The Hannover Medical School Enterprise Clinical Research Data Warehouse: 5 Years of Experience

Svetlana Gerbel[1[0000-0003-3430-5294]], Hans Laser[1[0000-0002-0958-8600]], Norman Schönfeld[1[0000-0003-3954-1915]], Tobias Rassmann[2[0000-0003-2996-9096]]

[1] Hannover Medical School, Carl-Neuberg-Str. 1, 30625 Hannover, Germany
[2] Volkswagen Financial Services AG, Gifhorner Straße 57, 38112 Braunschweig, Germany
gerbel.svetlana@mh-hannover.de

Abstract. The reuse of routine healthcare data for research purposes is challenging not only because of the volume of the data but also because of the variety of clinical information systems. A data warehouse based approach enables researchers to use heterogeneous data sets by consolidating and aggregating data from various sources. This paper presents the Enterprise Clinical Research Data Warehouse (ECRDW) of the Hannover Medical School (MHH). ECRDW has been developed since 2011 using the Microsoft SQL Server Data Warehouse and Business Intelligence technology and operates since 2013 as an interdisciplinary platform for research relevant questions at the MHH. ECRDW incrementally integrates heterogeneous data sources and currently contains (as of 8/2018) data of more than 2,1 million distinct patients with more than 500 million single data points (diagnoses, lab results, vital signs, medical records, as well as metadata to linked data, e.g. biospecimen or images).

Keywords: Clinical Research Data Warehouse, Secondary Use of Clinical Data, Data Integration, BI, Data and Process Quality, Text Mining, KDD, System Architecture.

1 Introduction

1.1 Data Warehouse and Secondary Use of Clinical Data

The secondary use of information means using the information outside the original purpose of use, e.g. using routine health care data for quality assurance or scientific purposes. The reuse of electronic health record (EHR) data for research purposes has become an important issue in the national debate [1, 2].

A typical large university hospital is characterized by a heterogeneous IT system landscape with clinical, laboratory and radiology information systems and further specialized information systems [3]. The IT system landscape of an university hospital usually includes systems for documentation and processing of data that are generated during the provision of health services (e.g. diagnostics and clinical findings) or the administration of patient data (e.g. master data). The totality of these systems is referred to as the Hospital Information System (HIS).

The Hannover Medical School (Medizinische Hochschule Hannover, MHH) is no exception in this context. In addition to the clinical and laboratory information system (HIS and LIS), the range of IT solutions used at the MHH also includes a number of individual solutions in the field of clinical research based on a wide variety of widely used database management systems (from Oracle, Microsoft, FileMaker etc.) as well as an unmanageable number of table-based documentation systems.

A universal approach for central data integration and standardization within an organization with heterogeneous databases is to build a data warehouse system based on database component consisting of consolidated and aggregated data from different sources. Data warehouse technology allows users to run queries, compile reports, generate analysis, retrieve data in a consistent format and reduce the load on the operative systems.

Already Teasdale et al. [4] and later Bonney [5] made clear that the reuse of clinical (primary) data is facilitated by Business Intelligence on the basis of a so-called "Research Patient Data Repository" or Clinical Data Warehouse. The use of Business Intelligence creates the basis for further extraction of empirical relationships and knowledge discovery in databases (KDD), like in the field of data science.

The relief of operative data processing and application systems is another central argument for the use of data warehouse technology. This makes it possible to execute requests for clinical data on a dedicated repository rather than at the expense of the operative systems [6].

In the clinical-university sector in Germany, there is a series of established data warehouse solutions for secondary data use. They are commonly described as Clinical Data Warehouses (CDW) [2]. In the IT environment of a hospital, the response times of operative systems (e.g. clinical workplace applications) are a particularly critical factor. Accessing the operational systems with real-time queries would increase the likelihood of non-availability. In addition to these drivers, the following typical application scenarios for secondary use [2, 3, 7, 8] have formed:

- Patient screening for clinical trials based on inclusion and exclusion criteria
- Decision support through comparison of diagnosis, therapy and prognosis of similar patients
- Epidemiological evaluations by examining the development of frequencies of clinical parameters (e.g. risk factors, diagnoses, demographic data)
- Validation of data in registers and research databases and their data enrichment with the aim of quality improvement

