A FEDERATED INFORMATION ARCHITECTURE FOR MULTINATIONAL CLINICAL TRIALS: STRIPA REVISITED

Prototype

Shen, Zhengru, Utrecht University, Utrecht, the Netherlands, z.shen@uu.nl

Meulendijk, Michiel, Utrecht University, Utrecht, the Netherlands, m.c.meulendijk@uu.nl

Spruit, Marco, Utrecht University, Utrecht, the Netherlands, m.r.spruit@uu.nl

Abstract

The Systematic Tool to Reduce Inappropriate Prescribing (STRIP) is a clinical intervention method crafted to deal with polypharmacy problems which are incurred by the concurrent use of multiple drugs. STRIP has been proven to be effective and is included in the Dutch national guideline for polypharmacy. To boost the usage of STRIP in clinical practices, a web application called the STRIP Assistant (STRIPA) was developed and further evaluated as user-friendly, efficient and effective by Dutch physicians. STRIPA has now evolved into a software tool that supports a large multinational randomized clinical trial (RCT). However, in order to successfully implement and use the application in such an RCT, several issues, including multilingual support, clinical data security, data accessibility and consistency, need to be addressed. In this paper, we present an overhauled STRIPA prototype with an lightweight data integration component that supports multinational implementations, ensures data consistency across countries, and maintains data accessibility and security. The component includes a high-level information architecture, data models redesigned to generalize data entities from all countries, and the ETL processes that integrate diverse data sources and transfer data between databases. Technical features of the application are tested during the implementation, and it also is about to be evaluated empirically as part of the RCT across four European countries.

Keywords: data integration, federated database architecture, ETL, medication review, polypharmacy.

1 Introduction

Polypharmacy, defined as the chronic use of multiple medications by a patient, often leads to severe clinical problems or accidents if it is inappropriate, including adverse drug effects, under prescribing, overtreatment, low patient compliance and decreased drug adherence (Björkman et al., 2002; Claxton et al., 2001; Munger, 2010; Steinman, et al., 2006; Wright, et al., 2009). The elderly, generally adults aged over 65 years, are the main sufferers of polypharmacy (Munger, 2010). The demographic shift towards an elderly population among developed countries brings necessity and urgency to the implementation of an effective solution to reduce the risks and maximize the benefits of polypharmacy.

The STRIP assistant (STRIPA) is a web application that was built to help physicians to perform medication reviews in the case of polypharmacy. A standardized medication review method named Systematic Tool to Reduce Inappropriate Prescribing (STRIP) is incorporated into the application. Meulendijk et al. presented main components of the application, including functional architecture, user interface, decision rule engine, and data formats (Meulendijk et al., 2015a). It has gained recognition as a user-friendly, effective and efficient tool among Dutch physicians (Meulendijk et al., 2015b; Meulendijk et al., 2016). Now, it is a part of a large-scale randomized clinical trial (RCT) that aims at investigating impacts of medication optimization in the multi-morbid elderly among four European countries: the Netherlands, Switzerland, Ireland and Belgium.

In order to implement and use the application across countries, relevant data from each country need to be collected and properly managed in the first place. These data contain medical data on medications, drug interactions and standardized medical terminology, and patient health records. Besides, usage data recorded in each country should be accumulated for further analysis. As the number of data sources grows, the complexity and diversity also increases. How to properly integrate and manage such diverse data sources becomes a challenge.

In this paper we present an overhauled prototype that deals with the above data integration issues so as to support conducting medication reviews in multiple countries. The prototype, named as STRIPA.EU, expands the previous version of STRIPA with a data integration component which mainly consists of a number of databases and ETL processes that populate the databases. The component is able to integrate necessary heterogeneous data sources from different countries, easily update the data as data sources renew, and gather country-specific usage data together into a centralized database. In general, the application works independently as a customized application for systematic medication reviews in each country. However, it also has a centralized database that maintains data consistency across countries and subsequently keeps research integrity and consistency of the large international RCT.

The remainder of the paper is organized as follows. At first, related work and the detailed design objectives of the present research are discussed. Second, the data integration component that extends STRIPA to multiple countries is elaborated from three aspects: data representation, integration architecture and the ETL implementation. Preliminary results of the implementation of STRIPA.EU are briefly discussed in the following. The final section presents contributions of this research and possible future research aspects.

