German Medical Data Sciences: Bringing Data to Life
R. Röhrig et al. (Eds.)

doi:10.3233/SHTI210061

Feasibility Queries in Distributed Architectures – Concept and Implementation in HiGHmed

Reto WETTSTEIN[a,1], Hauke HUND[b], Insa KOBYLINSKI[b],
Christian FEGELER[b] and Oliver HEINZE[a]
[a] *Department Medical Information Systems, Heidelberg University Hospital, Germany*
[b] *GECKO Institute, Heilbronn University of Applied Sciences, Germany*

Abstract. Medical routine data promises to add value for research. However, the transfer of this data into a research context is difficult. Therefore, Medical Data Integration Centers are being set up to merge data from primary information systems in a central repository. But, data from one organization is rarely sufficient to answer a research question. The data must be merged beyond institutional boundaries. In order to use this data in a specific research project, a researcher must have the possibility to query available cohort sizes across institutions. A possible solution for this requirement is presented in this paper, using a process for fully automated and distributed feasibility queries (i.e. cohort size estimations). This process is executed according to the open standard BPMN 2.0, the underlying process data model is based on HL7 FHIR R4 resources. The proposed solution is currently being deployed at eight university hospitals and one trusted third party across Germany.

Keywords. Feasibility queries, distributed processes, secondary use, medical informatics, BPMN, FHIR

1. Introduction

1.1. Background

Every day, enormous amounts of medical routine data are documented during patient care. This data promises to be of considerable importance for medical research [1, 2]. However, transferring data into a research context and making this data available to a researcher with reasonable effort (i.e. not manually) in adequate time is difficult [3].

To overcome those challenges, the HiGHmed consortium [4] has been established by the German Federal Ministry of Education and Research as part of the Medical Informatics Initiative (MI-I) [5]. At each university hospital involved in the project, a Medical Data Integration Center (MeDIC) based on open standards is currently being established. The goals of each MeDIC are to integrate data of primary medical information systems in one place, to facilitate data transfer into a research context and to provide data for research projects [6, 7]. To be able to use data stored in the MeDICs for

[1] Corresponding Author, Reto Wettstein, Department Medical Information Systems, Heidelberg University Hospital, Im Neuenheimer Field 130.3, 69120 Heidelberg, Germany; E-mail: reto.wettstein@med.uni-heidelberg.de.

research purposes, a researcher must have the possibility to determine by means of feasibility queries whether there is a sufficiently large cohort available for his specific research project. Because cohorts from one organization are often not large enough, a feasibility query should be executed across all organizations participating in the project. Therefore, a concept of a distributed feasibility process, aiming to calculate cohort sizes across multiple organizations has been developed, implemented and tested.

1.2. Requirements

A common high-level process template for feasibility queries has been developed by the MI-I National Steering Committee (NSG) for all four consortia participating in the MI-I [8]. Based on this template, the solution presented in this paper was adapted and refined to meet the requirements of HiGHmed. These requirements are:

1. The process should be automated and not have any centralized components.
2. User management should be independent of the process implementation at each participating organization.
3. Deployment of the process on the HiGHmed framework for data sharing must be possible.
4. Open standards should be used to define an interoperable data model.

In order to meet these requirements, different organization types and necessary communication messages where identified and an open data model standard supporting the process functionality by means of communication and process input/ output values was selected. Finally, the process was implemented, deployed on the HiGHmed framework for data sharing and tested with sample data across three participating organizations and one Trusted Third Party (TTP).

2. State of the Art

To support data exchange between members of the consortium, HiGHmed is developing an open source Data Sharing Framework (DSF) [9]. This framework implements a distributed process engine based on the Business Process Model and Notation (BPMN 2.0)[2] and the HL7 Fast Healthcare Interoperability Resources (HL7 FHIR R4)[3] open standards. Every participating organization runs a FHIR Endpoint accessible by other organizations and an internal Business Process Engine (BPE). The BPE executes BPMN processes in order to coordinate local and remote steps necessary to enable cohort size calculation and data sharing across institutional boundaries.

Architectures and processes for feasibility queries across institutional boundaries already exist. One example is the Clinical Communication Platform implemented by the German Cancer Consortium (DKTK) [10]. Automated requests for case numbers and feasibility of clinical studies are enabled via a central search function, accessing a database with a reduced set of patient data and biomaterial metadata. This database is regularly updated by the consortium members. Decentralized requests are handled via so

[2] https://www.omg.org/spec/BPMN/2.0
[3] https://www.hl7.org/fhir/R4

called bridgeheads deployed at each organization and only released after approval by the local committees.