In the review of Strasser [3] from 2010, the IT systems of 32 German university hospitals were evaluated with regard to the components of the HIS and the available data warehouse solutions for consolidating routine clinical data for secondary use of data. The results of this survey correspond to a survey conducted by the CIO-UK (Chief Information Officers - University hospitals) in 2011 in order to identify existing IT infrastructures and technology stacks in Germany that solve the challenges of data integration and data management for secondary data use with regard to Clinical Data Warehouse technology. The CIO-UK represents interests from the 35 university hospitals in Germany. In summary, it can be said that there is no universal solution to

implement a data warehouse technology for secondary use in Germany. There are numerous different implementations of CDWs in the community that use open source based frameworks such as i2b2[1] (incl. tranSMART) or proprietary development solutions with popular DBMSs (for example, Microsoft SQL Server, Oracle and PostgreSQL) [9, 10]. Generally speaking, currently i2b2 and OMOP[2]-based approaches appear to be the most widely used worldwide [11].

1.2 Background

MHH is one of the most efficient medical higher education institutions in Germany. As an university hospital for supramaximal care with 1,520 beds, the MHH treats patients who are severely ill. They benefit from the fact that the medical progress developed at the university is quickly available to the patients. Every year more than 60,000 people are treated in more than 70 clinics, institutes and research facilities; in the outpatient area there are around 450,000 treatment contacts per year [12].

Centralisation of the operational systems at the MHH is ensured by the Centre for Information Management (ZIMt). The ZIMt is responsible for the provision of the operational systems and ensures maintenance, support as well as the adaptation of the IT systems to the needs of the MHH. The ZIMt operates a class TIER 3 computer centre [13] at the MHH, i.e. a primary system availability of 99.982%. In addition, the MHH departments are certified according to DIN EN ISO 9001:2015. Thus, the highest demands are placed on the processes (SOPs) and offer patients as well as employees the assurance that they can rely on compliance with defined quality standards.

The Enterprise Clinical Research Data Warehouse (ECRDW) is an interdisciplinary data integration and analysis platform for research-relevant issues that has been available enterprise-wide since July 2013 [7]. The provision and support of the ECRDW as a central service at the MHH is carried out by the Division for Educational and Scientific IT systems of the ZIMt.

2 Material and Methods

2.1 Data Sources and Interfaces

The HIS of the MHH is operated by the ZIMt and consists of more than 50 subcomponents (e.g. Electronic Medical Record System (EMS), Laboratory Information Systems (LIS) and Radiology Information Systems (RIS)), which exchange EHR data via a communication server. The ECRDW integrates the EHR data of the HIS via existing HL7 interfaces, which are provided via a communication server, as well as via separate communication paths for systems that are not or no longer connected to the communication server (legacy systems).

---

1 https://www.i2b2.org/, https://www.i2b2.org/webclient/

2 https://www.ohdsi.org/data-standardization/

2.2 Selection Process and Evaluation of an Appropriate Data Warehouse Development Platform

The selection of a data warehouse technology was carried out between 2010 and 2011 by a working group of ZIMt and 12 other MHH departments (a total of 24 participants from the fields of IT, clinical, biometrics and clinical trial). The multi-stage selection process was divided into: product presentations, software implementations, workshops with suppliers, (weighted) evaluation of the tools by the working group members using the developed catalogue of requirements. The decisive selection criteria (inclusion and exclusion criteria) were as follows:

- Complete solution (data integration and analysis tools)
- Independence from the product vendor (autonomous development possible)
- Powerful data integration tool (ETL)
- Active community (knowledge bases, know how, support)
- Suitable license model

A total of five software vendors were evaluated. At the end of a 10-month selection process, the working group opted for Microsoft SQL Server BI data warehouse technology based on the selection criteria. The trend described in the Magic Quadrant of Business Intelligence Platforms 2011 by the Gartner Group market research report complemented our evaluation [14].

2.3 Development Architecture and Data Warehouse Approach

The development in the ECRDW is based on a three-tier-deployment-architecture: development, test and production environment (Fig. 1). The development environment consists of a database server and a dedicated server for version control in order to develop ETL processes and store the development artifacts in a version-secured manner. Systems of the development and test environment are virtualized to more dynamically distribute and economize resources. The test environment reflects the structure of the production environment in a virtual environment and is thus divided into a database server, analysis server and reporting server. Development statuses are first delivered and tested in the test environment. After successful test runs, the development artifacts (releases) are rolled out to the production environment (rollout) and tested again. The production environment therefore consists of a database server, analysis server, and reporting server. For performance reasons, the database server in the production environment is not virtualized.

