Integrated Data Repository Toolkit (IDRT)

A Suite of Programs to Facilitate Health Analytics on Heterogeneous Medical Data

C. R. K. D. Bauer1; T. Ganslandt2,3; B. Baum; J. Christoph; I. Engel; M. Löbe; S. Mate; S. Stäubert; J. Drepper; H.-U. Prokosch; A. Winter; U. Sax

1Department of Medical Informatics, University Medical Center Göttingen, Göttingen, Germany;
2Medical Center for Information and Communication Technology, Erlangen University Hospital, Erlangen, Germany;
3Chair of Medical Informatics, Friedrich-Alexander-University of Erlangen-Nuremberg, Erlangen, Germany;
4Institute for Medical Informatics, Statistics and Epidemiology (IMISE), University Leipzig, Leipzig, Germany;
5TMF – Technology, Methods, and Infrastructure for Networked Medical Research, Berlin, Germany

Keywords
Clinical data warehouse, controlled vocabularies, secondary use, data integration, information storage and retrieval

Summary
Background: In recent years, research data warehouses moved increasingly into the focus of interest of medical research. Nevertheless, there are only a few center-independent infrastructure solutions available. They aim to provide a consolidated view on medical data from various sources such as clinical trials, electronic health records, epidemiological registries or longitudinal cohorts. The i2b2 framework is a well-established solution for such repositories, but it lacks support for importing and integrating clinical data and metadata.

Objectives: The goal of this project was to develop a platform for easy integration and administration of data from heterogeneous sources, to provide capabilities for linking them to medical terminologies and to allow for transforming and mapping of data streams for user-specific views.

Methods: A suite of three tools has been developed: the i2b2 Wizard for simplifying administration of i2b2, the IDRT Import and Mapping Tool for loading clinical data from various formats like CSV, SQL, CDISC ODM or biobanks and the IDRT i2b2 Web Client Plugin for advanced export options. The Import and Mapping Tool also includes an ontology editor for rearranging and mapping patient data and structures as well as annotating clinical data with medical terminologies, primarily those used in Germany (ICD-10-GM, OPS, ICD-O, etc.).

Results: With the three tools functional, new i2b2-based research projects can be created, populated and customized to researcher’s needs in a few hours. Amalgamating data and metadata from different databases can be managed easily. With regards to data privacy a pseudonymization service can be plugged in. Using common ontologies and reference terminologies rather than project-specific ones leads to a consistent understanding of the data semantics.

Conclusions: i2b2’s promise is to enable clinical researchers to devise and test new hypothesis even without a deep knowledge in statistical programing. The approach presented here has been tested in a number of scenarios with millions of observations and tens of thousands of patients. Initially mostly observant, trained researchers were able to construct new analyses on their own. Early feedback indicates that timely and extensive access to their “own” data is appreciated most, but it is also lowering the barrier for other tasks, for instance checking data quality and completeness (missing data, wrong coding).

1. Introduction

In translational research settings an increasing amount of medical data is being gathered electronically, including both routine clinical documentation as well as the dedicated capture of research data in clinical studies. This development has enabled a multitude of novel applications that leverage these growing data repositories for secondary use [1]. Typical use cases include retrospective analyses for cohort discovery, feasibility studies or protocol optimization [2, 3]. Prospective approaches include the support of patient recruitment in clinical trials [4, 5], the re-use of routine data to support clinical trial execution or pharmacovigilance and disease registries [6]. De-
development of secondary use infrastructures has been pushed forward by several large-scale national and international projects like EHR4CR (Electronic Health Records for Clinical Research) [7], OHDSI (Observational Health Data Sciences and Informatics) [8], PCORnet (Patient-centered Clinical Research Network) [9] and i2b2 (Informatics for Integrating Biology and the Bedside) [10]. While some projects focus primarily on specific use cases (e.g. EHR4CR focus on clinical trial applications), the i2b2 platform has been established as a generic tool for warehousing and querying biomedical data.

The i2b2 framework has been developed as Open Source and was funded by the American National Institutes of Health. It provides intuitive means for iteratively building queries based on a simple concept hierarchy and a generic database schema. While the core platform contains only limited functionality for advanced analyses, many extensions have been developed, including phenotyping [11] and integration of statistics packages like R [12, 13] (currently 33 related projects listed on community wiki [14]).

The i2b2 platform has seen rapid uptake in international secondary use projects and has by 01/2015 been cited in 193 publications listed in PubMed, spanning projects in various medical specialties (e.g. [15, 16]), as well as methodological articles focusing on i2b2-based natural language processing [17–21] and extraction of temporal information from routine clinical data [22], among others. Additionally, it serves as a foundation for specialized software applications like e.g. tranSMART, which is developed in a public/private partnership between the EU-funded eTRIKS project [23] and Johnson & Johnson Inc. tranSMART [24, 25] re-uses the core database schema of i2b2 and extends it with dedicated structures and analysis functions for Omics data.