2 Status Quo and Design Objectives

2.1 Status Quo of Data Integration

Data integration aims at providing a unified view of heterogeneous data residing at different sources (Calì et al., 2002). There are three steps in data integration: extracting data from homogeneous or heterogeneous data sources, transforming data into a common structure and loading it into the final target (Muilu et al., 2007). These steps are commonly shortened as ETL. Data integration has become of great importance in the medical/clinical domain because of the tremendous increasing volume and complexity of data gathered at the medical community (Branson et al., 2008, Brazhnik & Jones, 2007). Many studies regarding integrating heterogeneous data sources in the areas of biomedical informatics and information technology have been conducted over the last decades (Karasavvas et al., 2004, Louie et al., 2007, Seoane et al., 2013). However, there is little literature focusing on integrating clinical data sources, especially pharmaceutical data from various countries. But the abundance of information about drugs and its heterogeneous nature among countries does pose a challenge to data integration (Meulendijk et al., 2015a).

Louie et al. (2007) divided data integration into two axes: integration architecture, and data representation. Integration architecture refers to the final destination into which data and metadata are loaded. There are two main types of integration architecture: data warehouse and federated database system (FDB) (Louie et al., 2007). The need for combining various independent data sources into a secure, but easily accessible data architecture has given rise to the development of federated database systems (Calì et al., 2002, Ziegler & Dittrich, 2004). A FDB system has a reliable infrastructure for data sharing and collaboration across disparate data sources while maintaining security and privacy locally. A number of aspects of such a system, such as privacy, security, access control and availability, have been well-studied both in research and in industry (Ansper et al., 2013).

In relational databases, data representation refers to a conceptual data model which represents the common schema of diverse data sources (Louie et al., 2007). For data integration, data models give a unified and structured view of the integrated and reconciled data sources, and also offer the elements

for expressing the queries over integrated systems (Calì et al., 2002). Identifying such a data representation from heterogeneous data sources is mostly time consuming and requires extensive human interactions (Zhao & Ram, 2007). Over the last three decades numerous methods have been proposed to facilitate this process (Batini et al, 1986, Clifton et al, 1998, Passi et al., 2002, Ramesh & Ram, 1995, Zhao & Ram, 2007). Data models for STRIPA.EU are created by detecting schematic correspondences among data sources at both structural and semantic levels.

2.2 Design Objectives

As stated by Meulendijk et al. et al. (2015a), data exchange and integration for STRIPA are difficult because of incompatible information systems and customized international classification standards for drugs and diseases in each country. In particular, since most countries use medical coding standards modified on the basis of international standards, or even completely disconnected ones, data encoded with country-specific codes are not fully interoperable. But decision rules of STRIPA are crafted with international codes, and they only recognize input encoded in such codes. Hence, semantic matching of data sources at instance level is necessary in data integration. Moreover, data of drugs and drug interactions from third-party companies vary a lot in terms of both schemas and formats. Last but not least, multilingualism of data sources adds another dimension of complexity to data integration. Table 1 demonstrates the diversity and complexity of data sources from different countries.

To overcome these issues, we developed a data integration component that is presented in three pillars: data representation, ETL processes and integration architecture. The component is added into STRIPA as a built-in tool. The design objectives when developing the new prototype were as follows: (1) to create a federated database system with a unified schema that supports implementation of the application in multiple countries; (2) to develop ETL processes that populate both a federated database and multiple localized databases in the predefined schema; (3) to regularly update and maintain medical data, including drugs and drug interactions, in each country; (4) to have an architecture which makes the application running in each country on customized local data sources and all the locally collected data remain under the control of the local authorities for the sake of privacy and security; (5) to create a unique identifier for each data item in the system.

