts and building ontology

We created an ontology to represent the privacy concepts we presented in Sect. 3 and BPMN elements. Figure 2 illustrates a part of our privacy-aware clinical workflow (PaCW) ontology. The foundational components of privacy rules are *user*, *data*, and *purpose*.

User is the set of individuals or organizations who can access the personal data. There are different roles of users both from the privacy point of view which are defined in the GDPR (*data*

Fig. 2 A Part of PaCW ontology

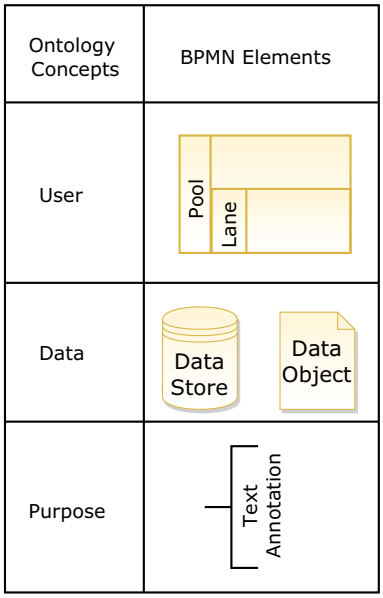

Fig. 3 Mapping between ontology concepts and BPMN elements

subject, data processor, data controller, data recipient, etc.) and from our running example (*patient, pediatrician, nurse, laboratory assistant,* etc.).

Data is categorized to adopt the right privacy measures suitable for the type of data to be protected. There are different data categories both from a privacy point of view which are defined in the GDPR (*personal data, sensitive data, identification data, anonymous data, public data,* etc.) and from our running example (*medical data, demographic data, contact data,* etc.).

Purpose specifies the reason for which data is collected, used, or disclosed. There are different purposes both from a privacy point of view which are defined in the GDPR (*public interest, public health, statistical analysis, research, marketing,* etc.) and from our running example (*hearing screening, blood screening, treatment,* etc.).

The following concepts support the foundational concepts described previously:

Retention defines the period of time for how long the data can be stored.

Condition defines the additional conditions within privacy rules regarding usage and disclosure of personal data. For instance, a privacy policy may specify that a particular data item can be accessed, but only with "opt-in" consent from the data subject.

Preference status specifies whether privacy preference holding a positive statement or not.

Preference duration is the duration of the privacy preference, which indicates how long the privacy preference is valid.

Entry date is the entry date of each data and privacy preference.

The above foundational components *user*, *data* and *purpose* can be represented in clinical workflows, using a set of BPMN constructs. Figure 3 shows the mapping between privacy concepts and BPMN elements. *User* can be mapped onto the BPMN *Pool* and *Lane* elements. *Data* can be mapped onto the BPMN *Data Object* and *Data Store* elements. BPMN *Message Flow* can also represent *Data* because it describes sharing data among different participants. However, we excluded it in this article as we initially focus only on the privacy rules regarding accessing data not sharing data. Finally, *Purpose* can be represented via the BPMN *Text Annotation* element.

5 Formalizing privacy rules

After creating the ontology, we built a formalization on top of this ontology so as to automate management of privacy in clinical workflows. We use our PaCW ontology to comprehend the semantic relationships among different elements. For instance; according to our ontology *pediatrician* is a *medical staff*. When a privacy rule is regarding *medical staffs*, it is also about its subclasses including *pediatricians*. We have privacy rules coming from privacy policies and privacy preferences.

5.1 Formalizing privacy policy rules

We categorized the privacy policy rules in three according to their types: *Consent privacy policy rules*, which are formally defined in Definition 1, state what kind of data operations requires an explicit consent. *Data minimization privacy pol-*

icy rules, which are formally defined in Definition 3, express the amount of personal data which should be revealed and processed. *Retention privacy policy rules*, which are formally defined in Definition 2, declare the retention period of data practices. We also show the formal representations of the privacy policy rules for newborn screening scenario in Table 1.

Definition 1 (*Consent privacy policy*) A consent privacy policy PC contains policy rules which are represented as 2-tuple pc = (purpose, requiresConsent), where:

- $purpose$ is the reason for which data is collected, used, or disclosed;
- $requiresConsent \in \{true, false\}$ specifies whether data processing requires consent or not;

Example 1 (*Consent privacy policy*) *(newborn-hearing-screening, true)* is the formal representation for *P1: "An explicit consent is required for newborn hearing screening."* in Table 1.

Definition 2 (*Retention privacy policy*) A retention privacy policy PR contains policy rules which are represented as 4-tuple pr = (user, purpose, data, retention), where:

- $user$ is the set of individuals or organizations who can access the personal data;
- $purpose$ is the reason for which data is collected, used, or disclosed;
- $data$ is a set of data objects;
- $retention$ defines the period of time the data is stored.

Example 2 (*Retention privacy policy*) *(hospital, result, newborn-hearing-screening, 3 years)* is the formalization for *P2: "A hospital can save results for the purpose of hearing screening with a retention limit by 3 years."* in Table 1.

Definition 3 (*Data minimization privacy policy*) A data minimization privacy policy PD contains policy rules which are represented as 4-tuple pd = (user, purpose, data, condition), where:

- $user$ is the set of individuals or organizations who can access the personal data;
- $purpose$ is the reason for which data is collected, used, or disclosed;
- $data$ is a set of data objects;
- $condition$ indicates additional conditions within the privacy policy regarding usage and disclosure of personal data. Each condition has a name and a state attribute, where condition.name is the name of the data within the condition and condition.state is the state of the data within the condition.