In the SHRINE project (Shared Health Research Network), i2b2 was extended to support queries across networked of federated i2b2 installations, which facilitates multi-center cohort discovery and disease registry use cases [6, 26, 27].

Notwithstanding the success of the i2b2 platform, several deficiencies have been identified by a requirements analysis (specified in the methods section). These deficiencies include a complex setup and configuration process, the lack of integrated tools for importing data in standard formats, limited inclusion of international standard terminologies and performance bottlenecks [28, 29]. The primary objective of this paper is to overcome these deficiencies by developing the Integrated Data Repository Toolkit (IDRT), a project funded by the German Technology and Method Platform for Networked Medical Research (TMF, BMBF MethInfraNet) [30] providing necessary extensions to the i2b2 platform. It aims to provide clinicians and translational researchers with tools and services for analyzing medical data by means of a central research data warehouse.

### Table 1 Requirements selected for implementation in IDRT project (CDISC = Clinical Data Interchange Standards Consortium, ODM = Operational Data Model, CSV = Comma Separated Value format, SQL = Structured Query Language, ICD10 = International Classification of Diseases version 10, TMF = German Technology and Method Platform for Networked Medical Research, PID = Patient Identifier, BMBF = German Federal Ministry for Education and Research)

<table>
<thead>
<tr>
<th>Requirement</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Tool for simplified setup and administration</td>
<td></td>
</tr>
<tr>
<td>2. Extensible Open source Extraction, Transformation and Loading (ETL) tool</td>
<td></td>
</tr>
<tr>
<td>3. Support for importing study data in CDISC ODM format</td>
<td></td>
</tr>
<tr>
<td>4. Support for importing tabular data in CSV and SQL formats</td>
<td></td>
</tr>
<tr>
<td>5. Support for importing biomaterial data</td>
<td></td>
</tr>
<tr>
<td>6. Frontend for editing i2b2 concept hierarchies</td>
<td></td>
</tr>
<tr>
<td>7. Support for non-US standard terminologies (e.g. ICD10)</td>
<td></td>
</tr>
<tr>
<td>8. Integration with TMF PID-Generator pseudonymization framework</td>
<td></td>
</tr>
<tr>
<td>9. Integration with BMBF Metadata Repository (MDR)</td>
<td></td>
</tr>
<tr>
<td>10. Best-practice documentation for performance tuning and ontology creation</td>
<td></td>
</tr>
</tbody>
</table>

### 2. Methods

#### 2.1 Requirements Analysis

An analysis concerning the gap between the i2b2 platform and community requirements was carried out in 2012 by means of a literature review [31], interviews with six German hospital site representatives already using i2b2, and a requirements gathering workshop with German i2b2 users. A set of 64 requirements were identified, including 10 with high priority selected for implementation in the project (Table 1). The full list of requirements is available at [32].

Additionally, 3 prototypical use cases for i2b2 in an academic setting were identified in which the IDRT requirements are exemplified (Figure 1):

- **Clinical Data Warehouse:** i2b2 is continuously updated with data from multiple routine clinical information systems, using tabular extraction from source system databases. The concept hierarchy makes use of pre-converted standard terminologies where applicable. The TMF PID-Generator is an open source component optionally used for record linkage using identifying data from disparate source systems or for pseudonymization [34, 35]. Data is provided for cohort discovery, recruitment support and retrospective analysis. In this context, i2b2 can also serve as a research query frontend for a pre-existing clinical data warehouse.

- **Research Data Repository:** Data from a completed clinical study is loaded into i2b2, using CDISC ODM (Operational Data Model by the Clinical Data Interchange Standards Consortium) [37] extraction from an electronic data capture system. The concept hierarchy can be manually edited to optimize usability and can be harmonized with a metadata repository (MDR). Data is provided for retrospective analysis.

- **Translational Research Unit:** Data from an ongoing translational research project is continuously loaded into i2b2, using tabular extraction for routine data, ODM extraction for study-specific data and biomaterial extraction for specimen data. Data is provided for quality assurance, data management and prospective analyses.
For the Clinical Data Warehouse use case, a common core dataset was identified that is universally used in Germany due to standardization in §21 of the German hospital reimbursement law [36]. The so-called "Paragraph 21" or §21 dataset includes structured data for demographics, diagnoses, procedures and additional data elements required for inpatient billing purposes. Due to its exact specification and broad availability, support for the §21 dataset could further raise the interest for i2b2 installations in German hospitals. The §21 dataset also mandates use of standard terminologies for diagnoses (ICD10-GM = International Classification of Diseases version 10 with German modifications) and procedures (OPS = German Operationen- & Prozedurenschlüssel, based on ICPM, the international classification of procedures in medicine), both of which are not included in the standard i2b2 distribution.