Supplier	Database / Standard	Scope	Format	Country	Languages
Z-Index	G-Standard	Medications, Clinical Interactions	Fixed Width	Netherlands	Dutch
RIVM	ICD-10	Episodes	XML	Netherlands	Dutch
APB	Delphi-Care	Medications, Clinical Interactions	Fixed Width	Belgium	French, Dutch
FOD	ICD-10	Episodes	CSV	Belgium	French, Dutch
HCI Solutions	INDEX	Medications, Clinical Interactions	XML	Switzerland	German, French
DIMDI	ICD-10	Episodes	XML	Switzerland	German
HelixHealth	Safescript	Medications, Clinical Interactions	Fixed Width	Ireland	English
WHO	ICD-10	Episodes	XML	Ireland	English
MedDRA MSSO	MedDRA	Adverse Events	CSV	Netherlands, Belgium, Switzerland, Ireland	Dutch, French, German, English
Regenstrief	LOINC	Measures, Laboratory Tests	CSV	Netherlands, Belgium, Switzerland, Ireland	Dutch, French, German, English

Table 1. Data sources from different countries

3 Data Integration Component

Details of the data integration component that supports and realizes the design objectives are described in continuation. It is elaborated from three main aspects: data representation, integration architecture, and the ETL processes.

3.1 Data Representation

As discussed in the previous section, determining data models that represent a common schema of all heterogeneous data sources is crucial for data extraction and transformation at the beginning of an ETL process. There are a great amount of methods or tools that have been developed for detecting schema-level correspondences. In this paper we identified common schema elements (i.e., data entities, relations and attributes) by matching schema elements of all the available data sources at both structural and semantic level. Meanwhile, unnecessary elements are filtered out from the identified common schema according to the system requirements.

To begin with, data required for the application are summarized into two categories: medical knowledge and patient data. Medication data, medical codes and decision rules constitute the knowledge base of STRIPA. As the core data component, medication data contain all available drugs, drug interactions and adverse events in a country. Decision rules in the application are implemented based on established clinical guidelines. Hospitals register a variety of information about a patient on patient health records, but STRIPA only requires patients' medical data on treatments, diagnoses and measurements. Moreover, patient health records need to be encoded with standardized medical codes because both decision rules and drug interactions are composed with standardized codes and they can only recognize encoded patient's information.

As shown in Table 1, medication data are provided by four different data suppliers. By matching data entities of all data sources, four common data entities are derived: medication, product, substance and interactions. Medication stores generic drugs and their substances are saved in substance. Besides, each generic drug might have a list of products in the market and this information is recorded in product. Interactions between two generic drugs are put into interactions. Figure 1 shows the relationship between these data entities and critical attributes of each data entity are also listed.