Example 3 (*Data minimization privacy policy*)

1. *(nurse, newborn-screening, blood-data, nurse.onDuty* $\wedge$ *birthDate.isSmaller3days)* is the formalization for *P3: "A nurse on duty can collect blood data of newborn babies for the purpose of newborn screening within three days after birth."* in Table 1.
2. *(Newborn-Screening-Laboratory, newborn-screening, blood-data, Newborn-Screening-Lab.isAssigned)* is the formal representation for *P4: "A nurse can transfer blood only to the assigned Newborn Screening Laboratories."* in Table 1.
3. *(lab-assistant, newborn-screening, (blood-data*$_1$, …, *blood-data*$_k$*), examination.isUnsatisfactory)* is the formal representation for *P5: "A laboratory assistant can request repetitive blood collection only if the examination is unsatisfactory."* in Table 1.
4. *(pediatrician, newborn-screening, examination-result, examination-result.isAbnormal)* is the formalization for *P6: "A pediatrician can access the result of the lab examination only if the result is abnormal."* in Table 1.

5.2 Formalizing privacy preference rules

We formally defined privacy preference rules in Definition 4. In Example 4, we present the formal representations of the preference rules for newborn screening scenario in Table 2.

Definition 4 (*Privacy preference*) A privacy preference R contains preference rules which are represented as 8-tuple r = (dataSubject, user, purpose, data, condition, duration, status, entryDate), where:

- $dataSubject$ is the set of individuals whom personal data is about;
- $user$ is the set of individuals or organizations who can access the personal data;
- $purpose$ is the reason for which data is collected, used, or disclosed;
- $data$ is a set of data objects or data categories;
- $condition$ indicates additional condition(s) within privacy preference regarding usage and disclosure of personal data;
- $duration$ is the duration of the privacy preference;
- $status \in \{true, false\}$ specifies whether privacy preference holding a positive statement or not;
- $entryDate$ indicates the entry date of the privacy preference.

Example 4 (*Privacy preference*)

1. *(Alice, health-insurance-company, any, newborn-screening-data, any, any, false, 2019-03-01)* is the formalization for *R1: "Alice does not give consent for her newborn baby's screening data to be shared with health insurance companies on March 1, 2019."* in Table 2.
2. *(Alice, any - Bob, any, newborn-screening-data, any, 6months, false, 2019-03-01)* is the formal representation for *R2: "Alice prefers that only the pediatrician Bob can access her newborn baby's screening results for 6 months on March 1, 2019."* in Table 2.
3. *(Alice, any, research, newborn-screening-data, any, any, false, 2019-03-01)* is the formal representation for *R3: "Alice does not give consent for her newborn baby's screening data to be used for the purpose of research on March 1, 2019."* in Table 2.
4. *(Alice, medical-staff, any, newborn-medical-data, emergency, any, true, 2019-03-01)* is the formal representation for *R4: "Alice gives consent that medical staff can access her newborn baby's medical data without any restrictions for emergency cases on March 1, 2019."* in Table 2.

6 Checking compliance with privacy principles

We denominate Clinical Workflows which we used as inputs as Data-Aware Workflows. In order to check the compliance, we formally defined Data-Aware Workflows in Definition 5 by extending the core BPMN process formal definition given by [4]. Data-Aware Workflow is a directed graph with nodes (components) $\mathcal{C}$ and arcs (sequence flows) $\mathcal{F}$. For any node $c \in \mathcal{C}$, set of ingoing sequence flows of c are given by $getIncomingFlows(c) = \{f \in \mathcal{F} | f_2 = c\}$ and set of outgoing sequence flows of c are given by $getOutgoingFlows(c) = \{f \in \mathcal{F} | f_1 = c\}$. Also, for any component node $c \in \mathcal{C}$, the name of c is given by $c.name$. [ ] symbol is used to illustrate partitioning. For instance; $\mathcal{T}[\mathcal{T}^{\mathcal{D}}]$ means that $\mathcal{T}$ can be partitioned into $\mathcal{T}^{\mathcal{D}}$. Figure 4 shows the core BPMN elements we used in our Data-Aware Workflow definition.

We assume that for each data operation task (BPMN tasks linked to a data object) in a BPMN workflow, it is known which data used for which purpose explicitly. For this purpose, we use BPMN text annotation elements as (D,p) with D referring a set of data and p as the *purpose* for accessing data. For the sake of simplicity, we assume that there is exactly one purpose for each data operation task. However, our approach can be adapted for more than one purposes. Figure 5 shows different versions of the data operation tasks we defined in our Data-Aware Workflow definition.