A similar approach was followed by the trans-European project Electronic Health Records for Clinical Research [11]. Feasibility queries, created by a researcher in the so called workbench, are stored in a central message orchestrator and retrieved using polling mechanism by local endpoints that reside inside a university hospitals network. These endpoints execute the queries and return the results to the orchestrator for aggregation. Final results are then returned to the workbench for presentation to the researcher.

Another example is the Clinical Research Platform implemented by the German Centre for Cardiovascular Research (DZHK) [12], having a central data management system that is supplied by several organizations and updated at regular intervals. This enables feasibility queries across institutional boundaries in a centralized manner.

To the best of our knowledge, no solution exists fulfilling all requirements of HiGHmed. In this paper we present a distributed process for feasibility queries that is independent of the disease under investigation, does not require the use of any centralized data storage components and is fully automated.

3. Concept

Figure 1 and 2 illustrate the BPMN models which consist of three distinct organization types, each shown in a pool. The upper pool represents the leading/requesting organization (the request itself is submitted by a researcher). This organization type coordinates the feasibility query. The lower pool shows all query-receiving organizations. They process the query locally and provide the results to a TTP, represented by the middle pool. The TTP does not hold any data permanently, but rather aggregates the received results and calculates the final cohort size, which is eventually transmitted to the requesting organization.

The feasibility query process consists of the following steps:

First, a researcher authenticates himself against a cohort browser, provided locally at his organization. This browser is regarded as an external service which is not part of the process implementation but rather provides the possibility to define the feasibility query and to send the initial message trigger to the BPE in order to start the distributed process. The definition of the query has to contain inclusion and exclusion criteria at least for each cohort that should be analyzed in the research project. In addition, the researcher can configure his request using two parameters. He can define whether patient consent checks (Figure 1) and/or privacy preserving record linkage (Figure 2) should be performed. After that, the browser should convert the inclusion and exclusion criteria of each requested cohort into a query that can be executed on the MeDIC repositories.

The next process step identifies the query-receiving organizations that will be involved in the calculation of the cohort size. Currently, the request is sent to all members participating in the infrastructure because of security considerations (see section lessons learned). Afterwards, the feasibility request is forwarded to the TTP and the participating query-receiving organizations. The TTP receives all correlation keys (i.e. one for each participating query-receiving organization) defined by the leading organization. Each key is unique for each feasibility request and query-receiving organization. Only after a result is available for each correlation key or if a timer of five minutes expires, the process can continue at the TTP with aggregation of the results. Simultaneously, the que-

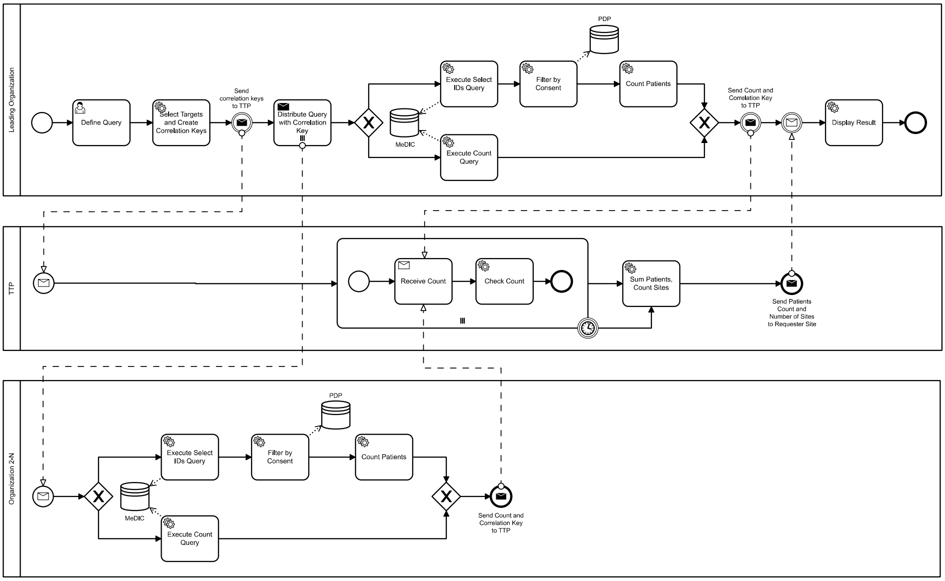
Leading Organization
Define Query
Select Targets and Create Correlation Keys
Send correlation keys to TTP
Distribute Query with Correlation Key
Execute Select IDs Query
MeDIC
Execute Count Query
Filter by Consent
PDP
Count Patients
Send Count and Correlation Key to TTP
Display Result
TTP
Receive Count
Check Count
Sum Patients, Count Sites
Send Patients Count and Number of Sites to Requester Site
Organization 2-N
Execute Select IDs Query
MeDIC
Execute Count Query
Filter by Consent
PDP
Count Patients
Send Count and Correlation Key to TTP