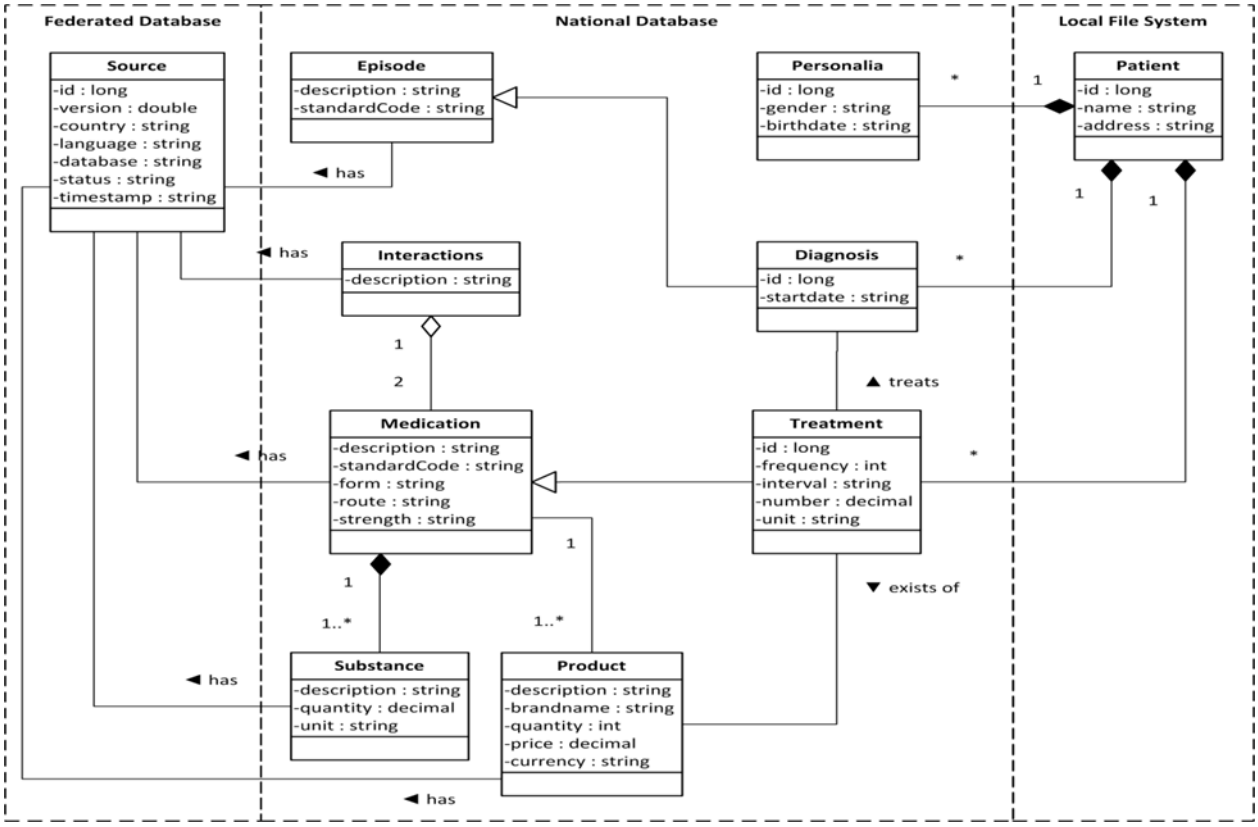

Figure 1. Class Diagram depicting main data models of STRIPA

Patient health records are represented by three data entities: diagnoses, measurements and treatments. Standardized medical codes that are utilized to encode patient health records include LOINC, SNOMED, modified and translated versions of ICD-10 in each country. They are represented with four data entities: *episodes*, *dosage-units*, *dosage-intervals* and *measures*. *Episodes* is used to instantiate *diagnoses* with ICD-10 codes. For instance, one record of patients' diagnoses refers to an episode identifier given the patient's number and a timestamp. *Measures* is extracted from LOINC and it relates to patients' *measurements* the same as *episodes* to *diagnoses*. Together with *medications*, *dosage-intervals* and *dosage-units* generate instances of *treatments* of patients. Figure 1 demonstrates how patient health records are modelled with three layers of data entities. Splitting patient health records into a set of data entities reduces data redundancy in the first place because of database normalization. It also provides an easy way of updating medical codes in each country. For example, adding new ICD-10 codes into *episodes* requires no modification for *diagnoses* and the rest of database.

Since each country regularly renews its medication data by adding new medications or removing old ones, medication data in the application need to be updated accordingly. In order to keep track of medication data during a series of updates, a unique identifier for each data item is created and recorded in sources before data items are added into relevant databases. Table *sources* also stores other metadata which keeps medical knowledge data manageable. Figure 1 shows attributes of *sources* and relationships between *sources* and other data entities. In addition, to maintain the privacy of patients patients' identification information, such as name, address, is stored as *patient* in local file system within hospitals, and the rest are put into *personalia* in national databases. They are linked with a unique patient identifier.

3.2 Integration Architecture

Integration architecture provides an outline of databases in a system from a data integration perspective. Data flows are also included to give a dynamic view of how data are transferred in-between databases. Building a reliable integration architecture is crucial for any system with distributed and autonomous databases in different locations. For STRIPA.EU a federated database system was developed instead of a data warehouse for a number of reasons. At first, unlike a data warehouse, the database in each country is not a subset of a centralized database, whereas each contains all data necessary for the application and takes nothing from the centralized one. Secondly, patient health records collected locally need to remain under the control of the local authorities for the sake of privacy and security. Only collected research data in each country are allowed to be transferred into the centralized database. Finally, a federated database system grants autonomy to the database in each country, which guarantees national databases having power to adapt themselves in response to changes.

Figure 2 depicts the integration architecture that was designed for STRIPA.EU. It is a federated database system which is composed of a federated database and a number of autonomous national databases. As shown in the figure, each country has its own national database, and private patient data is extracted and saved in the local file system within hospitals. Physical independence of the local file system prohibits unauthorized access and use of locally collected private patient data in each country. National databases are autonomous and independent of each other. The autonomy of national databases ensures that countries could customize their databases from various aspects such as schema and data instances. Because of the high independence of national databases, the application in each country has a high system availability. In particular, database breakdown in any country does not affect other countries. The functions of the federated database are twofold: storing and managing metadata for data in all countries, and gathering user log data for further user behaviour analysis (Pachidi et al., 2014). There are two ETL processes for the implementation of STRIPA.EU in each country. The first ETL process takes the aforementioned external data sources and transfers them into a national database while writing metadata into the federated database. Another ETL is designed to accumulate application usage data from all countries. A detailed description of ETL processes will be given in the following.