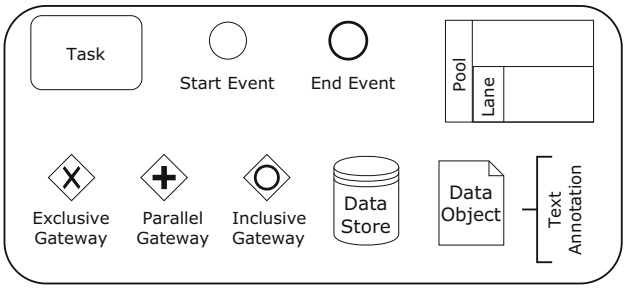

Fig. 4 Core BPMN elements

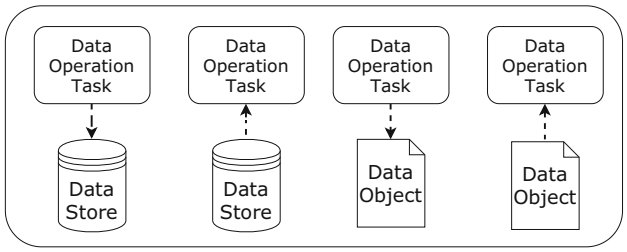

Fig. 5 Data operation tasks in BPMN

Definition 5 (*Data-aware workflow*) A data-aware workflow model is a tuple, $\mathcal{W} = (\mathcal{C}[\mathcal{T}[\mathcal{T}^{\mathcal{D}}], \mathcal{E}[e^{\mathcal{S}}, \mathcal{E}^{\mathcal{E}}], \mathcal{D}, \mathcal{G}[\mathcal{G}^{\mathcal{E}}, \mathcal{G}^{\mathcal{I}}, \mathcal{G}^{\mathcal{P}}]], \mathcal{L}, laps, \mathcal{F}[\mathcal{F}^{\mathcal{D}}], \lambda)$ where:

- $\mathcal{C}$ is a set of components which can be partitioned into disjoint sets of tasks $\mathcal{T}$, events $\mathcal{E}$, data objects $\mathcal{D}$, and gateways $\mathcal{G}$,
- $\mathcal{T}^{\mathcal{D}} \subseteq \mathcal{T}$ is a set of data operation tasks,
- $e^{\mathcal{S}} \in \mathcal{E}$ is a start event, $\mathcal{E}^{\mathcal{E}} \subseteq \mathcal{E}$ is a set of end events,
- $\mathcal{G}$ can be partitioned into disjoint sets of exclusive gateways $\mathcal{G}^{\mathcal{E}}$, inclusive gateways $\mathcal{G}^{\mathcal{I}}$, parallel gateways $\mathcal{G}^{\mathcal{P}}$,
- $\mathcal{L}$ is a set of lanes and pools,
- $laps : \mathcal{C} \rightarrow \mathcal{L}$ is a partial function which maps some of the components onto set of lanes and pools,
- $\mathcal{F} \subseteq \mathcal{C} \times \mathcal{C}$ is the control flow relation, i.e. a set of sequence flows connecting components, where $\mathcal{F}_1$ is the source component and $\mathcal{F}_2$ is the target component of the sequence flow $\mathcal{F}$,
- $\mathcal{F}^{\mathcal{D}} \subseteq \mathcal{F}, \mathcal{F}^{\mathcal{D}} \subseteq (\mathcal{T}^{\mathcal{D}} \times \mathcal{D}) \cup (\mathcal{D} \times \mathcal{T}^{\mathcal{D}})$ is a set of input and output data associations,
- $\lambda : F^{D} \rightarrow (D, p)$ is a annotation function which maps control flows to the tuples $(\mathcal{D}, p)$ where $\mathcal{D}$ is a set of data objects and $p \in \mathcal{P}$ is a purpose from universe of purposes $\mathcal{P}$; $\lambda_1(\mathcal{F}^{\mathcal{D}}) = \mathcal{D}$ and $\lambda_2(\mathcal{F}^{\mathcal{D}}) = p$.

We define a *case* including data-aware workflow, privacy policy and privacy preference in Definition 6. For instance, Newborn Hearing Screening Procedure Clinical Workflow, privacy policy taken from maternity clinic and privacy preferences taken from parents can be defined as a *case*.

Definition 6 (*Case*) A case is a 4-tuple, $\mathcal{S} = (\mathcal{DS}, \mathcal{W}, \mathcal{P}, \mathcal{R})$ where:

- $\mathcal{DS}$ is a data subject,
- $\mathcal{W}$ is a data-aware workflow,
- $\mathcal{P}$ is a set of privacy policy rules in which $\mathcal{P} = \mathcal{PC} \cup \mathcal{PR} \cup \mathcal{PD}$, and
- $\mathcal{R}$ is a set of privacy preference rules.

We define *privacy-aware case* as cases which are compliant with privacy principles in Definition 7.

Definition 7 (*Privacy-aware case*) A case $\mathcal{S}$ is privacy-aware iff

- $\mathcal{W}$ is compliant with purpose specification principle,
- $\mathcal{W} \wedge \mathcal{PC} \wedge \mathcal{R} \wedge \mathcal{DS}$ is compliant with consent check principle,
- $\mathcal{W} \wedge \mathcal{PD}$ is compliant with data minimization principle, and
- $\mathcal{W} \wedge \mathcal{PR}$ is compliant with limited retention principle.

6.1 Checking compliance with purpose specification principle

In order to be compliant with the purpose specification principle, each of the data operation task in data-aware workflow model $\mathcal{W}$ has to have at least one specific purpose. We define the compliance check for purpose specification principle in Definition 8.

Definition 8 (*Compliance with purpose specification principle*) A data-aware workflow model $\mathcal{W}$ is compliant with purpose specification principle iff $\forall f \in \mathcal{F}^{\mathcal{D}}, \lambda_2(f) \neq \emptyset$.

6.2 Checking compliance with data minimization principle

According to data minimization principle, healthcare providers should access the personal data only if it is required to perform their duties. We have data minimization privacy policy rules $\mathcal{PD}$ which express the amount of personal data which should be revealed and processed. In order to check the compliance of a data-aware workflow model $\mathcal{W}$ with data minimization privacy principle during design time; we can check whether the data used by data operation tasks is necessary for the stated purpose according to $\mathcal{PD}$. We define the compliance check with data minimization principle during design time in Definition 9. It simply checks whether there is a $pd \in \mathcal{PD}$ such that its *user* component (pd_1) is equal to the *user* component ($laps(f_1)$) in $\mathcal{W}$, its *purpose* component (pd_2) is equal to the *purpose* component ($\lambda_2(f)$) in $\mathcal{W}$, and its *data* component (pd_3) is a superset of the *data* component ($\lambda_1(f)$) in $\mathcal{W}$.