Figure 1. BPMN model of the feasibility process using the optional consent check parameter.

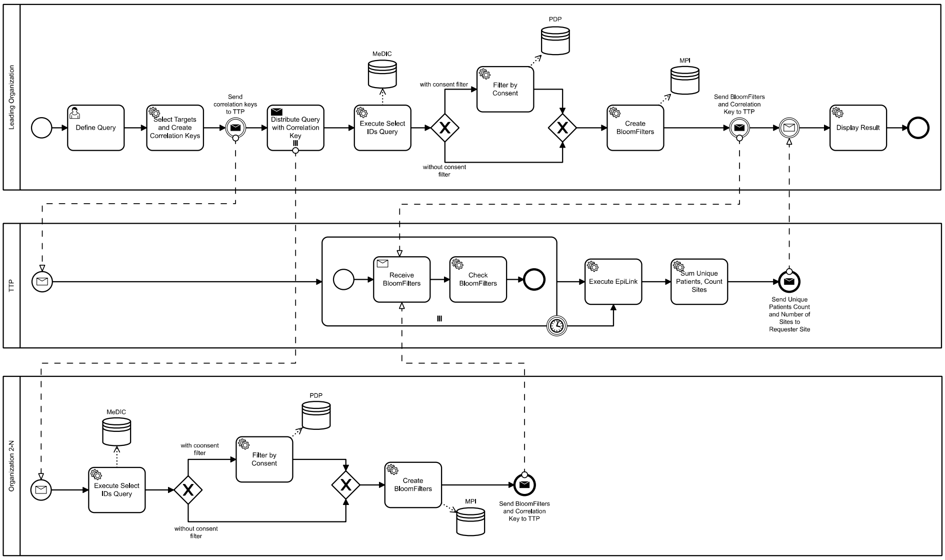

Figure 2. BPMN model of the feasibility process using the optional record linkage parameter.

ry-receiving organizations receive a message containing their individual correlation key and the feasiblity queries.

Following, the request-receiving organizations calculate the cohort size by executing each supplied query and send the results, together with the organizations' correlation key, to the TTP. If no parameter is selected, each query result is calculated using a count query. If at least one of the two parameters is activated, the query is modified to return the Patient Identifier (PID) instead of the count result. In the case that the consent check parameter is activated, it is verified that for each returned PID a rule exists in the Policy Decision Point (PDP), allowing access to the patients' data. All PIDs withstanding this check are summed up and sent to the TTP. If the record linkage parameter is activated, no number is transmitted to the TTP. Instead, the PID is used to retrieve the patients' demographic data from the Master Patient Index (MPI) to generate a privacy preserving bloom filter. All bloom filters are then sent to the TTP.

The TTP temporarily stores all results. When a result is available for each correlation key or the timer expires, the aggregated cohort size is calculated and a range filter is applied. The final result includes the multicentric cohort size and the number of request-receiving organizations that responded with a cohort size bigger than zero. These two results of each query are then transmitted to the leading organization and displayed to the researcher. In the case of the record linking parameter, the TTP executes a linkage algorithm, which uses the bloom filters to check whether a patient is the same person across multiple organizations [9]. The final cohort size of each query is then calculated based on the linkage result and sent to the leading organization, so that the results can be presented to the researcher.

4. Implementation

All data of the feasibility process is modelled using HL7 FHIR R4 resources. For the representation of inclusion and exclusion criteria of each cohort, the FHIR Group resource was selected. Since this resource does not contain a field for mapping the query corresponding to the criteria, an extension was defined. The query language is dependent on the data model used for storage of medical data in the MeDIC repositories. In the case of HiGHmed this is an openEHR repository, therefore an AQL query has to be built based on the defined criteria. Other query languages for other data models, such as CQL for FHIR repositories or FHIR path queries, are alternative options.

