A Generic Data Harmonization Process for Cross-linked Research and Network Interaction*

Construction and Application for the Lung Cancer Phenotype Database of the German Center for Lung Research

D. Firnkorn1; M. Ganzinger1; T. Muley2,4; M. Thomas3,4; P. Knaup1

1Institute of Medical Biometry and Informatics, Heidelberg University, Heidelberg, Germany;
2Translational Research Unit, Thoraxklinik at University Hospital Heidelberg, Heidelberg, Germany;
3Department of Oncology, Thoraxklinik at University Hospital Heidelberg, Heidelberg, Germany;
4Translational Lung Research Centre Heidelberg (TLRC-H), Member of the German Centre for Lung Research (DZL)

Keywords
Lung cancer phenotype data, data integration, research collaboration, research data warehouse

Summary
Objective: Joint data analysis is a key requirement in medical research networks. Data are available in heterogeneous formats at each network partner and their harmonization is often rather complex. The objective of our paper is to provide a generic approach for the harmonization process in research networks. We applied the process when harmonizing data from three sites for the Lung Cancer Phenotype Database within the German Center for Lung Research.

Methods: We developed a spreadsheet-based solution as tool to support the harmonization process for lung cancer data and a data integration procedure based on Talend Open Studio.

Results: The harmonization process consists of eight steps describing a systematic approach for defining and reviewing source data elements and standardizing common data elements. The steps for defining common data elements and harmonizing them with local data definitions are repeated until consensus is reached. Application of this process for building the phenotype database led to a common basic data set on lung cancer with 285 structured parameters. The Lung Cancer Phenotype Database was realized as an i2b2 research data warehouse.

Conclusion: Data harmonization is a challenging task requiring informatics skills as well as domain knowledge. Our approach facilitates data harmonization by providing guidance through a uniform process that can be applied in a wide range of projects.

1. Introduction

Common lung diseases are considered worldwide as a leading cause of death. In particular, lung cancer and the chronic obstructive pulmonary disease (COPD) are among the top ten of observed death cases since 1990 [1]. Affected patients who have to live with respiratory illnesses are suffering from an increasingly diminishing quality of life because of exhausting treatments and breathing ailments [2, 3]. Since the amount of deaths caused by lung diseases has grown over the last 20 years, research networks have been founded to combat new cases and to improve patient care in the future.

In Germany, the Federal Ministry of Education and Research (BMBF) initiated the German Center for Lung Research (DZL) [4] in 2011 as a research network with the aim to examine and treat lung diseases respecting eight disease areas, among others COPD and lung cancer. Also, the DZL was established as a meta-network of five regional networks, like the Translational Lung Research Center Heidelberg (TLRC-H). Each of them is a network of research institutions and (university) clinics, with long-term experience with lung patients. One of the biggest challenges when founding a research network like the DZL is to provide a platform for collaborative work [5]. Joint data analysis is a key requirement for cross-linked research. Within a pilot project in the DZL disease area Lung Cancer (DZL LuCa), the phenotypic data collections of LungenClinic Großhansdorf (ARCN), the Thoraxklinik Heidelberg and University Hospital Munich (CPC-M) are integrated into a central research data warehouse (RDW) to enable cross-site data analysis.

Plenty of phenotype data have been collected at the three DZL LuCa sites over decades in heterogeneous data formats. While the main data source at all sites is the local electronic medical record (EMR) sys-
tem, data definitions are not identical because the EMR systems were built by independent manufacturers and fulfill site-specific requirements. We call the metadata of these data sources source data elements (SDEs). The SDEs have not been harmonized before to support collaborative research. That means metadata, definitions of SDEs, and documentation strategies differ among the sites. Additionally, coding standards are lacking [6] which complicates joint data analysis [7]. Therefore, a harmonization process is necessary to be able to semantically map SDEs of each participating site to common data elements (CDEs), which are approved as consensus for DZL LuCa [8].

There are several approaches to match SDEs from different source systems in the field of Business Intelligence [9]. Nevertheless, there is a need to establish a reference process on how to harmonize and integrate CDEs into a RDW to enable overarching data analyses in medical research networks.

### 2. Objectives

The objectives of our paper are: i) Development of the generic Harmonization Process (gHarP) for reaching consensus on CDEs, ii) creation of a common lung cancer phenotype dataset according to gHarP; and iii) construction of the Lung Cancer Phenotype Database (LCPD) as a RDW for the disease area lung cancer.

### 3. Methods and Tools

We divided the data integration process for LCPD into three phases according to a German white paper by Fraunhofer Institute for Software and Systems Engineering (ISST) [10]:

**Phase 1: Schema matching.** We define schema matching as “the task of identifying semantic correspondences between metadata structures” [11]. The result is a set of correspondences of the SDEs at the three LuCa sites and a set of CDEs which constitute the common lung cancer phenotype dataset.