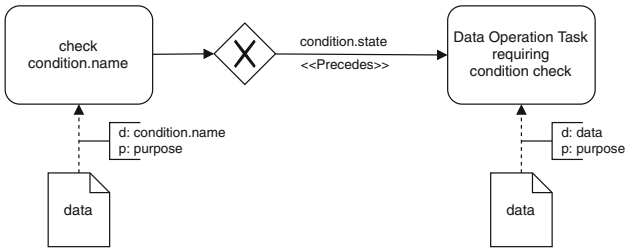

Fig. 6 Condition check pattern

Our compliance check definitions are built on top of our PaCW ontology. We used PaCW to reason the semantic relationships and hierarchy among different elements. In Definition 9 when we state that the *user* component of $\mathcal{PD}$ and the *user* component of $\mathcal{W}$ are equal, it does not mean that they must be exactly the same. These definitions also consist of the subclasses of the classes. For instance; assume there is a privacy policy rule "Any medical staff can access any sensitive data". According to this rule, a nurse can access any medical data because *nurse* is a *medical staff* and *medical data* is a *sensitive data*.

Some data minimization privacy policy rules require conditions. For instance, *P6: "A pediatrician can access the result of the lab examination only if the result is abnormal."* in Table 1 has a condition formally defined as *examination-result.isAbnormal*, where *condition.name* is *examination-result* and *condition.state* is *isAbnormal*. For such data operation tasks, we must check whether the condition is fulfilled. We design a condition check pattern as it is illustrated in Fig. 6. $<< Precedes >>$ operator under the arc is represented in a way similar to [5]. A *precedes* B means that B can occur only if A occurred before.

Algorithm 1: Condition Check

```
1  def SCAN(W, c, p, cond):
2      sflist ← getIncomingFlows(c)
3      foreach sf ∈ sflist do
4          if sf_1 ∈ G^E ∧ sf.name = cond.state then
5              sf'list ← getIncomingFlows(sf_1)
6              if ∃sf' ∈ sf'list|sf'_1 ∈
               T^D ∧ sf'_1.name.equals('check' + cond.name) then
7                  sf''list ← getIncomingFlows(sf'_1)
8                  if ∃sf'' ∈ sf''list|λ_1(sf'') ⊇ cond.name∧
                   λ_2(sf'') = p then
9                      return true
10         else if sf_1 ! = e^S then
11             SCAN(W, sf_1, cond, p)
12         else
13             return false
```

In order to check whether the condition is satisfied for a data operation task, we define a function SCAN (see Algorithm 1). It takes inputs $\mathcal{W}$ as a data-aware workflow, c as a component, p as a purpose and $cond$ as a condition; and returns $true$ or $false$ as output. It checks whether a condition check step precedes c by traversing all sequence flows.

Definition 9 (*Compliance with data minimization principle*) A data-aware workflow model $\mathcal{W}$ and a data minimization privacy policy $\mathcal{PD}$ is compliant with data minimization principle iff

- $\forall f \in \mathcal{F}^{\mathcal{D}} \implies \exists pd \in \mathcal{PD} : pd_1 = laps(f_1) \wedge pd_2 = \lambda_2(f) \wedge pd_3 \supseteq \lambda_1(f) \wedge \text{SCAN}(\mathcal{W}, c, \lambda_2(f), pd_4) = true$, where $c = f_1$ if $f_1 \in \mathcal{T}^{\mathcal{D}}$ else $c = f_2$.

6.3 Checking compliance with consent check principle

According to the consent check privacy principle, some data operation tasks are lawful only if the data subject has given consent to this processing. We have consent check privacy policy rules $\mathcal{PC}$ which declare what type of data practices require an explicit consent.

During design time, in order to check the compliance of a data-aware workflow model $\mathcal{W}$ with consent check privacy principle; we find out whether a consent check step precedes each of the data operation tasks which require consent. This pattern is illustrated in Fig. 7. Consent check is considered as a special case of condition check in which $condition.name$ is $consent$ and $condition.state$ is yes. In order to check whether a consent check step precedes a data operation task, we use SCAN function with the condition $consent.yes$. We formally define the compliance check for consent check principle during design time in Definition 10.

Definition 10 (*Consent check during design time*) A data-aware workflow model $\mathcal{W}$ and a consent privacy policy $\mathcal{PC}$ compliant with consent check principle iff

- $\forall f \in \mathcal{F}^{\mathcal{D}} : (\lambda_2(f), true) \in \mathcal{PC} \implies \text{SCAN}(\mathcal{W}, c, \lambda_2(f), consent.yes) = true$, where $c = f_1$ if $f_1 \in \mathcal{T}^{\mathcal{D}}$, else $c = f_2$.

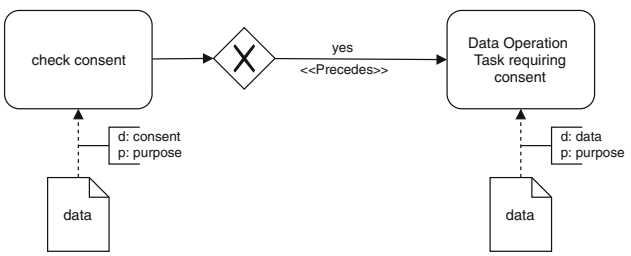

Fig. 7 Consent check pattern

During runtime, we can check whether a data subject gives consent to a specific data operation via checking their privacy preferences. We define the compliance check for consent check principle during runtime in Definition 11. For all input or output data associations, we check whether there exists a positive privacy preference rule of the data subject $\mathcal{DS}$. Some privacy preference rules contain conditions, therefore we also check whether the condition is met through 1 function. The duration of the consent is another important aspect which we looked at. *today − data.entryDate <= duration* checks whether this privacy preference rule is still valid.