The different cohorts are grouped together with a FHIR ResearchStudy resource. This resource adds the possibility to include further information about the research project such as a description and information about the researcher. This information is optional for feasibility queries, but will be mandatory for any subsequent data sharing request. Furthermore, the resource contains a reference to the selected TTP and all query-receiving organizations that are involved in the request, again using an extension.

All messages between organizations are transmitted as FHIR Task resources. This resource consists of an instantiating URI, defining the subprocess that has to be started, and a field representing the current state of the subprocess (i.e. requested, in progress, completed, failed). In addition, information about the sending and the message-receiving organization are stored as references. The input fields of the resource are used to define the data needed to execute a subprocess. The output fields represent either the results of the subprocess or possible errors during execution. Each Task resource, regardless of the subprocess being initiated, contains input fields for a message name, a business key and a correlation key. The business key identifies a feasibility query request. The correlation key is different for each query-receiving organization and used to match results at the TTP to a business key. Depending on the subprocess instantiated by the resource, additional inputs are supplied such as the references to the Group and ResearchStudy resources or the configuration parameters consent check and record linkage. The interval to receive results from individual request-receiving organizations at the TTP was set to five minutes.

For execution of the feasibility process on the DSF, the Task resources are sent to the individual FHIR Endpoints of the organizations and forwarded from there to the BPE. The BPE then executes the subprocess and sends the result to the FHIR Endpoint of the result receiving organization, again using a FHIR Task resource. For the exact structure of the DSF reference implementation and the handling of the FHIR Task resources, the reader is referred to [9].

The open source reference implementation of the feasibility process in the Java programming language and the corresponding (data-) model specifications in FHIR and BPMN can be found on GitHub[4]. Implementation of the feasibility process was possible without changes to the BPE using the included plugin interface.

5. Lessons Learned (Discussion)

All four requirements defined in the introduction could be met. In contrast to already published tools for feasibility queries across institutions (see section state of the art), this version follows a fully decentralized approach without any central component, where all data remains at the recording organization. Each organization can assume a coordinating role. Therefore, a researcher can submit the feasibility query requests at his own organization's feasibility browser instead of using a centrally available service. As a consequence, there is no need to set up a central user management system because already established systems at each organization can be used.

Apart of the initial step by the researcher manually defining the queries in the cohort browser, the process is fully automated.

When submitting a feasibility query, a researcher can configure the request using two parameters. The consent check parameter is introduced because other laws may be applicable to patient data, making it available without patient consent (e.g. in infection control studies). The record linkage parameter has been added to identify patients who have received treatment in several hospitals to ensure they do not appear multiple times in a feasibility result. This parameter is especially necessary for retrospective studies that are concerned with rare diseases and small cohorts. For large cohorts, this calculation-intensive step may be omitted. By using bloom filters and a linkage algorithm, patient privacy can be ensured and identifying demographic patient data does not leave any organization.

Using profiled FHIR resources as data model in conjunction with predefined code systems ensures interoperability on a semantic level. By applying a generic approach to define feasibility queries in FHIR Group resources, different data models and their corresponding query languages, such as CQL or path queries for a FHIR repository or AQL queries in an openEHR repository, can be supported by the process. This means

[4] https://github.com/highmed/highmed-dsf

that the process is independent of the repository data model. Based on the coding of the query, the appropriate repository can be addressed. The automated generation of queries as part of a cohort browser should also be available in the future. An advantage of using FHIR resources to define the cohort definitions and the process tasks is that an audit trail is inherently provided, allowing to track how many requests are made, when requests are made, what organization made how many requests, how often a certain organization responded, etc.

If the optional information about the query submitting researcher in the FHIR ResearchStudy resource is not supplied, the researcher initiating the feasibility process is only known to the leading organization. The request-receiving organizations can only identify the leading organization. Access to a cohort browser used to start the distributed feasibility process needs to be restricted locally for authorized researchers. This decision is made to ensure fast response times and quick cohort size searches across the organizations participating within the infrastructure. If there is a specific reason why the researcher must be known at a request-receiving organisation, a request can be submitted to the leading organization to release this information.