For the Research Data Repository use case ODM was chosen as it is widely used and provides a comprehensive dataset containing both descriptive data of a clinical trial (visits, form and item metadata) and documented content in a single, well-defined file that can be exported from many electronic data capture systems. The concept hierarchies derived from raw ODM-extracted data, however, are not conducive to end-user querying and need to be manually edited for better usability. Analysis of data elements imported from multiple electronic data capture systems is often hampered by lack of precise definitions, missing annotations or disparate value sets. The German National Metadata Repository (MDR) [33, 37] is currently being established based on the ISO 11179/3:2013 standard to make high-quality data elements, code lists or entire forms centrally available. The MDR provides a REST-based (Representational State Transfer) application programming interface (API) to automate retrieving forms, data elements or validate.

For the Translational Research Unit use case, biomaterial data is additionally imported. Biomaterial typically requires multiple attributes to be associated to each specimen. The generic entity-attribute-value database schema originally used by i2b2 could only associate attributes to patient and encounter entities, but did not support the storage of nested sub-attributes of a given fact. Since version 1.6, i2b2 has been extended with modifier codes to support this additional level of association. However, the ETL (extract, transform, load) pathway needs to be adapted to support modifier codes, and a modifier-aware export plugin needs to be implemented, as the existing i2b2 data export tools don’t adequately process modifier code-assigned data.

2.2 Architecture

A multi-tiered software architecture (Figure 2) was designed to accommodate the use cases. Source data can be imported into a staging area using extensible "extractor" plugins to support various formats. All extractor plugins share a common loading pathway which is re-used across different source data formats. The imported data is split into metadata for the i2b2 concept hierarchy and patient data. Optionally the concept hierarchy can be edited to harmonize data elements and improve usability for end users. Modifications made at this step are applied when staged data is transformed and finally loaded into an i2b2 instance for user querying.
All ETL pathways and individual extraction plugins in this architecture were created through Talend Open Studio. A graphical user interface was realized within the Eclipse Rich Client Platform framework. Both tools are based on the Java programming language.

A Web Service-based SOAP interface (Simple Object Access Protocol) to the TMF PID-Generator was integrated into the ETL pathway, allowing on-the-fly generation of pseudonyms from configurable identifying records in the source datasets.

Talend Open Studio provides an intuitive graphical frontend to visually design complex data flows while at the same time exposing the underlying Java program code for low-level customization. Components designed in Talend Open Studio can be exported as JAR (Java Archive) files that can be re-used in other Java projects. An integrated user frontend called IDRT Import and Mapping Tool was developed using the Eclipse RCP Framework to incorporate both the Talend Open Studio based ETL components and functions to edit i2b2 concept hierarchies. A Web Service-based REST interface to the MDR Metadata Repository was implemented to make concept definitions and value sets available in the Import and Mapping Tool. A plugin for the standard i2b2 Web Client was imple-
mented in JavaScript. This IDRT i2b2 Web Client Plugin leverages the i2b2 API to support a meaningful export of modifier-coded data and gives a quick view/export option for browsing patient data associated with conducted queries.

A generic approach to extract biomaterial data was developed, based on a concept hierarchy covering typical data elements used in biobanking (core specimen attributes based on the SPREC Standard PREanlytical Code, storage hierarchy, additional flexible specimen metadata attributes). The approach requires the implementation of a platform specific “driver” to interface a specific biobank management system. Within the project, a driver was developed for the Starlims Biorepository. Drivers for other biomaterial management systems can be easily developed to transform data into the generic target structure.

Functionality to automate the setup and configuration of i2b2 server instances was implemented separately as a Bash-based shell script. A modular approach was chosen to facilitate support for multiple versions of i2b2, databases and operating systems.