Definition 11 (*Compliance with consent check principle during run time*) A data-aware workflow model $\mathcal{W}$, a consent privacy policy $\mathcal{PC}$ and a privacy preference $\mathcal{R}$ for a data subject $\mathcal{DS}$ is compliant with consent check principle iff

- $\forall f \in \mathcal{F}^{\mathcal{D}} : (\lambda_2(f), true) \in \mathcal{PC} \implies \exists r = (\mathcal{DS}, laps(f_1), \lambda_2(f), data, cond, duration, true) \in \mathcal{R}$: data $\supseteq \lambda_1(f) \wedge$ today - data.entryDate <= duration $\wedge$ $\text{SCAN}(\mathcal{W}, c, \lambda_2(f), cond) = true$, where $c = f_1$ if $f_1 \in \mathcal{T}^{\mathcal{D}}$ else $c = f_2$.

6.4 Checking compliance with limited retention principle

According to the limited retention privacy principle, data should be stored no longer than necessary. We have retention privacy policy rules $\mathcal{PR}$ which declare the retention period of data practices.

During design time, in order to check the compliance of a data-aware workflow model $\mathcal{W}$ with limited retention privacy principle; we find out whether a retention check step precedes each of the data operation tasks. Retention check can be considered as a special case of condition check in which $condition.name$ is $data.$'$retention$' and $condition.state$ is *today − data.entryDate <= data.retention*. In this formalization, " symbol used to define texts. For example; *retention* is a variable, but '$retention$' is used as a text, not a variable. Figure 8 illustrates the retention check pattern for the retention privacy policy rule P2.

P2: "A hospital can save results for the purpose of hearing screening with a retention limit by 3 years."

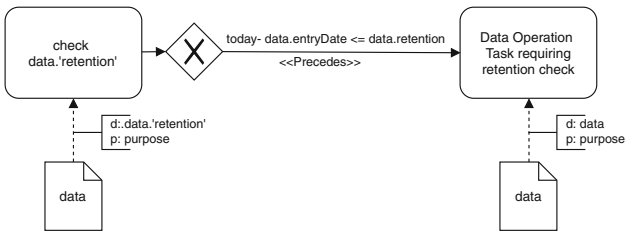

Fig. 8 Example retention check pattern

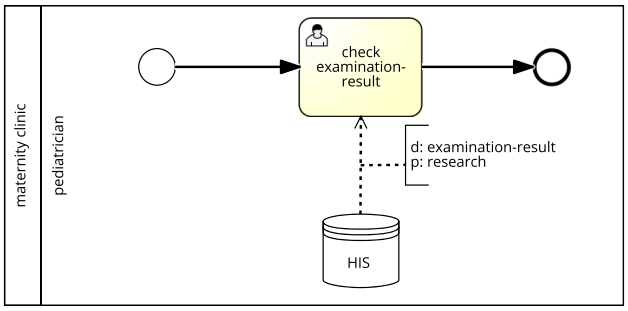

Fig. 9 Newborn screening—data minimization problem before transformation

In order to check whether there is a retention check step before a data operation task, we use Algorithm 1. We formally defined the compliance check for limited retention principle in Definition 12 which is very similar to our 'Compliance with Data Minimization Principle' which is defined in Definition 9. The only difference is the condition input for the function SCAN.

Definition 12 (*Compliance with limited retention principle*) A data-aware workflow model $\mathcal{W}$ and a retention privacy policy $\mathcal{PR}$ is compliant with limited retention principle iff

- $\forall f \in \mathcal{F}^{\mathcal{D}}, pr \in \mathcal{PR} : pr_1 = laps(f_1) \wedge pr_2 = \lambda_2(f) \wedge pr_3 \supseteq \lambda_1(f) \implies \textsc{Scan}(\mathcal{W}, c, \lambda_2(f), cond) = true$, where $c = f_1$ if $f_1 \in \mathcal{T}^{\mathcal{D}}$ else $c = f_2 \wedge$ cond.name = $pr_3.'retention' \wedge$ cond.state = today - pr_3.entryDate <= pr_4.

7 Transformation

When there is privacy violation according to the privacy compliance check, we aim to apply transformation. Section 6 explains the privacy compliance check algorithms. The transformation is intended to work on a set of pre-defined transformation rules. For each of the privacy principles, we worked on possible transformation ways. In this section, the initial transformation ideas will be presented. Since it is an ongoing research, these transformation strategies might alter.

1. Purpose specification In order to be compliant with the purpose specification principle, there should be at least one specific purpose for each data operation tasks. If there is no purpose specified, there is a privacy violation and the transformation action could be returning a message event as a warning to the users which states there should be a specific purpose.

2. Data minimization In order to make sure that users do not hold more data than required; we checked whether the data usage is limited to what is necessary to properly fulfill the purpose via Definition 9. If there is no data specified, there is a privacy violation and the transformation action could be returning a message event as a warning to the users which states there should be a specific data. When user access more data than required, we might warn the user about this violation and limit the data operation according to data minimization privacy policy. Another possible violation can be triggered by not holding the condition. In this case, we might add a condition check step before the data operation task requiring the condition check as it is shown in Fig. 6.