The requirement for sending a feasibility query to all organizations is based on security considerations: If a request could only be sent to a subset of the organizations, an equation system could be built using multiple feasibility queries to track how many patients are treated for a certain disease by one organization, as long as the requesting researcher knows the cohort size at his own organization. This information could be used for economical bench marking of the organization. However, this is not intended in the case of a process execution where the requesting researcher is not known by all organizations and is therefore prevented by sending the request to all organizations.

The range filter is a similar security measure ensuring again, that in case of small cohort sizes or only a few responding organizations, no conclusions can be drawn about the individual case numbers at each organization.

After deployment on the DSF, the process could be tested for a small dataset with three participating organizations and one TTP. It is unknown, how response times will vary for larger data sets and more participating organizations. This also depends on the hardware resources supplied to DSF deployments.

Finally, this process should serve together with another NSG template [8] as a starting point for designing and implementing the actual data sharing process in HiGHmed. It will include steps for decision making by a use and access committee, the negotiation of a data use contract with the researcher and the encryption and merging of data from different organizations.

6. Conclusion

This paper proposes a solution for automated and distributed feasibility queries calculating cohort sizes across multiple institutions using the open standards BPMN 2.0 and HL7 FHIR R4. Two parameters *consent check* and *record linking* allow researchers to configure requests according to the research projects needs. With this implementation there is no need for any central data storage component. Therefore, identifying data does not have to leave the institutions and data privacy regulations are acknowledged. The feasibility process will be deployed at all HiGHmed organization in the near future. Due to the generic approach and the provided open source reference implementation, this process can also be used independently outside the HiGHmed consortium.

Acknowledgements

The project is funded by the German Federal Ministry of Education and Research (BMBF, grant id's: 01ZZ1802A and 01ZZ1802E).

The authors would like to thank all committers that contributed to the open source reference implementation and to test the current release.

Conflict of Interest

The authors state that they have no conflict of interests.

References

[1] L.V. Rasmussen, The Electronic Health Record for Translational Research, *J Cardiovasc Transl Res*, **7.6** (2014) 607–614.

[2] PricewaterhouseCoopers, Transforming Healthcare through Secondary Use of Health Data, (2009).

[3] K. Dentler, A. ten Teije, N. de Keizer, and R. Cornet. Barriers to the Reuse of Routinely Recorded Clinical Data: a Field Report, *Stud Health Technol Inform*, 192 (2013) 313–317.

[4] B. Haarbrandt et al., HiGHmed - An Open Platform Approach to Enhance Care and Research across Institutional Boundaries, *Methods Inf. Med.*, 57.S 01 (2018) e66–e81.

[5] P. Knaup, T.M. Deserno, H.-U. Prokosch, and U. Sax, Implementation of a National Framework to Promote Health Data Sharing, *Yearb Med Inform*, 27.01 (2018) 302-304.

[6] S.L. Aguduri et al., Modeling Clinical Data Transformation for a Medical Data Integration Center: An openEHR Approach, *64. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (GMDS)*, DocAbstr.113 (2019).

[7] N. Yüksekogul et al., ETL-Processes for a Medical Data Integration Center – First Experiences from the Heidelberg University Hospital, *64. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (GMDS)*, DocAbstr.112 (2019).

[8] T. Wendt et al. (AG Data Sharing MI-I), Prozessmodelle des Data Sharing im Rahmen der Medizininformatik-Initiative, *unpublished*, (2019).

[9] H. Hund, R. Wettstein, M. Heidt, and C. Fegeler, HiGHmed Data Sharing Framework (HiGHmed DSF), (2020). https://github.com/highmed/highmed-dsf/ (accessed July 17, 2020).

[10] M. Lablans, E. Schmidt, and F. Ückert, An Architecture for Translational Cancer Research As Exemplified by the German Cancer Consortium, *JCO Clinical Cancer Informatics*, 2 (2018) 1-8.

[11] Y. Chen, R. Bache, S. Miles, M. Cuggia, and A. Taweel, SOA-based Platform for Automating Clinical Trial Feasibility Study, *Proceedings of the IADIS International Conference e-Health*, (2013) 87-94.

[12] German Centre for Cardiovascular Research (DZHK), Clinical Research Platform (CRP), (2020). https://dzhk.de/en/research/clinical-research/clinical-research-platform/ (accessed July 17, 2020).