Based on the community requirements gathered at the initial workshop, a set of standard terminologies was defined, including the ICD10-GM diagnosis classification, OPS procedure classification and LOINC laboratory codes. Individual extractors were implemented in Talend Open Studio to convert each terminology from its source format into i2b2 concept hierarchies. A re-useable approach based on CLaML (Classification Markup Language) [38] was used for ICD10-GM and OPS, the other terminologies were implemented using the CSV extractor. Based on existing documentation about the i2b2 database schema and the experiences made during preparation of the set of IDRT standard terminologies, a best-practice document on the creation of i2b2 concept hierarchies was compiled.

i2b2 uses a generic database schema based on an Entity-Attribute-Value approach [39], which allows the import of a wide array of source data without the need to modify the database or software. However, in i2b2 all fact data is stored in a single observation table (called observation_fact), which can lead to performance bottlenecks with large datasets. Several application- and database-level optimizations were devised and tested against a large production dataset (101,060,156 fact records, 19GB raw CSV source data). Tests were carried out on a virtual machine which was reset to its initial state for each test run by means of snapshots to ensure identical environment. A best-practice document was compiled to summarize recommendations for performance optimization of i2b2 installations.

3. Results

An integrated toolkit of administration and ETL functions, terminology extractors and documentation was successfully implemented to close gaps identified in the i2b2 ecosystem. The toolkit has been made publicly available as open source under GNU GPLv2 [40] at the i2b2 community wiki [41] with the i2b2 Wizard licensed under GNU GPLv2 and all ETL created using Talend Open Studio as GNU LGPL.

3.1 Administration and Configuration Tools

The i2b2 Wizard provides an easy-to-use frontend for rapidly setting up a full i2b2 server instance, including the installation of all dependent software modules, setting of necessary configuration variables and pre-loading of default database schemas. Through its modular approach, it currently supports several i2b2 versions (versions 1.6 and 1.7), several widely used Linux distributions (Ubuntu LTS 12.4, 14.04 and CentOS) and underlying database servers (Oracle, MS SQL, PostgreSQL). Configuration functions allow the quick setup of i2b2 projects and designation of users and access permissions. Support for installing and configuring federated i2b2 architectures based on SHRINE has also been implemented. Installations done by the i2b2 Wizard can be adapted to local requirements. For example, it supports custom naming of database schemas in order to simplify integration with databases or to allow sharing of one database by several i2b2 servers. Its modular architecture allows for easy integration of support for other types and versions of software components by duplicating and modifying existing scripts.

3.2 Extraction, Transformation and Loading Tools

The IDRT Import and Mapping Tool provides an integrated graphical user interface that functions as a hub for the ODM and tabular extractors developed in the project (Figure 3). All configured i2b2 databases are shown in a server browser for selection as source or destination of an ETL job. All configuration data is stored for later re-use.

The ODM extractor allows importing ODM version 1.2 and upwards files exported by various Electronic Data Capture system platforms into i2b2 without the need for manual configuration. It leverages the study metadata information contained in ODM files to automatically construct a comprehensive concept hierarchy of study visits, forms, item groups, items and value sets that reflect the structure of the associated electronic case report forms. The extractor also supports repeatable sub-forms, which are implemented using modifier codes.

The tabular extractor enables users to import table-based data either from text files in CSV (comma-separated values) format or from Java Database Connectivity compliant databases. The extractor requires a metadata configuration file for each data source to describe the data types, role-playing columns (e.g. patient or encounter number) and optional column aliases, which can be automatically created by the Import and Mapping Tool. When importing from a database, table metadata is automatically queried in order to pre-assign column datatypes. The tabular extractor generates a basic 3-level concept hierarchy consisting of source table names, column names and values imported from the respective columns. For source columns designated as numeric or text, values are automatically stored in the respective dedicated columns of the i2b2 observation_fact table, and metadata xml definitions are generated in the associated entries of the concept hierarchy.

The biomaterial extractor requires more extensive configuration, as no widely ac-
cepted standard format for biomaterial metadata has yet been established. It is thus not available in the Import and Mapping Tool frontend, but needs to be configured and executed in a Talend Open Studio environment. The extractor provides a pre-configured concept hierarchy containing placeholders for core specimen attributes (e.g. material type, volume), the SPREC standard, storage hierarchy and a branch for project specific metadata. The Starlims Biorepository driver implemented in the project supports extraction of all the attributes from a standard installation of the platform.

For the German §21 standard dataset, a pre-packaged set of configuration files is provided that automates the comprehensive import of all required source files and standard terminologies into an i2b2 instance. Apart from execution through the Import and Mapping Tool frontend, all extractors are provided as separate packages that can be imported into Talend Open Studio. This allows for the implementation of bespoke ETL jobs adapted to local needs as well as the automation of data loading in production environments. In this environment, integration of the TMF PID-Generator into ETL jobs is also possible, to generate pseudonyms during import on the fly or provide record linkage across multiple data sources.