We give an example showing how the transformation approach works for this privacy principle. Figure 9 illustrates a BPMN diagram for newborn screening which consists of a privacy violation because of the noncompliance with the privacy rule *P6: "A pediatrician can access the result of the lab examination only if the result is abnormal."*

Accessing *examination-result* without the condition check results in a privacy violation and the transformation action to solve this problem is to add condition check task. Figure 10 illustrates the BPMN diagram for newborn screening after this transformation. The transformed parts are shown in green.

3. Consent check Some tasks can be legitimate only with an explicit consent of a data subject. If there is a data operation task requiring explicit consent and if there is no consent

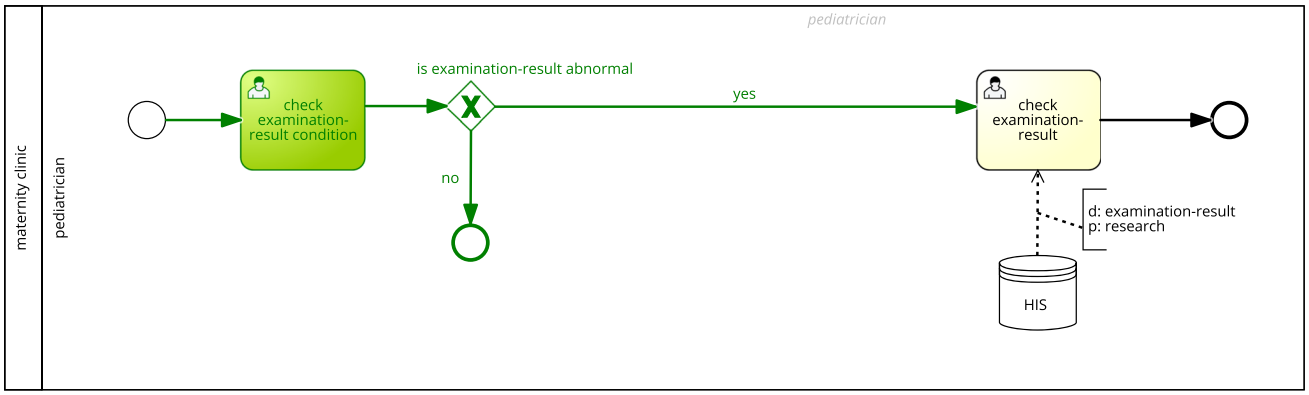

Fig. 10 Newborn screening—data minimization problem after transformation

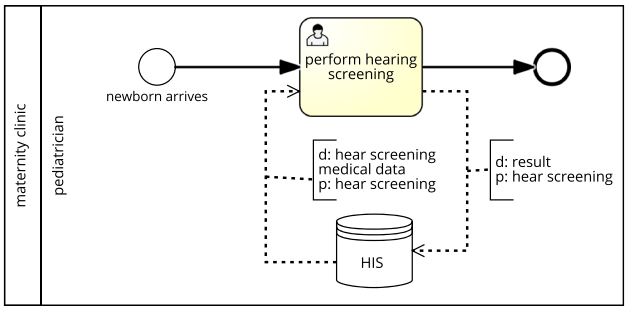

Fig. 11 Newborn hearing screening—consent problem before transformation

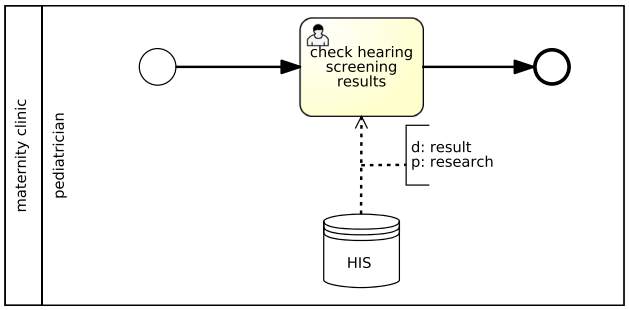

Fig. 13 Newborn hearing screening—retention problem before transformation

check task before that data operation task, there is a possible privacy violation. The transformation action could be adding a consent check task beforehand as it is illustrated in Fig. 7.

We give an example showing how the transformation approach works for this privacy principle. Figure 11 illustrates a BPMN diagram for newborn hearing screening which consists of a privacy violation because of the noncompliance with the privacy rule *P1: "An explicit consent is required for the purpose of newborn hearing screening."*

Accessing data for newborn hearing screening without consent results in a privacy violation and the transformation action to solve this problem is to add consent check task. Figure 12 illustrates the BPMN diagram for newborn hearing screening after this transformation. The transformed parts are shown in green.

4. Limited retention period According to the limited retention privacy principle, data should be stored no longer than necessary. In Sect. 6.4, we explain how to check the compliance with limited retention principle. If there is no retention check step before data operation task, it might result in a possible privacy violation. In order to have a privacy-aware worfklow, we could add a retention check step beforehand.

In GDPR, limited retention principle is closely related to a legal right of data subjects which is called as 'right to be forgotten'.

> [...] the controller shall have the obligation to erase personal data with- out undue delay where one of the following grounds applies: (a) the personal data are no longer necessary in relation to the purposes for which they were collected or otherwise processed; (b) the data subject withdraws consent [...] [Article 17, §1].