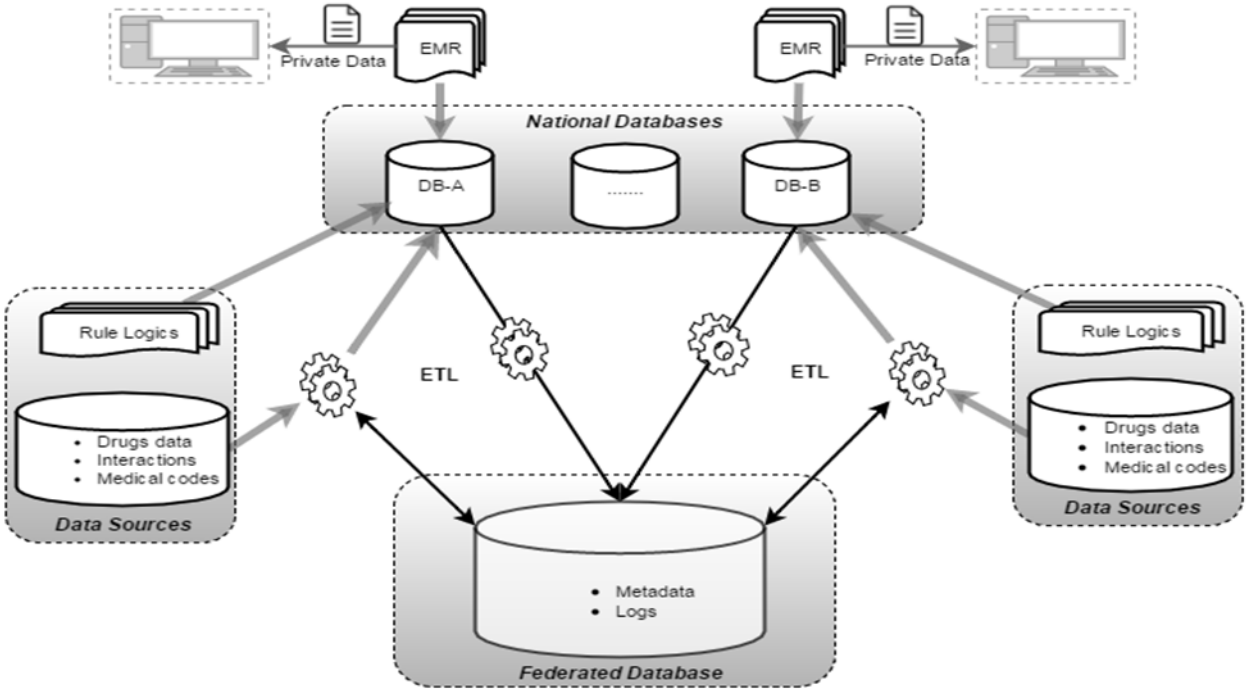

Figure 2. Information Architecture for data integration in STRIPA.EU

3.3 ETL

Determining data models and designing integration architecture only lay the groundwork for the federated database system. In order to provide the client side with a reliable data service, all relevant data need to be imported and heterogeneous data sources, including the most important medication data, need to be transferred using well-designed ETL processes. Therefore, ETL plays a decisive role in the implementation of STRIPA.EU.

There are a number of challenges ahead when it comes to constructing a successful ETL implementation: 1) significant dissimilarity between external data sources and target data structure; 2) external data sources are encoded in diverse formats, including CSV, fixed width and XML; 3) ETL needs to be autonomous, preferably just a click away. A powerful open source tool (Talend Open Studio for Data Integration) has been used to operationalize the ETL processes. It is capable of tackling a variety of data formats. And its graphic interfaces make ETL easier to follow and evaluate by non-IT experts. In the following paragraph an example of the ETL processes developed for medication data in the Netherlands is elaborated to give a glimpse of the complexity.