Figure 3 The IDRT Import and Mapping Tool with the Import Browser (1) and the Mapping Editor (2). The Import Browser lists i2b2 projects which can be filled with data via a context menu, prompting a wizard interface for the selected data type. The Mapping Editor is divided into two windows: the source data (left) and the customized target i2b2 ontology (right). In this example biospecimen data (3) was imported with i2b2 modifiers, clinical data (4) from a CSV file, study data (5) via CDISC ODM and the German ICD-10-GM structure (6) with the IDRT routines. The target ontology (7) is based on a study structure imported from the BMBF Metadata Repository. After copying and (re-)structuring the source data, the ICD-10-GM diagnose of a patient from the clinical data (5) was automatically mapped into the full ICD-10-GM terminology (8) using the previously imported structure (6). The source for all the patient data mapped to an i2b2 concept can be viewed in a tooltip (9). The new target ontology can be saved (10) for later modification and uploaded with all the necessary patient data (e.g. observation_fact) to another i2b2 project selected from the Import Browser.
3.3 Terminology Import and Ontology Editing Tools

Customized extractors are provided in the Import and Mapping Tool as well as through Talend Open Studio packages to import the following terminologies from their respective source distributions into i2b2:

- International Statistical Classification of Diseases and Related Health Problems 10th revision German Modification (ICD-10 GM): diseases, signs and symptoms
- Operationen- und Prozedurenschlüssel (OPS): German modification of the ICPM and official classification of operational procedures
- International Classification of Diseases for Oncology (ICD-O-3): tumor diseases
- TNM Classification of Malignant Tumors (TNM): cancer staging and origin
- Logical Observation Identifiers Names and Codes (LOINC): laboratory observations
- Medical Dictionary for Regulatory Activities (MedDRA): adverse event classification
- German diagnosis-related groups (G-DRG): reimbursements for hospital cases
- Standard PREanalytical Code (SPREC): preanalytical coding for biospecimens

An interface to the German Metadata Repository (MDR) is provided through its REST API, allowing users to transfer single items or complete forms or trial structures into an i2b2 concept hierarchy with the Import and Mapping Tool.

Graphical functions for editing concept hierarchies are provided in the Import and Mapping Tool to allow users to restructure source data to better meet research requirements (Figure 3). The editor allows removing irrelevant items, rearranging, renaming and grouping items into folders. Multiple identical data elements imported from separate sources can be merged into unified items. Date fields can be used to assign start and end timestamps and make use of novel temporal query functions in i2b2. Users can also combine terminology structures like ICD-10 with i2b2 items, which will automatically rearrange the items into the respective terminology structure space. All modifications performed in the Import and Mapping Tool are stored in the database, can be versioned and will be automatically loaded when resuming work. Uploading the enhanced version of the i2b2 concept hierarchy will not only fill the ontology table, but also all other necessary tables that are linked to the ontology, such as the observation_fact and the dimension tables.

3.4 Data Export Tool

An export plugin is provided for the i2b2 Web Client, allowing users to interactively define the dataset to be exported, including support for modifier-coded data. The IDRT i2b2 Web Client Plugin leverages metadata from the i2b2 concept hierarchy to collate related data items in tables, e.g. to bundle modifier-coded data elements belonging to the same clinical finding. Additionally, the plugin applies inclusion/exclusion criteria to the data before export, so the result set is not just restricted to the patient set, but also to the individual data elements actually matching the query. Data can be viewed and sorted on screen and exported in CSV format.

3.5 Best-practice Documentation

Documentation is always crucial for a software product, but even more in a setting where clinical hypotheses might depend on the consistency of data and the comprehensibility of the user interface. To lower the barrier for new users and administrators, a number of best practice guidelines have been developed. First, constructing an i2b2 ontology is no easy task with regards to the feature-rich data model: naming of folders, concepts and codes, usage of concept qualifiers (modifier codes), user-defined SQL operators, embedding XML code for realizing normal ranges, partitioning of i2b2 ontologies, user rights, or multi-axial hierarchies. Therefore, a white paper for building i2b2 ontologies is provided to the community.

A best practice document describing possible performance optimizations and their benefits against standard i2b2 setups is provided. The largest performance gain (80x) could be achieved by partitioning the observation_fact table (range-based on the concept_cd column), which however requires a database server that supports partitions. The second largest performance gain (77x) could be achieved by index modification of the observation_fact table (replacement of a compound index with two separate indexes for concept- and modifier-codes). Usage of the total_num field of the i2b2 ontology table, which has been described as a useful optimization [42] surprisingly did not provide a notable performance gain. However, filling the total_num field is still recommended as it is used by the i2b2 web client to display overall patient counts per folder and leaf in the concept hierarchy. A comprehensive presentation of all optimizations and the testing methods can be found online in the best practice document [43].

3.6 Application of IDRT Components in the Defined Use Cases

All parts of the IDRT were constructed to give quick and easy access to an otherwise laborious task of installing and filling a research data warehouse. The tools were tested with real life data in different use cases and with data types.