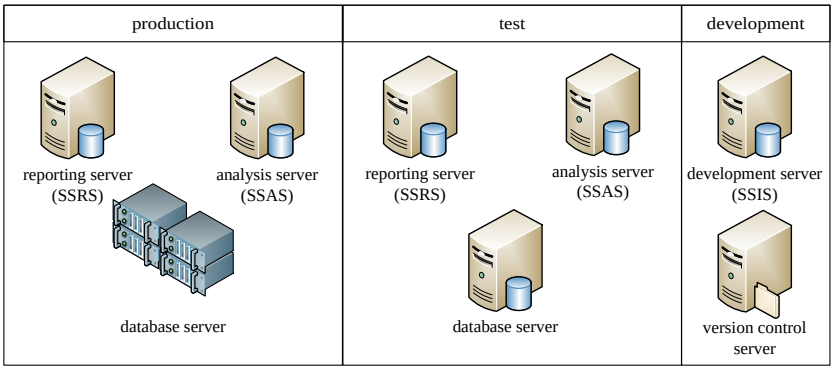

Fig. 1. The ECRDW three-tier-deployment-architecture

The ECRDW is based on the Microsoft SQL Server architecture. The SQL Server serves as database management system and core data warehouse repository. The additional service SQL Server Integration Services (SSIS) is used for data integration via the ETL (Extract, Transform, Load) process and for developing and providing Business Intelligence (BI) solutions SQL Server Analysis Services (SSAS). The data modelling is done in concepts in a relational data model and is based on the Inmon architecture [15]. This means that data is first merged into a consolidated layer, which forms the basis for departmental views of the data sets (known as data marts). This approach makes it possible to create a comprehensive respective global scheme for the data captured at the MHH.

Since the primary data is collected for clinical purposes and for billing, a preprocessing is necessary, e.g. to reduce the heterogeneity of the system-specific data models and to check data integrity and consistency. In addition, clinical data in a HIS is exchanged typically among the primary systems via HL7. Logically, this leads to redundant storage of information. Thus, master data transmitted via the HL7 message type ADT (Admission, Discharge, Transmission) to each primary system is duplicated. For each single fact from the clinical documentation there is a primary leading operative system at the MHH. To avoid possible ambiguities or redundancies (requirement for entity matching) in data integration, the leading operative system is identified in each integration project. Information from other primary systems of the HIS that consume the primary data of another system is not integrated. If two primary systems represent a similar concept (e.g. laboratory findings), the semantic integration of both systems into the same concept of the ECRDW takes place. If duplicates of the information are produced, the primary operative system for this information is identified again. Deduplication thus takes place within the ETL process. Primary data from the various data sets of the operative systems are incrementally integrated into a central repository using data warehouse technology. Depending on the possibilities of the HIS subsystem, this process takes place daily or weekly in an incremental loading process via ETL into the ECRDW core repository. After consolidation and standardization in modelled standard concepts (e.g. historized master data), central use cases can be served by providing targeted data selections.

2.4 Methods to Ensure Data Protection

In Germany, the evaluation of primary data arising in the context of treatment is regulated by aspects of data protection at EU, federal and state level. In addition there are the legal regulations (SGB V, SGB X, infection protection, cancer early detection and cancer register, hospital laws etc.). This special feature is reflected in the possibilities of using these data for research purposes (secondary data use) and inevitably leads to a limitation of the use cases for such data. In secondary data analysis, the right to informational self-determination of the individual must always be protected and weighed against the right to freedom of science and research [16].

With the entry into force of the EU Basic Data Protection Regulation (GDPR) (EU 2016/679) on 25 May 2018, the processing of genetic, biometric and health data is prohibited under Article 9(1) of the GDPR, unless the person has explicitly consented to the use of the data. The current Lower Saxony Data Protection Act (§ 13 NDSG) states that a person must have given consent to the use of personal data for scientific purposes. A data protection and access concept must therefore be defined in order to comply with data protection regulations. The implementation is described in 3.4.

2.5 ECRDW Use Cases

Typical application scenarios of a clinical data warehouse [2, 7, 8] were implemented at the MHH in three central application cases.