Medication data in the Netherlands are provided in positional flat files by the G-Standaard. There is a significant structural difference between data sources and the above data model. In specific, generic medications needed in the national database do not exist in data sources, whereas only medicinal products, coupled with product relevant information, such as product name, form, route, etc. are stored separately in different files. ETL should generate generic medications by grouping medicinal products based on their substances. Figure 3 conceptualizes the ETL process with the Process-Deliverables Diagram (Weerd & Brinkkemper, 2008).

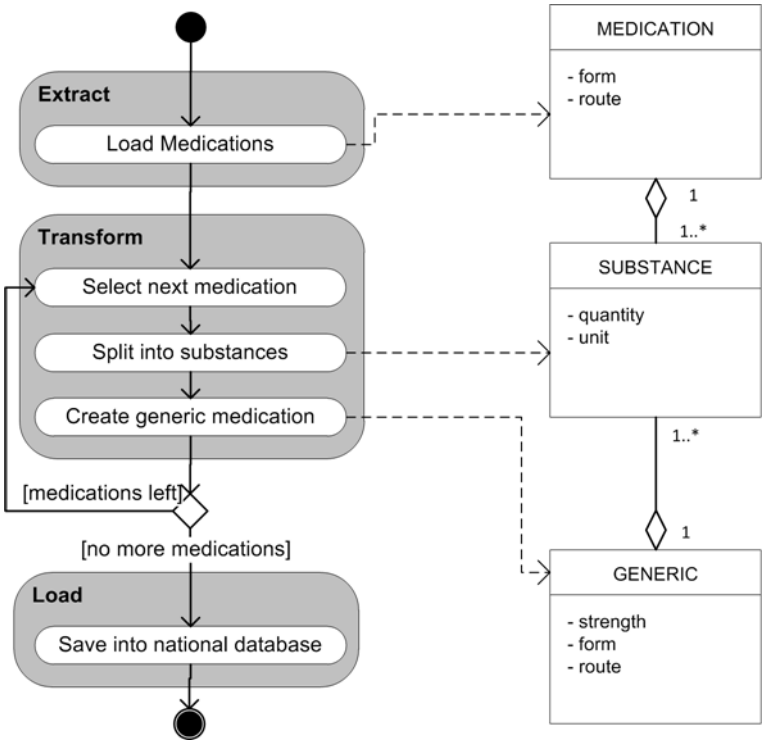

Figure 3. The ETL process of extracting medication data in the Netherlands

As mentioned above, the regular updates of medication data require us to renew drugs and drug interactions within the system accordingly. And the updating process has been built as a part of the ETL implementation. The following process deliverable diagram illustrates the process of updating a database with its new released external data sources. It is coded in a python script and automated with a batch file that loads and directly runs the script.