3.7 Clinical Data Warehouse

The Clinical Data Warehouse use case was implemented at Erlangen University Hospital, providing i2b2 as an additional query frontend for research applications. The i2b2 Wizard was used for setup and configuration, and the ICD, OPS and LOINC standard terminologies were integrated, resulting in a concept hierarchy with 118,490 nodes. The platform is currently running 14 i2b2 instances containing either full departmental datasets or project-related subsets. The size of datasets varied between 1,617 patients/47,453 facts for a small department and 226,457 patients/3,681,370 facts for a large project.

The §21 extractor was used at Erlangen, Göttingen and Leipzig University Hospitals to import benchmarking datasets into i2b2. Each site imported a locally generated §21 benchmarking dataset into the platform with data sets ranging from 42,000 to
47,000 patients with around 53,000 cases and up to 1.3 million individual observations.

3.8 Research Data Repository

The trial database of the Competence Network for Congenital Heart Defects [44] – available in CDISC ODM files – was integrated with i2b2 [28]. The source data inherent ODM Ontology hierarchy can be used for composing selections and filters on the data. Over 16000 data sets with roughly 80 concepts were imported and partially combined with imaging metadata.

For the Integrated Research and Treatment Center for Sepsis Control and Care Jena [45], a research database was created combining exported clinical data in CDISC ODM format from 12 clinical trials running in the CDMS OpenClinica with those of the central Sepsis Registry as well as data from the patient data management system, resulting in a research repository containing 31 medical concepts and 1.4 mio. data points.

Several Research Repositories were implemented at Erlangen University Hospital: Within the multicenter cloudhealth project funded by the German Federal Ministry of Economic Affairs and Energy, the i2b2 Wizard was used to setup multiple i2b2 instances. The ODM extractor was used to import 585,248 free text pathology reports for analysis in the Pathology use case and 67,344 free text surgery reports and coded diagnosis/procedure data in the Endoprosthetics registry use case. IDRT tools are being used in several ongoing local and national projects, including the "Clinical Data Intelligence" Big Data project (i2b2 Wizard, ODM and SQL extractors), MELC biobank (i2b2 Wizard, CSV and biomaterial extractors) and CONKO 007 biobank (i2b2 Wizard, ODM and biomaterial extractors).

3.9 Translational Research Unit

For the Clinical Research Group 241 [46] which focusses on the longitudinal course of psychosis, the trial data from the local team in Göttingen can be accessed as CDISC ODM from the secuTrial System (IAS GmbH). The ODM file contains data for over 700 patients and defines about over 1500 items, resulting in over 1.8 million patient facts (many items being "repeatable") within the file. Using the ODM extractor uploading this data set to i2b2 takes about 25 minutes and results in nearly 1.3 million patient facts in the i2b2 database (discarding some null values and not research related items created by the secuTrial System). Furthermore biomaterial metadata was added to this trial data via the CSV extractor, adding roughly 800,000 additional facts as nested data in i2b2. With the help of the Import and Mapping Tool mapping editor the i2b2 ontology was optimized by grouping often used search terms and renaming items for a more user-friendly browsing. One goal was to use the i2b2 platform as an easy tool for quality assurance and data management, being able to query over all available data and substituting missing features in the data capture tools.

4. Discussion

With the Integrated Data Repository Toolkit (IDRT) project we intended to develop a suite of programs to facilitate health analytics on heterogeneous medical data complementary to the established i2b2 stack. The goal was to give immediate and effortless access to a clinical data research platform for merging and querying data from different sources.

Therefore we implemented and introduced an integrated toolkit into practical use to facilitate the creation and maintenance of i2b2-based biomedical data marts in secondary use [47] and translational projects [47]. In order to lower the barrier to entry for this platform we provide a simplified setup tool (including SHRINE) as well as loading pipelines for various widely used data formats. Most of the central requirements from i2b2 literature review and stakeholder interviews like graphical user interfaces, pre-configured ETL pipelines and terminology support were implemented. The i2b2 Wizard with its setup and configuration functions thus significantly lowers the barrier to entry for sites that need to provide and i2b2-based query framework based on individual platform and configuration requirements.

The ODM extractor provides a large coverage of clinical trial data from multiple systems. The ODM extractor was successfully tested with secuTrial and OpenClinica. The tabular extractor covers data from any system using a relational database as well as manually exported data from clinical or research scenarios. As an important German use case the importer for the §21 (German Hospital Reimbursement Act) dataset allows a “quick win” introduction of the i2b2 platform with immediate availability of clinical data for querying. Within the constraints of data privacy regulations, more than 1.7 Mio data sets from German University hospitals could be made available in the §21 format for 2011 [48]. Even though the §21 dataset is specific to Germany, the approach of adding support for broadly accepted national core datasets should be considered for other platforms as well. Similar datasets should be available in other countries where reimbursement is based on Diagnosis-related Groups (e.g. USA, Australia) and other countries as well (UK National Health Service, France).