It spells out an obligation that the personal data must be deleted when those purpose are no longer applicable or when the data subject withdraws consent. We worked on how to model right to be forgotten. The first method could be erasing data. Apart from erasing the data, there are three general methods to handle the obligation for right to be forgotten. The

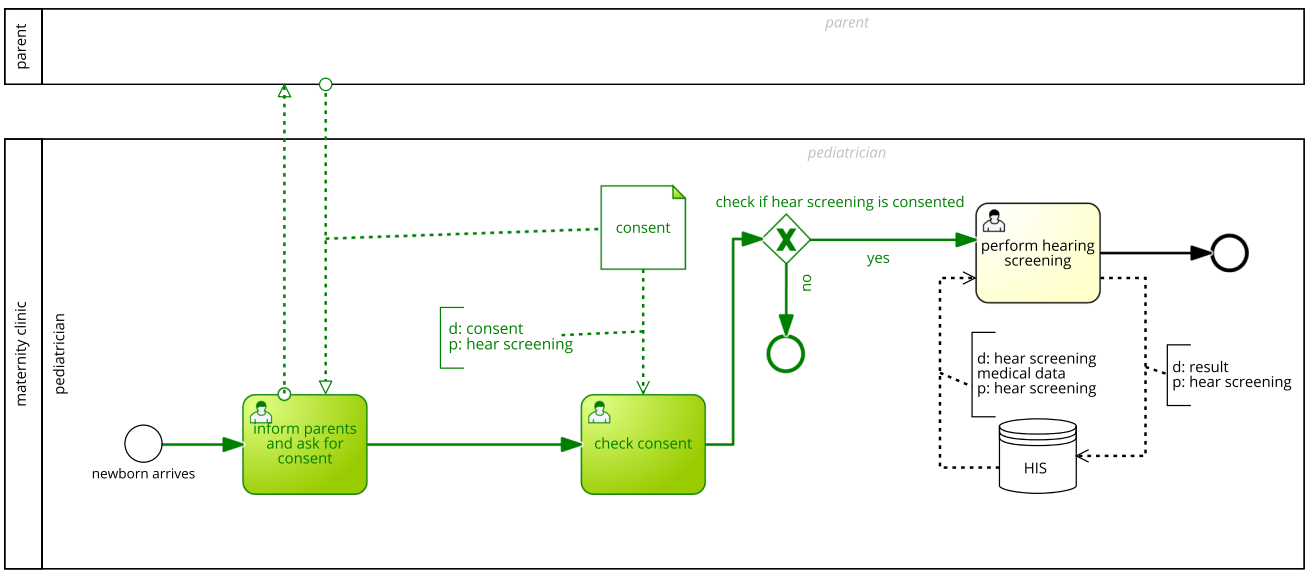

Fig. 12 Newborn hearing screening—consent problem after transformation

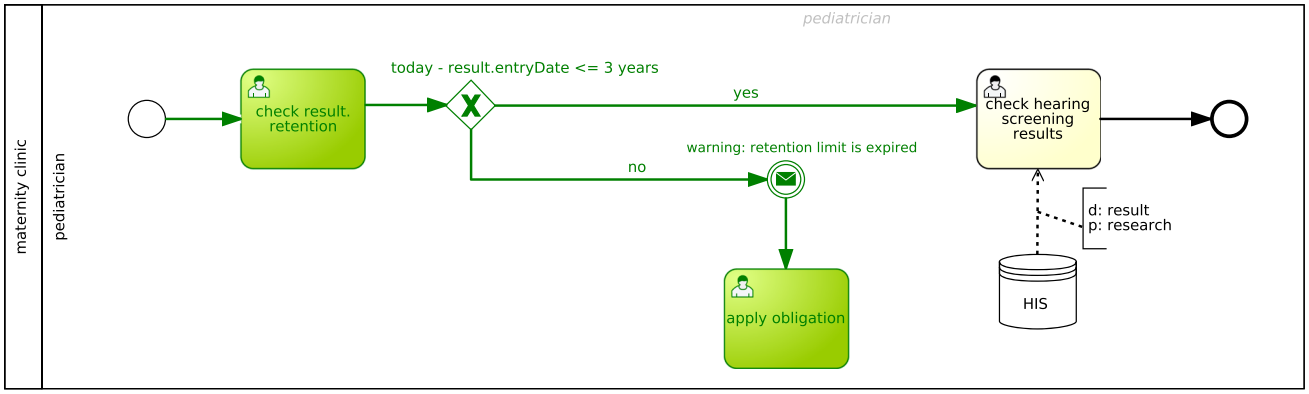

Fig. 14 Newborn hearing screening—retention problem after transformation

methods are stated in [6] as follows: (i) *Denying access to data:* a requester does not have data access. (ii) *Withholding the location of data:* a requester cannot access the data without the right location. (iii) *Rendering data uninterpretable:* the application cannot interpret the data

The transformation action for the obligation could be warning user about the violation and adding a apply obligation task. We give an example showing how the transformation approach works for this privacy principle. Figure 13 illustrates a BPMN diagram for newborn hearing screening which may consist of a privacy violation because of the possible noncompliance with the privacy rule *P2: "A hospital can save results for the purpose of hearing screening with a retention limit by 3 years."*

Accessing data after retention period results in a privacy violation and the transformation action to solve this problem could be to add retention check task. Figure 14 illustrates the BPMN diagram for newborn hearing screening after this transformation. The transformed parts are shown in green.

8 Related work

So far, there is no study that fully outlines transforming clinical workflows into privacy-aware ones automatically. However, there are studies that are related to modeling privacy requirements in business processes [7–9] and in access control systems [10]. In database community, there is significant work on privacy-aware technologies which contribute to our approach for the negotiation of personal data between data subjects whom personal data is about and data processors who processes the data (P3P [11], EPAL [12], Hippocratic data-bases [13]).