Screening: By means of a so-called anonymous cohort identification, researchers have the possibility to define a cohort via a data warehouse on the basis of inclusion and exclusion criteria. The criteria are used to calculate quantities for e.g. patients, cases and laboratory values on the basis of the EHR data and to be able to provide a statement on the feasibility of the research question. A screening for clinical studies is possible analogously and can contribute to the reduction of the sometimes time-consuming research on EHR data [17-20].

Epidemiological Study: Similar to screening in a clinical study, data collection can also be very time-consuming when performing a retrospective data analysis (epidemiological study). Medical findings are sometimes only available as PDF documents in the central archive for patient files after completion of treatment. Through the use of a data warehouse, a wealth of information about the entirety of the patients of a hospital, the clinical pictures and the context-specific final results of the therapy (e.g. condition at discharge) can be provided, which are available in the various application systems of a HIS.

Validation and Data Enrichment: Research and registry databases often suffer from manual data entry (so-called "media discontinuity"). As a result, errors, typos, incomplete or erroneous data collection are a challenge that many such data collections have to overcome [1]. The use of data warehouse technology can make a decisive contribution to correcting errors in existing information. In addition, data from a register can be completed by adding additional data from the database of a data warehouse (e.g. risk factors from EHR data) [21].

2.6 Information Quality Scheme: Process Chain, Artefact, Relativity (PAR)

In order to improve the information quality in the analysis of large amounts of data, the criteria for mapping the information quality of Wang and Strong [22] were examined and modified for transferability to a clinical research data warehouse.

The result was a modified three-dimensional information quality scheme (Process chain, Artefact, Relativity, PAR) with a total of 26 criteria. During the development of the ECRDW, a subset of information quality criteria of the PAR scheme which are assigned to the sub-process of processing in the process chain dimension and by the definition of templates has been implemented. The aspect of reuse was the key point of the process chain dimension. These are 12 information quality criteria: standardization, source traceability, loading process traceability, processing status, reference integrity, uniform presentation, data cleansing scope, degree of historization, no-redundancy, performance and restartability [23].

3 Results

3.1 Content of the ECRDW

The MHH ECRDW has been continuously integrating data from MHH's primary systems into a relational, error-corrected and plausibility-tested data model since it went live in 2013 (see Table 1).

Table 1. Content of the ECRDW (as of July 2018)

Domains (millions)	07/2013	07/2014	07/2015	07/2016	07/2017	07/2018
Biospecimen	-	-	-	-	0,01	0,04
Demographic data	1,97	2,12	2,28	2,47	2,64	2,93
ICD diagnosis	-	6,65	7,75	8,92	9,88	11,92
Intensive care	-	-	-	-	-	228,23
Laboratory findings	-	167,50	186,96	208,42	236,24	287,56
Movement data	-	-	-	-	14,89	16,98
Radiological findings	-	-	-	-	-	0,97
Risk factors	-	-	-	-	0,03	0,05

As of July 2018, the ECRDW repository contains data from more than 2 million patients, more than 11 million diagnoses and more than 6 million cases with approximately 500 million data points.

The ECRDW currently collects administrative information such as demographic data, movement data, visit data, diagnoses (ICD-10GM), risk factors and severity of the disease from the SAP i.s.h. system. The EMS (SAP i.s.h.med) is used to load reports, findings and discharge letters. Metadata for biosamples is provided from the MySamples and CentraXX systems. Intensive care data originates from a legacy system (COPRA) as well as from the operative ICU system (m.life). The ECRDW receives information on findings from the laboratory via LIS (OPUS::L). Metadata on radiological examinations (including findings) are provided

via a RIS (GE Centricity). Cardiovascular data, findings and values of cardiological echocardiographies as well as cardiac catheter examinations originate from another RIS (IntelliSpace CardioVascular). Depending on the source system, different times are therefore possible for the start of digital recording. The earliest capture times are shown in Table 2 analog to Table 1.