With ODM support covering data generated from research projects, §21 import capabilities allow the streamlined import of routine clinical care data into an i2b2 database. While the limited amount of data elements and their focus on billing could be considered a disadvantage, the IDRT platform provides all necessary tools to extend this core dataset with additional health record data.

The biomaterial data extractor addresses the lack of established standard format through definition of a generic concept hierarchy for biomaterial metadata. Respective drivers were successfully implemented for one commercial system with the possibility to adapt it to other systems.

A major concern after importing data is the integration of data from different sources. Thus the support of terminologies and ontologies is a key success factor. Using the i2b2 concept hierarchy we integrated standard terminologies (e.g. LOINC, MedDRA, TNM), which avoids recurring work at each site to convert into i2b2 format. However, several terminologies of interest could not be added, e.g. because of li-
censing issues in Germany and size (SNOMED CT). Additionally the integration with the German Meta Data Repository (MDR) [37] facilitates the re-use of well-documented concepts and value sets for querying.

The Import and Mapping Tool closes an important gap in the i2b2 infrastructure by allowing users to interactively modify concept hierarchies and create mappings between source & destination structures which are then applied during the ETL phase not only to structural but patient data as well. All components are available as Talend Open Studio packages for integration into individual ETL pathways, e.g. for automation, customizing or when pre- or post-processing are required. The editor also enables the use of the imported terminologies by mapping patient data directly onto them, either by use of customizable regular expressions or manually via a drag and drop user interface.

One of the biggest obstacles of transferring clinical data into the i2b2 platform was the loss of information about interlinked data like a medication and the dose of this medication giving to a patient in one particular instance. Even though the i2b2 modifier system was introduced in i2b2 version 1.6 to expand the i2b2 star schema with the ability to store such nested data, the documentation and community extension enabling modifiers were lacking. Implementing support for i2b2 modifiers in all the general data extractors and the editor broadens the usefulness of potential data sources significantly. Moreover the advanced query interface introduced by the i2b2 core team in conjunction with modifiers lets the researchers formulate complex queries. Quick access to results of these queries is giving via the IDRT Web Client plugin with its emphasized modifier support.

The i2b2 platform lacks native support for omics data, and its basic star schema is not suitable for most omics data types. Although projects like ONCO-i2b2 [49] have successfully expanded i2b2 to handle subgroups of omics data, these extensions are either not yet released to the public or not updated for the latest i2b2 releases. The i2b2 roadmap contains several extensions to address this issue in future releases, including interfaces to NoSQL-type databases for storage of large-scale omics data [50]. These developments may be accelerated by the “Big Data to Knowledge” (BD2K) grant [51, 52]. For the time being, one possible alternative examined by Candel et al. [53] is the tranSMART platform [24], an accessible and generic system based on i2b2 with broad omics data support. TranSMART expanded the i2b2 star schema with additional tables and created its own server and search tools. To test the flexibility of our ETL approach we successfully created experimental modified loading functionality for the Import and Mapping Tool and its ontology editor to work with the current tranSMART release (1.2).

The IDRT toolkit fulfils important requirements for integrated data repositories described in our research consortia and in literature [54]. The Desiderata given by Huser [54] formulate basic requirements for re-use & integration (fulfilled by IDRT through extractors) and maintenance (through i2b2 wizard & configurable ETL extractors). IDRT supports the desiderata for a single patient identifier (through PID generator inclusion), fact-nesting (through modifiers, also in Web Client/Export-Plugin), semantic integration & terminology model (through terminologies & MDR interface).

Shin [3] classifies i2b2 as a clinical research data warehouse, which contrary to a clinical data warehouse is used for research only. Shin further defines five desired characteristics of such a research data warehouse, with IDRT supporting the data extraction characteristic, but not covering honest broker or special functions for institutional review boards, which are gaps in the current i2b2 ecosystem that could be closed by modular extensions. In the current version, IDRT does not cover context representation, documentation & metadata and IDR historical evolution, however the i2b2 database schema does not yet provide a means for storing such information.