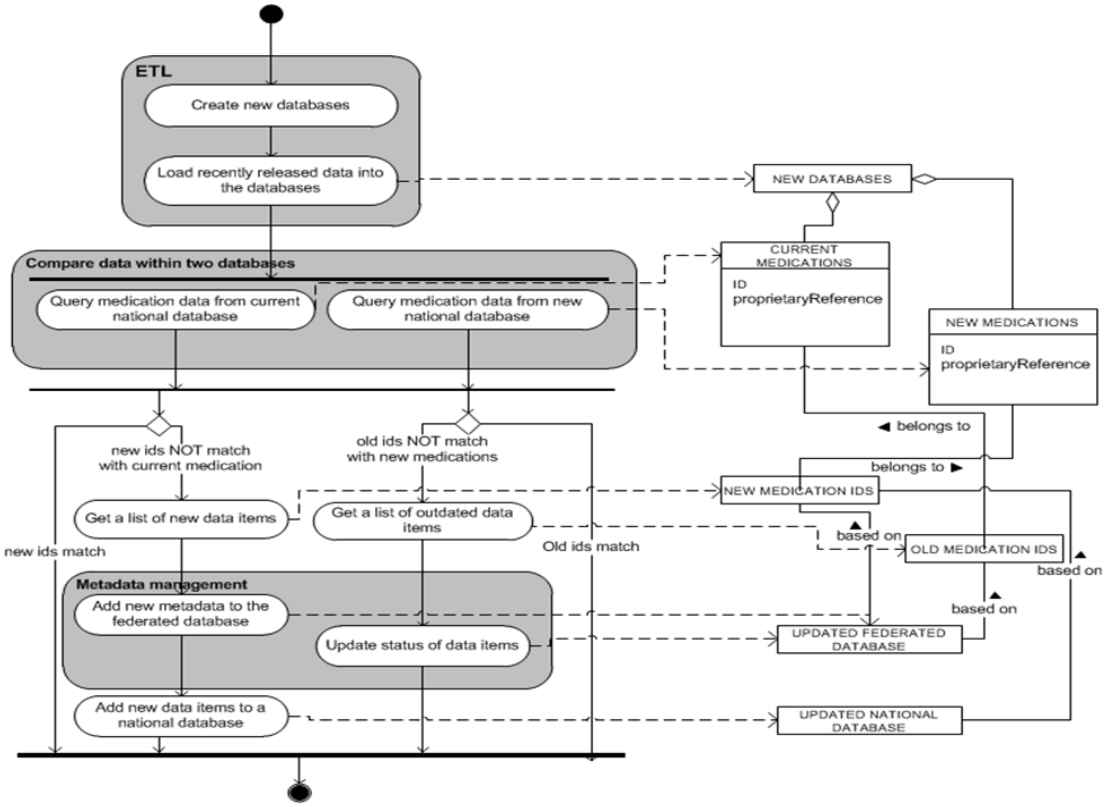

Figure 4. Process Deliverable Diagram depicting the data updating process

4 Evaluation & Contribution

Evaluation of the application has been carried out or is currently planned in multiple steps. At first, ETL processes are tested on data completeness, data consistency and data integrity. During the development of ETL processes, random manual comparisons of data between databases and data sources have been performed. More systematic testing tools and methods will be developed to assess ETL performance in terms of efficiency and maintenance before the implementation of the software. The next step of the evaluation focuses on the integration architecture. How well this architecture supports the implementation and use of the application in multiple countries is investigated. Security and privacy issues of the clinical trials are properly addressed with the architecture. Moreover, having integrated data sources from four data sources in the given architecture shows its technical feasibility and suitability for such multinational clinical trials.

The successful development of STRIPA.EU and initial feedback collected from physicians suggest that the proposed data integration component fulfils the requirements of multinational clinical trials. These requirements, such as healthcare interoperability, privacy and security, are very common among many other large-scale clinical trials. Therefore, this leads us to believe that the federated information architecture and data integration methods have a good potential to be reused in other clinical trials.

Acknowledgements

This work is part of the project “OPERAM: OPtimising thERapy to prevent Avoidable hospital admissions in the Multimorbid elderly” supported by the European Commission (EC) HORIZON 2020, proposal 634238, and by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 15.0137. The opinions expressed and arguments employed herein are those of the authors and do not necessarily reflect the official views of the EC and the Swiss government.

References

Albrecht, A., & Naumann, F. (2008). Managing ETL Processes. *NTII*, 8, 12-15.

Ansper, A., Buldas, A., Freudenthal, M., & Willemson, J. (2013). Protecting a Federated Database Infrastructure Against Denial-of-Service Attacks. *Critical Information Infrastructures Security (pp. 26-37)*. Springer International Publishing.

Batini, C., Lenzerini, M., & Navathe, S. B. (1986). A comparative analysis of methodologies for database schema integration. *ACM computing surveys (CSUR)*, 18(4), 323-364.

Björkman, I., Fastbom, J., Schmidt, I., & Bernsten, C. (2002). Drug-drug interactions in the elderly.*The Annals of Pharmacotherapy*, *36*(11), 1675-81.

Branson, A., Hauer, T., McClatchey, R., Rogulin, D., & Shamdasani, J. (2008). A data model for integrating heterogeneous medical data in the Health-e-Child Project. *Studies in health technology and informatics*, 138, 13.

Brazhnik, O., & Jones, J. F. (2007). Anatomy of data integration. *Journal of biomedical informatics*, 40(3), 252-269.

Calì, A., Calvanese, D., De Giacomo, G., & Lenzerini, M. (2013). Data integration under integrity constraints. *In Seminal Contributions to Information Systems Engineering (pp. 335-352)*. Springer Berlin Heidelberg.

Claxton, A., Cramer, J., & Pierce, C. (2001). A systematic review of the association between dose regimens and medication adherence. *Clinical Therapeutics*, *23*(8), 1296-310.

Clifton, C., Housman, E., & Rosenthal, A. (1998). Experience with a combined approach to attribute-matching across heterogeneous databases. *In Data Mining and Reverse Engineering (pp. 428-451)*. Springer US.