Table 2. Earliest time of recording for each domain

Domains	Minimal date (month/year)
Biospecimen	10/2012
Demographic data	09/1986
ICD Diagnosis	01/2007
Intensive care	05/2005
Laboratory findings	06/2000
Movement data	04/2008
Radiological findings	12/2013
Risk factors	07/2007

3.2 Projects Implemented with ECRDW

The ECRDW of the MHH is productive since 07/2013 and provides data on research-relevant issues. In the period from 07/2013 to 07/2018, 48 project inquiries from 39 departments of the MHH were registered. Three project requests are annually recurring data deliveries (e.g. register implementation). Table 3 shows an overview of the registered projects in the period from 2013 to 2018 for the three central use cases (screening, epidemiological study, validation and data enrichment).

In some projects, text analysis methods were used to obtain additional features from full-text documents, such as radiological findings and discharge letters, and to make them available with data from the ECRDW's structured databases.

Researchers are able to use innovative methods, such as medical data mining, to identify new hypotheses about the amount of data due to the very large amount of data per project.

From a total of 33 MHH clinics, 16 clinics (48%) submitted an evaluation request to the ECRDW in 2018 (by July). In some projects, in addition to providing data, a screening was carried out in advance to check the feasibility of the project request. In relation to the provision of data for epidemiological study requests (31 projects) and data enrichment (16 projects), however, only in 9 projects a screening for patient data has been carried out.

Table 3. Number and nature of ECRDW-based research projects (2014-2018)

Year	# Projects	# Departments	Screening	Epidemiological study	Validation and data enrichment
2014	4	4		3	1
2015	4	4		1	3
2016	13	9	2	8	5
2017	8	6	2	6	2
2018	19	16	5	13	5
Total	48	39	9	31	16

3.3 Data and Process Quality

In order to ensure data and process quality, the process chain dimension "processing" from the developed PAR scheme was completely implemented in the ECRDW. The following criteria, among others, were taken into account:

Traceability of the Data Origin (Data Linage / Data Provenance): Additional columns (load date and update date) in the tables as reference to the source system of every record were added.

Traceability of the Loading Process: Errors that occur during load jobs can be assigned to an unique error table referencing additional tables providing information about job definition and run. Incorrect records remain in a staging table for each entity and are deleted only after successful loading in the core data warehouse.

Restart Capability: Integration jobs can be repeated at any time, since the data is only deleted from the staging area when the record is successful load in the data warehouse and the update of the data warehouse only takes place when the ETL process is complete and the temporary target tables are merged into the real tables.

Standardization of the Development of ETL Pipelines and Modeling: For the data integration we use three templates (Staging, Historization and Update) for the ETL processes, which are already predefined and only need to be adapted.

Referential Integrity: The artificially generated primary and foreign keys are based on adequate hash function.

Redundancy-Free: Duplicates are recognized between loading processes and within a loading process using a hash value. The script for managing duplicates is integrated into the standardized templates.

Performance: To optimize performance, lookup tables are created before the start of the loading processes (using the hash function for reference checks) and then truncated again.

Data Cleansing Scope: An own developed model for error codes is used which classifies errors at attribute level (or finer) and monitors them in an error reporting system for each subject area.

Time Variance: The changes of data over time (historization) are tracked via the concept of a temporal database.

Standardized reports generated on the basis of the error records are held in the staging area. These are automatically distributed to the ECRDW team from the source system (BI) so that the cause of the errors can be identified and eliminated. For the long-term analysis, the errors are classified as persistent and BI procedures are applied to these error tables to identify error constellations and proactively avoid them in the sense of datamining.

3.4 Data Protection and Security

In coordination with the data protection officer of the MHH, a data protection concept was developed that defines processes for data use and access. The data protection concept provides the storage of health data in pseudonymised form. The pseudonymisation of health data is a necessary step towards compliance with data protection in order to protect patients from the identification of their person. Instead of patient identifying data (IDAT/PID), pseudonyms (surrogate keys) are used. The pseudonyms are administered and assigned in the ECRDW.

The patient's consent to the use of his/her data for scientific purposes is registered with the MHH treatment contract when the patient is admitted. The current data protection concept stipulates that the patient's consent must be given for any data processing for scientific purposes. This is taken into account in every step of a data provision process.

To ensure data security, all ECRDW systems are backed up daily via a central backup concept. Security authentication and authentication of ECRDW users takes place via the central MHH Active Directory. A transaction log archives all queries and executing users.