**Phase 2: Semantic mapping.** In this step the mapping rules are defined. They enable the source data from each site to be automatically transformed into data according to the CDEs.

**Phase 3: Transformation.** This step comprises the complete extract-transform-load (ETL)-process, which integrates the data of the three LuCa sites into LCPD.

We developed gHarP as a systematic approach for phase 1 and 2 which can be applied in a wide range of data integration projects. Since this generic process is an outcome that can be used independent of LCPD we describe the process in section 4.1 of the results part of this paper.

### 3.1 Tool for Supporting the Harmonization Process

As tool for supporting phase 1 and 2 we developed a spreadsheet-based solution, which we call harmonization table (HT). Within the HT, a comprehensive view is given for each SDE from each site and they are listed according to their correspondences to the CDEs. This means the HT is an overarching documentation containing on the one hand all source and target metadata definitions. On the other hand, each site gets an overview of the mapping rules and the transformation of their local data into the target dataset. Furthermore, the HT is used to configure and parameterize an automatic ETL process enabling rapid prototyping and supporting the deployment of the entire project. An example of our HT is provided as supplementary material.

### 3.2 Tools to Set up LCPD

Key requirements for LCPD were data import according to the specifications in the HT with the help of an ETL tool, query execution, sub-cohort extraction by performing queries with selectable CDEs, and export of sub-cohort data to a structured text file for further analysis.

Due to these requirements the opensource RDW Informatics for Integrating Biology & the Bedside (i2b2) [12, 13], has been chosen as tool for cohort management. In phase 3, for data import and export, we utilized the likewise open-source tool Talend Open Studio [14], which is a component-based ETL programming environment, generating Java code for further manual adaptions. For data export we developed a tool called Generic Case Extractor (GCE), which has also been implemented with Talend Open Studio.

### 4. Results

In section 4.1, we describe gHarP for reaching consensus on CDEs and for preparing local data sources. Afterwards, in section 4.2 we describe how we applied the reference process in the disease area lung cancer and give a brief overview of the parameter domains of the resulting common lung cancer dataset. Section 4.3 closes this chapter with the description of the export functions of LCPD.

#### 4.1 Reference Process for CDE Development

The following roles participate in gHarP:

- **Central Data Manager** – The network has to appoint a central data manager who is responsible for the conduction of the harmonization process and for the cooperation with the other stakeholders.
- **Local Data Manager** – Each site of the research network has to appoint a local data manager.
- **Domain Expert** – The domain expert is usually a physician or biologist, who contributes to the process with specific domain knowledge, like oncology etc.
- **Medical Researcher** – The network has been established as a medical research network with medical researchers as principal investigators at each site.
- **IT-Expert** – The IT-expert can support the process by providing tools to support the harmonization process and by supervising their application.

To reach consensus, a data harmonization committee has to be established. In smaller consortia each role from each site should participate in the committee. In larger consortia the representatives of each role should elect two or three delegates as members of the data harmonization committee.
The gHarP consists of the following steps:

Step 1. **Identification of major semantic concepts**: Discussion of potential future research questions and important concepts among medical researchers. Domain experts specify general parameters necessary to describe a common patient cohort. The results should be documented by the central data manager.

Step 2. **Definition of CDEs**: Find a parameter consensus in accordance to the results of step 1. Provide a formal definition of CDEs including definitions of items, data types respectively allowed values. Supported data types are integer, decimal, categorical, date, and string. Since the value domains of string parameters cannot be checked during the ETL process in a later step, it is advantageous to use more constraining data types. To avoid defining arbitrary data models, reusing established standards (e.g. [15, 16]) is preferable. Using standards also helps in the future integration of similar data sets on the same basis from external sources. Otherwise, a leading data set among the participating sites might be a suitable basis. Existing coding standards, such as the Logical Observation Identifiers Names and Codes (LOINC) [17] or the Systematized Nomenclature of Human and Veterinary Medicine Clinical Terms (SNOMED CT) [18] should be used as far as possible to define a CDE for each parameter. If no coding standard provides sufficient definitions or a standard cannot be used (currently, there is no national license of SNOMED CT for Germany, e.g.), a CDE has to be defined from scratch for a specific project. CDEs are discussed inside the data harmonization committee and the results are documented by the central data manager. To reach consensus it might be necessary to have several drafts of the CDEs and discussion rounds.

Step 3. **Communication of CDEs to all stakeholders**: This can take place by sending a paper-based or electronic document to each site. Optionally, it could be provided on a website or central server or by an electronic tool like a metadata-repository [19].