Karasavvas, K. A., Baldock, R., & Burger, A. (2004). Bioinformatics integration and agent technology. *Journal of Biomedical Informatics*, 37(3), 205-219.

Louie, B., Mork, P., Martin-Sanchez, F., Halevy, A., & Tarczy-Hornoch, P. (2007). Data integration and genomic medicine. *Journal of biomedical informatics*, 40(1), 5-16.

Meulendijk, M. C., Spruit, M. R., Jansen, P. A., Numans, M. E., & Brinkkemper, S. (2015a). STRIPA: a rulebased decision support system for medication reviews in primary care. *Proceedings of the European Conference on Information Systems (ECIS) 2015*.

Meulendijk, M. C., Spruit, M. R., Drenth-van Maanen, A. C., Numans, M. E., Brinkkemper, S., Jansen, P. A., & Knol, W. (2015b). Computerized Decision Support Improves Medication Review Effectiveness: An Experiment Evaluating the STRIP Assistant's Usability. *Drugs & Aging*, 1-9.

Meulendijk, M., Spruit, M., Willeboordse, F., Numans, M., Brinkkemper, S., Knol, W., Jansen, P., & Askari, M. (2016). Efficiency of clinical decision support systems improves with experience. *Journal of Medical Systems*, *40*(4), 1–7.

Muilu, J., Peltonen, L., & Litton, J. E. (2007). The federated database–a basis for biobank-based post-genome studies, integrating phenome and genome data from 600 000 twin pairs in Europe. *European Journal of Human Genetics*, 15(7), 718-723.

Munger, M. A. (2010). Polypharmacy and Combination Therapy in the Management of Hypertension-in Elderly Patients with Co-Morbid Diabetes Mellitus. *Drugs Aging*, 27 (11), 871-883.

Pachidi,S., Spruit,M., & Weerd,I. van der (2014). Understanding Users' Behavior with Software Operation Data Mining. *Computers in Human Behavior, 30*, Special Issue: ICTs for Human Capital, 583–594

Passi, K., Lane, L., Madria, S., Sakamuri, B. C., Mohania, M., & Bhowmick, S. (2002). A model for XML Schema integration. *In E-Commerce and Web Technologies (pp. 193-202)*. Springer Berlin Heidelberg.

Ramesh, V., & Ram, S. (1995). A methodology for interschema relationship identification in heterogeneous databases. *Proceedings of the Twenty-Eighth Hawaii International Conference (IEEE)* Vol. 3, pp. 263-272.

Seoane, J., Aguiar-Pulido, V., R Munteanu, C., Rivero, D., R Rabunal, J., Dorado, J., & Pazos, A. (2013). Biomedical data integration in computational drug design and bioinformatics. *Current computer-aided drug design*, 9(1), 108-117.

Steinman, M., Landefeld, C., Rosenthal, G., Berthenthal, D., Sen, S., & Kaboli, P. (2006). Polypharmacy and prescribing quality in older people. *Journal of the American Geriatrics Society*, 54(10), 1516-23.

Weerd, I. Van De, & Brinkkemper, S. (2008). Meta-modeling for situational analysis and design methods. *Handbook of Research on Modern Systems Analysis and Design Technologies and Applications*, 38–58.

Wilkinson, K., Simitsis, A., Castellanos, M., & Dayal, U. (2010). Leveraging business process models for ETL design. *In Conceptual Modeling–ER 2010 (pp. 15-30).* Springer Berlin Heidelberg.

Wright, R., Sloane, R., Pieper, C., Ruby-Scelsi, C., Twersky, J., Schmader, K., & Hanlon, J. (2009). Underuse of Indicated Medications Among Physically Frail Older US Veterans at the Time of Hospital Discharge: Results of a Cross-Sectional Analysis of Data From the Geriatric Evaluation and Management Drug Study. *The American Journal of Geriatric Pharmacotherapy*, *7*(5), 271-280.

Zhao, H., & Ram, S. (2007). Combining schema and instance information for integrating heterogeneous data sources. *Data & Knowledge Engineering*, 61(2), 281-303.

Ziegler, P., & Dittrich, K. R. (2004). Three decades of data integration-All problems solved?. *In IFIP congress topical sessions* (pp. 3-12).