4 Discussion and Conclusion

The use of a data warehouse technology as the basis for the implementation of multidisciplinary data integration and analysis platform results in some significant advantages for clinical research, among others:

- Relief for the operative health care systems
- Support in the planning and implementation of studies
- Data enrichment of research databases with quality-assured information from central systems (e.g. laboratory systems, administrative systems, OR systems)
- Integration and storage of historic data repositories which are not usable for IT (legacy systems)

In addition, the research repository with consolidated data from different domains (HIS and research systems), offers a more complete data set and thus a basis for investigating relationships and potential patterns between disease progression and measures. The using of medical data mining methods based on the extensive and retrospective data of an ECRDW can serve as a valuable resource for the generation of innovative knowledge in all areas of medicine [24].

The development of a translational data integration and analysis platform is however a long-term process. Although the MHH ECRDW project was initiated in 2010 (by a core development team of two persons), it was made available for all health researchers in June 2013. The selection of the appropriate data warehouse development platform, data modelling, a development of an integration strategy, data protection, security and use and access concepts and finally testing, validation and maintenance phases are the time consuming but needed phases of an iterative development process. Another very important issue is the involvement of experts from different fields (computer scientists, physicians, statisticians, privacy protect officer, management etc.) in different phases of development.

Semantic modelling of clinical concepts as well as analysis of unstructured data and OMICs are further key issue. At the MHH, additional tools for text analysis based on Natural Language Processing are currently being developed for the scientific use of findings that are only available in (semi-)structured form (such as medical letters and findings).

The FAIR Data Principles published in 2016 [25] define fundamentals that research data and research data infrastructures must meet in order to ensure sustainability and reusability. As part of a research infrastructure, the ECRDW has been respecting these principles in some aspects since 2013:

- *Findable*: data and metadata machine-readable and searchable through central database management system
- *Accessible*: use and access concept; data provision in standard formats or standard interfaces (CSV, ODBC, HTTPS)
- *Interoperable*: structuring of data by standard vocabularies, classification systems (ICD, LOINC, OPS, etc.); metadata on semantics between data sets
- *Reusable*: metadata can be exported machine-readable

Due to the actuality of the FAIR Data Principles, the future development of the ECRDW should also take into account these principles. The developed PAR scheme complements this approach.

Even in the age of knowledge graphs, data integration remains a major challenge [26]. With the increasing number of source systems, complexity has always increased. Challenges in data preparation and harmonization still apply to new big data technologies as well. We expect that the use of a data warehouse-based solution as an already consolidated and plausibility checked source will probably lead to a simplification by merging with further data sources (e.g. when consolidating with web sources) [27].

Intense exchange and collaborations within the national projects and facilities is required to take advantage of synergy effects. The involvement of ECRDW is an important factor of the sustainability concept within the Data Integration Centres of the Medical Informatics Initiative [28] and further national and international projects (e.g. German Biobank Alliance [29] and EHR4CR [30]).