Step 4. **Matching of SDEs**: Local data managers at each participating site list SDEs in correspondence to the CDEs with a systematic metadata description. Matching-discrepancies which can be resolved unambiguously by transformation at local sites are not regarded as semantic conflict. Missing SDEs have to be identified. The result of the matching process has to be communicated to the central data manager. This can be supported by a tool provided by the IT-expert which standardizes the description of correspondences.

Step 5. **Review matching results**: The central data manager reviews and summarizes the matching results of each site and resolves the conflicts in cooperation with the local data managers. This is in analogy to step 4, but semantic conflicts must be resolved. For some of the required source data it might be necessary to extract those from local patient records. Data sources are now prepared for ETL-process of an RDW.

Step 6. **Approval of the common data set**: The consensus on CDEs has to be finally reviewed by the data harmonization committee, approved by all stakeholders and published.

Step 7. **Specification of SDEs**: Definition and communication of the final specification of the data formats of the local data sources by the central data manager in consultation with the local data managers and/or IT-expert.

Step 8. **Preparation of local data sources**: Local data managers adapt the local data sets according to the final CDEs and the requirements of step 7. Transformation of source data can either take place at the respective site by the local data manager or during the central ETL process of the RDW. The final description of the SDEs has to be provided by the local data managers. This is in analogy to step 4, but semantic conflicts must be resolved. For some of the required source data it might be necessary to extract those from local patient records. Data sources are now prepared for ETL-process of an RDW.

If it is planned to add more locations as a part of the RDW in the future, they have to use the results of the harmonization phase and start with step 8. Tools, provided by the IT-expert might be used. Figure 1 illustrates the described steps of gHarP.
4.2 Development of Common Lung Cancer Dataset

4.2.1 Application of Reference Process

As DZL LuCa is a rather small consortium with three sites, the roles central data manager and IT-expert are fulfilled by the same person. In the data harmonization committee each role of each site is represented. The committee members met either in telephone conferences or face-to-face.

In step 1 we performed a face-to-face group discussion between medical researchers and domain experts from ARCN, TLRC-H and CPC-M. The result was the specification of ten parameter domains and the identification of four patient cohorts inside LuCa. This has been documented and communicated by the central data manager to the local data managers at each site.

Within step 2 we identified 285 CDEs inside the parameter domains which describe the four patient cohorts. As there has been no standardized data set for LuCa, the CDEs have been created utilizing LOINC, ICD 10, ICD-O and well-known parameter definitions among the sites. In step 3 the CDE document was provided in a central data repository accessible for each site.

4.2.2 Overview of the Resulting Data Set

With the gHarP approach, we established a lung cancer dataset covering ten domains: location, lung functions, patient information, diagnostics, therapies, laboratory values, toxic exposure, questionnaires, tumor characteristics, and follow-up. Across the domains, 302 CDEs have been defined, whereas 285 could be harmonized. Source data, whose CDEs cannot be harmonized are present in free text, concerning 17 CDEs. For some free text SDEs it is possible to derive structured SDEs with

Steps 4 and 5 served for the determination of all SDEs that match according to the defined 285 CDEs. During these steps, all local data managers examined each source dataset with the contained SDEs and matched them according to their values, representations and measurement units (e.g. synonyms, date formats or centimeter respective meter) to get a global semantic for the resulting CDEs. As an example: The information about a patient's general health condition was regarded important for LCPD by all sites. Two sites document this information according to the Eastern Cooperative Oncology Group (ECOG) and one according to the Karnofski-Index [20, 21]. Therefore, the central data manager defined a mapping rule to map the Karnofski-Index ranges to the ECOG categories. Altogether, the local data managers created a list with more than 544 SDEs arising from the sites for semantic mapping. Inside our harmonization tool HT, the central data manager created a sub-table for each site and mapped the matched SDEs syntactically and semantically to its corresponding CDE inside another sub-table that comprises the target LuCa data set. The results of the mapping activities were discussed among all data managers in five telephone conferences and two in person meetings. This feedback resulted in three revision cycles of the HT.

Step 6: To achieve a harmonized lung cancer dataset, the central data manager defined for each SDE a tuple of nine parameters inside the HT: Source variable name, clinic definition (unique meaning at a site), data type (categorical, integer, decimal, date, string/free text), values, format (e.g. date format), units or ranges, origin document and mapping rule. A CDE tuple is a set of ten parameters wherein some are similar to the SDE tuple, such as target variable name, common definition, data type, values, format and units or ranges. Additionally, it contains the following parameters: dataset (groups CDEs to a domain), concept code (unique identifier of a CDE inside LCPD) and plausibility checks. A field for comments is optional in both, CDE and SDE tuple. The completed HT is adopted by the data harmonization committee and is published within the central LuCa repository.