The creation of all IDRT tools was accompanied by field testing with varied medical data following the three introduced use cases. We iterated preliminary results with different research consortia according to the spiral approach described in Anderson 2012 [2] not only for stress testing and real life usage, but to recognize the advantage of minimizing the effort of implementing and updating the technical infrastructure i.e. the ETL in said spiral. Performance testing also included an assessment of the underlying database structures of i2b2, uncovering several optimizations that enabled query speed improvements of up to 80×, which have been made available to the community through a best practice document. While these optimizations have so far only been evaluated on an Oracle 11g platform, they should be transferable to other i2b2-supported database platforms.

Research data in patient oriented settings have to be assessed in the context of informed consent and privacy. Therefore mechanisms for pseudonymization, data separation, and k-anonymity are crucial for research data management toolsets. Even though we showed how such services can be easily incorporated into our toolset with the integration of the TMF PID Generator, other tools for privacy mechanisms should be integrated.

5. Conclusions

Research data warehouses and research databases provide a unique benefit by enabling scientists to query diverse data generally spread over different platforms and formats. The i2b2 platform adds more advantages with its easily expanded plugin structure, adapting for instance advanced analytics functions for specific research projects [49]. As an outcome of the IDRT project we provide tools and best practice for the most demanded functions according to our initial requirements analysis which including a designated i2b2 workshop. These include the installation/configuration and straightforward populating the platform from common data sources with additional options for refining data with structural and annotation information.

The approach presented here has been tested in a number of scenarios with millions of observations and tens of thousands of patients. Although no formal evaluation on the user experience has been performed yet, trained researchers were able to construct new analyses on their own. Early
feedback indicates that timely and extensive access to their “own” data is appreciated most. Furthermore, the IDRT is also lowering the barrier for the application of more specific tasks, for instance checking data quality and completeness (missing data, wrong coding, cf. [2]).

For future projects we have to focus on the roadmap of both i2b2 and tranSMART to the end of offering and maintaining a sustainable, update-proof infrastructure for translational research from the ETL pipeline, the ontology mapping tool up to the export interfaces for further investigation. The extension of the IDRT toolset over the course of the project to support multiple Linux distributions as well underlying databases shows that the platform is modular and extensible. Making the source code available on a public open source repository [55] will facilitate ongoing development through the i2b2 community. Documentation, compiled applications and a pre-installed virtual machines have been made available through the i2b2 community website [41] to provide a comprehensive starting point for interested users.

Acknowledgments

This research was funded by the TMF under BMBF grant 01GI1003. Additional support was provided by ZKS2 (BMBF/DLR grant 01KN1102), IFB Adipositas (BMBF/DLR grant 01EO1001) and sysINFLAME (BMBF grant 01ZZX1306C).

References

4. Cugia M, Besana P, Glasspool D. Comparing semi-automated systems for recruitment of pa-
6. Natter MD, Quan J, Ortiz DM, Bouvaros A, Ilo-
wite NT, Inman CJ et al. An i2b2-based, generaliz-
able, open source, self-scaling chronic disease reg-
12. Segagni D, Ferruzzi F, Larizza C, Tibullo V, Napo-
13. Weinlisch B, Mate S, Prokosch HU, Ganslandt T, Toddenroth D. “R-Scriptlets” für i2b2-Endan-
wender. GMDS 2014. 59. Jahrestagung der Deuts-
chen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e.V. (GMDS). Göt-
tingen, 07–10.09.2014. Düsseldorf: German Medi-
cal Science GMS Publishing House; 2014. Doc-
Astr. 346 2014.

20. Uzuner O, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and re-
22. Sun W, Rumbakshy A, Uzuner O. Evaluating tem-
poral relations in clinical text: 2012 i2b2 Chal-
24. Athey BD, Braxenhalter M, Haas M, Guo Y. tranSMART: An Open Source and Community-
27. Amin W, Tsui FR, Borromeo C, Chuang CH, Espi-
29. Deshmukh VG, Meystre SM, Mitchell JA. Evaluat-
32. Löbe M, Stäubert S, Winter A. Integrated Data Research Repository Toolkit (IDRT) – Deliverable D1.1: Aktualisierung und Formalisierung der beste-

IDRT-D1.1-Deliverables-IDRT-D1.1-Anforderung
6W1jaVaBE).
33. Stausberg J, Löbe M, Verplancke P, Dreppe J, Herre H, Löffler M. Foundations of a metadata re-
ev.de/Themen/Projekte/V015_01_PID_Generator.

Datenlieferung/Datensatzbeschreibung. 
42. Optimizing Query Performance with the Ontology Total_Num field – i2b2 Developer’s Forum – i2b2 Wiki [cited 2015 Aug 25 (archived at: http://www.webcitation.org/6b2e9h8b)]. Available from: https://community.i2b2.org/wiki/display/DevForum/Optimizing+Query+Performance+with+the+Ontology+Total_Num+field (V.04).

© Schattauer 2016