References

1. Meystre, S.M. et al.: Clinical Data Reuse or Secondary Use: Current Status and Potential Future Progress, 10.15265/IY-2017-007, 2017. 10.15265/IY-2017-007.
2. Tolxdorff, T., Puppe, F.: Klinisches Data Warehouse, Informatik Spektrum, no. 39, pp. 233-237, June 2016. 10.1007/s00287-016-0968-3.
3. Strasser, N.: RDA - Systeme an Universitätskliniken in Deutschland, 2010.
4. Teasdale, S. et al.: Secondary uses of clinical data in primary care, Informatics in Primary Care, vol. 15, no. 3, pp. 157-66, 2007. 10.14236/jhi.v15i3.654.
5. Bonney, W.: Applicability of Business Intelligence in Electronic Health Record, Procedia - Social and Behavioral Sciences, no. 73, pp. 257-262, 2013. 10.1016/j.sbspro.2013.02.050.
6. Mucksch, H., Behme, W.: Das Data Warehouse-Konzept, 4th ed., Eds. Wiesbaden: Gabler Verlag, 2000.
7. Gerbel, S., Laser, H., and Haarbrandt, B.: Das Klinische Data Warehouse der Medizinischen Hochschule Hannover, Forum der Medizin_Dokumentation und Medizin_Informatik, no. 2, pp. 49-52, 2014.
8. Dugas, M., Lange, M., Müller-Tidow, C., Kirchhof, P., Prokosch, H.U.: Routine data from hospital information systems can support patient recruitment for clinical studies, Clinical trials, pp. 183-189, Apr. 2010. 10.1177/1740774510363013.
9. Dugas, M., Lange, M., Müller-Tidow, C., Kirchhof, P., and Prokosch, HU.: Routine data from hospital information systems can support patient recruitment for clinical studies, Clinical trials, pp. 183-189, Apr. 2010.
10. Murphy, S.N., Weber, G., Mendis, M., Gainer, V., Chueh, H. C., Churchill, S., & Kohane, I.: Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2), Journal of the American Medical Informatics Association, 2010.
11. Jannot, A-S. et al.: The Georges Pompidou University Hospital Clinical Data Warehouse: A 8-years follow-up experience, International Journal of Medical Informatics, no. 102, pp. 21–28, Feb. 2017.
12. Medizinische Hochschule Hannover. (2016) Jahresbericht der MHH. https://www.mh-hannover.de/fileadmin/mhh/bilder/ueberblick_service/publikationen/Jahresbericht_2016_GESAMT_FINAL_31.08.2017.pdf, last accessed 2018/09/15.
13. OVH SAS. (2018) Understanding Tier 3 and Tier 4. https://www.ovh.com/world/dedicated-servers/understanding-t3-t4.xml, last accessed 2018/09/15.
14. Gartner Inc. (2018) Magic Quadrant for Analytics and Business Intelligence Platforms. https://www.gartner.com/doc/reprints?id=1-4RVOBDE&ct=180226&st=sb, last accessed 2018/09/15.
15. Inmon, W.H.: Building the Data Warehouse, 3rd ed., Robert Ipsen, Ed.: John Wiley & Sons, Inc., 2002.
16. Arbeitsgruppe Erhebung und Nutzung von Sekundärdaten (AGENS), et al. (2012) Deutsche Gesellschaft für Epidemiologie e. V. http://dgepi.de/fileadmin/pdf/leitlinien/GPS_fassung3.pdf, last accessed 2018/09/15.
17. Thadani, S.R., et al.: Electronic Screening Improves Efficiency in Clinical Trial Recruitment," Journal of the American Medical Informatics Association, vol. 16, no. 6, pp. 869-873, 2009. 10.1197/jamia.M3119.
18. Sandhu, E., et al.: Secondary Uses of Electronic Health Record Data: Benefits and Barriers, Joint Commission Journal on Quality and Patient Safety, no. 38, pp. 34 - 40, 2012. 10.1016/S1553-7250(12)38005-7.

19. Beresniak, A., et al.: Cost-benefit assessment of using electronic health records data for clinical research versus current practices: Contribution of the Electronic Health Records for Clinical Research (EHR4CR) European Project, Contemporary Clinical Trials, no. 46, pp. 85-91, 2016. 10.1016/j.cct.2015.11.011.
20. Coorevits, P., et al.: Electronic health records: new opportunities for clinical research, Journal of Internal Medicine, no. 274, pp. 547–560, 2013. 10.1111/joim.12119.
21. Just, B.H., et al.: Why Patient Matching Is a Challenge: Research on Master Patient Index (MPI) Data Discrepancies in Key Identifying Fields, Perspectives in Health Information Management, 2016. PMC4832129.
22. Wang, R., Strong, D.: Beyond Accuracy: What Data Quality Means to Data Consumers, Journal of Management Information Systems, vol. 12, pp. 5-33, 1997. 10.1080/07421222.1996.11518099.
23. Rassmann, T.: Entwicklung eines Verfahrens zur integrierten Abbildung und Analyse der Qualität von Forschungsdaten in einem klinischen Datawarehouse, Dissertation 2018.
24. Prather, J.C., et al.: Medical data mining: knowledge discovery in a clinical data warehouse, Proceedings of the AMIA Annual Symposium, pp. 101-105, 1997. PMC2233405.
25. Wilkinson M.D., et al.: The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data 3, 10.1038/sdata.2016.18, 2016.
26. Konsortium HiGHmed: Heidelberg - Göttingen - Hannover Medical Informatics, http://www.medizininformatik-initiative.de/de/konsortien/highmed, last accessed 2018/07/31.
27. German Biobank Alliance (GBA), https://www.bbmri.de/ueber-gbn/german-biobank-alliance/, last accessed 2018/07/31.
28. Electronic Health Records for Clinical Research - (EHR4CR), http://www.ehr4cr.eu/, last accessed 2018/07/31.