The sub-tables containing the SDEs of each site in the HT are considered as specification for the data format of the local data sources and their preparation, as described in steps 7 and 8. Adaptions of the local databases have been performed by extending database tables according to missing SDEs. Then, the source data of the identified patient cohorts has been extracted by the local data managers which are then used for data integration inside the developed central ETL process from the IT-expert.

The completed HT is used to parameterize generic ETL processes that have been implemented for data integration (cf. section 4.3).
distinct value domains. If the diversity of free text inside SDEs is too high, this is not possible and their contribution to the target dataset and subsequent analyses was discussed again.

4.3 Lung Cancer Phenotype Database: Data Integration

We implemented a two-dimensional ETL process for data import: The first dimension is responsible for the integration of the metadata structure shown as a hierarchical tree in LCPD. This structure consists of the CDEs defined as the target dataset inside the HT for lung cancer phenotype data. As we are utilizing i2b2 as RDW this tree structure is called i2b2 ontology. The second dimension maps and transforms the incoming source data according to the metadata information provided by each SDE-CDE tuple inside the HT. After the transformation, the data is loaded with the help of a bulk-load process into LCPD. In Figure 2, we give an overview over the described ETL process.

5. Discussion

In this work, we introduced the LCPD for the disease area LuCa inside the DZL as a RDW. We provide gHarP and show its application for the development of a harmonized lung cancer dataset. Our gHarP approach comprises tasks, which are independent from any technical RDW solution. Thus, it can be adopted in other data integration projects and domains. In DZL LuCa we experienced, that the iterative character of this approach enables a very flexible and at the same time rapid development of a harmonized dataset and a data integration project. Agarwal et al. also suggest an incremental data harmonization process, but in a more approximate manner with the help of machine learning tools [22]. However, in DZL LuCa it was important to include domain experts to validate the correctness and semantics of the target dataset.

Our harmonization tool HT is a spreadsheet based solution, which was well accepted by all stakeholders, because the format is in widespread use in biomedical research. The HT evolved to become a comprehensive communication platform among all roles and sites participating in gHarP. The HT promoted critical debates in respect of the own data situation locally but also site-overarching. As it contains all required SDEs from each site gathered during the project, a site could easy recapitulate where they have a more complex representation of SDEs compared to other sites. In this situation, a certain reduction of the complexity was necessary to fulfill the defined requirements for the harmonized LuCa dataset. Additionally, it became obvious at which point a site has an insufficient representation of SDEs according to the defined CDEs. Here, subsequent improvements of the local databases had to be performed.

Nevertheless, data harmonization is a complex and challenging task. In the future, this could be comfortably supported by a distinct software application. The user interface design is challenging and should be at least as intuitive as our spreadsheet-based tool HT. It should also act as a central documentation repository and support the parameterization of automated ETL processes. An approach fulfilling these requirements partially is DataShaPER [23]. The software is provided as an online service and at present it cannot be installed locally. However, DataShaPER does not help with parameterizing ETL processes.

Another approach for data integration or harmonization is the use of dataspaces [24]. In this method, it is possible to bring data from different sources together regardless of their schema with the use of merging routines. This could be a great benefit for dynamic and growing environments, where different data sources must be available very fast for common analyses. But it is unclear, if a semantic-preserving mapping of SDEs to automatically generated CDEs during merging can be guaranteed. This was regarded as an important prerequisite for networked research in DZL LuCa. openEHR is a promising approach for defining CDEs. Nevertheless, since all participating sites were already documenting data in a structured format, we decided to use the already available parameters as starting point for the harmonization process and not openEHR archetypes. A next step of our research will be to compare our CDEs with available openEHR archetypes and to submit the results of the comparison to the openEHR archetype review process. We already did this in the field of neonatology [25]. AT least, establishing RDWs in clinical environments is still a challenging task due to the diversity of data-related characteristics [26] compared to other domains like finance [27].

6. Conclusion

Data integration between different institutions is a challenging process. With the LCPD project we showed a practicable way to perform cross-site data integration and harmonization. The developed lung cancer dataset describes core CDEs, which might be also used at other DZL locations. This is a great advantage, because those datasets can be integrated into LCPD without further harmonization efforts. We also plan to publish the CDEs as a blueprint for databases in the domain of lung cancer. Future work comprises also the integration of genotype data to be able to correlate them with the already gathered phenotype data and to support the understanding of lung cancer also at the genome level.

Acknowledgments

The DZL is funded by the German Federal Ministry of Education and Research (BMBF, Funding number: 82DZL00402). The authors would like to thank the DZL database in the domain of lung cancer. Future work comprises also the integration of genotype data to be able to correlate them with the already gathered phenotype data and to support the understanding of lung cancer also at the genome level.

References