Moreover, researchers have proposed enhancements to Hippocratic database systems to ensure minimal disclosure in respect with privacy policies stated by enterprises and customer preferences [14,15]. In [15] customers are asked to express their privacy preferences in the form of privacy penalties associated with each personal data item. The aim is to find the optimal path with the maximum privacy protection with criterion of the smallest privacy penalty. As a future work, we might adapt their "Minimal Disclosure Algorithm" into our privacy preference setting.

In order to to ensure and maintain compliance of business processes models with certain regulations, there are two main approaches which are design-time compliance checks and run-time compliance checks. Kalenkova et al. [16] introduces event-logs and how to use them for the run-time compliance checks. Event-logs can be used to discover what actually happens during run-time. In our research, we focused on compliance check during design time. We would like to examine run time compliance via integrating event logs.

9 Conclusion

In healthcare domain, how to safeguard data privacy of patients challenging. Healthcare providers have to process sensitive medical data compliantly with binding privacy regulations such as the GDPR. In this paper, we give a formal introduction to our approach towards transforming non-privacy-aware BPMN-based clinical workflows into privacy-aware ones. We explained how to detect privacy problems in clinical workflows and try to resolve them, through algorithms that incorporate privacy policies and privacy preferences.

Since it is ongoing research, there are future works. Our transformation algorithm is a work in progress and we contemplate to finalize it. We are also interested in working on demonstrating the soundness and completeness of the transformed privacy-aware clinical workflows. Our initial idea regarding soundness property of transformed workflows is not checking soundness of the workflow after the transformation, but instead, we aim to provide transformations in a way that do not destroy the soundness property of existing workflows.

Acknowledgements This work is supported by DFG Research Group "Service-oriented Architectures for the Integration of Software-based Processes, exemplified by Health Care Systems and Medical Technology" (SOAMED).

References

1. EU General Data Protection Regulation (GDPR). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ 2016 L 119
2. Guarda P, Zannone N (2009) Towards the development of privacy-aware systems. Inf Softw Technol 51(2):337–350
3. European Commission (2015) Special Eurobarometer 431: Data protection. http://ec.europa.eu/commfrontoffice/publicopinion/archives/ebs/ebs_431_en.pdf. Accessed 09 Jan 2019
4. Dijkman RM, Dumas M, Ouyang C (2007) Formal semantics and analysis of BPMN process models using Petri nets. Queensland University of Technology, Tech. Rep
5. Awad A, Decker G, Weske M (2008) Efficient compliance checking using BPMN-Q and temporal logic. In: International conference on business process management. Springer, pp 326–341
6. Vijfvinkel MM (2016) Technology and the right to be forgotten. Master's thesis, Radboud University
7. Mülle J, von Stackelberg S, Böhm K (2011) Modelling and transforming security constraints in privacy-aware business processes. In: 2011 IEEE international conference on service-oriented computing and applications (SOCA). IEEE, pp 1–4
8. Labda W, Mehandjiev N, Sampaio P (2014) Modeling of privacy-aware business processes in BPMN to protect personal data. In: Proceedings of the 29th annual ACM symposium on applied computing. ACM, pp 1399–1405
9. Bartolini C, Muthuri R, Santos C (2015) Using ontologies to model data protection requirements in workflows. In: JSAI international symposium on artificial intelligence. Springer, pp 233–248
10. Belaazi M, Rahmouni HB, Bouhoula A (2015) An ontology regulating privacy oriented access controls. In: International conference on risks and security of internet and systems. Springer, pp 17–35
11. Cranor L (2002) Web privacy with P3P. "O'Reilly Media, Inc."
12. Ashley P, Hada S, Karjoth G, Powers C, Schunter M (2003) Enterprise privacy authorization language (EPAL). IBM Research
13. Agrawal R, Kiernan J, Srikant R, Xu Y (2002) Hippocratic databases. In: VLDB'02: Proceedings of the 28th international conference on very large databases. Elsevier, pp 143–154
14. LeFevre K, Agrawal R, Ercegovac V, Ramakrishnan R, Xu Y, DeWitt D (2004) Limiting disclosure in hippocratic databases. In: Proceedings of the 30th international conference on very large databases. VLDB Endowment, pp 108–119
15. Massacci F, Mylopoulos J, Zannone N (2006) Hierarchical hippocratic databases with minimal disclosure for virtual organizations. VLDBJ 15(4):370–387
16. Kalenkova AA, van der Aalst WMP, Lomazova IA, Rubin VA (2017) Process mining using BPMN: relating event logs and process models. Softw Syst Model 16(4):1019–1048

Saliha Irem Besik is a research assistant and a Ph.D. student in Computer Science Department at Humboldt-Universität zu Berlin, Germany. She received a M.Sc. degree in Computer Engineering from Middle East Technical University, Turkey. Her current research interests cover database systems, privacy, business process modeling, and clinical informatics.

Johann-Christoph Freytag Ph.D. (Harvard University) is currently Full Professor for Databases and Information Systems (DBIS) at Humboldt-Universität zu Berlin. Freytag's research interests include all aspects of query processing and query optimization in object-relational database systems, new developments in the area of database systems (such as semi-structured data and data quality), privacy in database systems, and various aspects of big data as well as applying database technology to domains such as GIS, genomics, text, and bioinformatics/life science. He has published extensively over the years reaching almost 100 publications in the area of database systems, data management, query optimization, privacy, bio informatics, and related areas.