Supporting Reuse of EHR Data in Healthcare Organizations: The CARED Research Infrastructure Framework

Vincent Menger[1,2], Marco Spruit[1], Jonathan de Bruin[3], Thomas Kelder[4] and Floor Scheepers[2]
[1]Department of Information and Computing Sciences, Utrecht University, Princetonplein 5, Utrecht, The Netherlands
[2]Department of Psychiatry, University Medical Center Utrecht, Utrecht, The Netherlands
[3]Department of Information and Technology Services, Utrecht University, Utrecht, The Netherlands
[4]EdgeLeap B.V., Utrecht, The Netherlands

Keywords: EHR, Data Management, Infrastructure, Open Source, Repeatability, Data Preparation, Data Analysis.

Abstract: Healthcare organizations have in recent years started assembling their Electronic Health Record (EHR) data in data repositories to unlock their value using data analysis techniques. There are however a number of technical, organizational and ethical challenges that should be considered when reusing EHR data, which infrastructure technology consisting of appropriate software and hardware components can address. In a case study in the University Medical Center Utrecht (UMCU) in the Netherlands, we identified nine requirements of a modern technical infrastructure for reusing EHR data: (1) integrate data sources, (2) preprocess data, (3) store data, (4) support collaboration and documentation, (5) support various software and tooling packages, (6) enhance repeatability, (7) enhance privacy and security, (8) automate data process and (9) support analysis applications. We propose the CApable Reuse of EHR Data (CARED) framework for infrastructure that addresses these requirements, which consists of five consecutive data processing layers, and a control layer that governs the data processing. We then evaluate the framework with respect to the requirements, and finally describe its successful implementation in the Psychiatry Department of the UMCU along with three analysis cases. Our CARED research infrastructure framework can support healthcare organizations that aim to successfully reuse their EHR data.

1 INTRODUCTION

The digitization of our society is rapidly creating opportunities to use new resources for research in healthcare in the form of routinely collected large datasets (Murdoch and Detsky, 2013; Priyanka and Kulennavar, 2014). Meanwhile, recent advancements in Machine Learning and Big Data analytics enable unlocking the potential value of these datasets (Groves *et al.*, 2013). While current research in healthcare is predominantly based on Randomized Controlled Trials (RCTs) and Cohort Studies (CSs), a data analytics approach on the other hand integrates real time data from various sources within a health care organization, such as structured patient records, unstructured text notes, lab measurements, financial data, and various others (Badawi *et al.*, 2014; Friedman *et al.*, 2014). This approach can have substantial benefits in addition to RCT and CS study designs, in terms of cost-effectiveness, sample size and reduction of selection bias (Raghupathi and Raghupathi, 2014; Gandomi and Haider, 2015). In the near future, this may allow a transition to a data driven healthcare, where using real time clinical data for supporting important decisions in the care process becomes the norm, and research using data analytics becomes an important driver of new insights into aetiology and treatment of disease (Murdoch and Detsky, 2013).

In this context, the various types of EHR data that have been gathered for the sake of delivering care to patients have become an important asset to healthcare organizations, that is nowadays typically available in a digital format (Dean *et al.*, 2009). Healthcare organizations have therefore started to assemble their EHR data in data repositories in order to apply data analytics techniques to them (Lokhandwala and Rush, 2016; Obermeyer and Lee, 2017). Several challenges in management and analysis of data have subsequently emerged, not only of a technical nature (e.g. secure storage of data, data pre-processing and data analysis), but also

Menger, V., Spruit, M., de Bruin, J., Kelder, T. and Scheepers, F.
Supporting Reuse of EHR Data in Healthcare Organizations: The CARED Research Infrastructure Framework.
DOI: 10.5220/0007343900410050
In *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2019)*, pages 41-50
ISBN: 978-989-758-353-7

of an organizational nature (e.g. combining data from different sources, effective collaboration between researchers, and reproducibility of research) and an ethical nature (e.g. legal regulations and patient privacy concerns) (Hersh *et al.*, 2014; Safran, 2014; Meystre *et al.*, 2017). To mitigate these challenges, a dedicated infrastructure consisting of appropriate hardware and software components is essential for gaining reliable and secure access to the data, and for sharing knowledge about the structure and meaning of data (Jensen, Jensen and Brunak, 2012; Danciu *et al.*, 2014). Current data repositories however are often based on Data Warehouse (DWH) technology which falls short in addressing most of these challenges, leading to scattering of both data and knowledge about data within an organization (Roski, Bo-Linn and Andrews, 2014; George, Kumar and Kumar, 2015).

A dedicated research infrastructure for reusing EHR data improves on this situation by providing a unified data management practice for all researchers involved, ranging from data sources on the one hand to applications in clinical practice and clinical research on the other (Hersh *et al.*, 2013). For example, it helps reduce errors in analysis, improves possibilities for collaboration among researchers of various disciplines, and leads to more efficient use of time and resources in the long term (Pollard *et al.*, 2016). Although the benefits of such an infrastructure are apparent, a general framework for an infrastructure to support reusing EHR data has not yet been proposed. Our study aims to provide such a framework. Because research infrastructure software packages need to interoperate with a large variety of health IT systems, databases, EHR software from different vendors, and other local standards (Hammami, Bellaaj and Kacem, 2014; Kasthurirathne *et al.*, 2015), our study will address this problem on the conceptual level rather than offering a software solution. Individual healthcare organizations can subsequently use existing tools within their organization, supplemented with (open source) software packages to implement an infrastructure consisting of appropriate hardware and software components based on our proposed framework, that is interoperable with their current systems and practices.

In this study, we will thus identify the most important requirements for an infrastructure for reusing EHR data by means of expert interviews in the University Medical Center Utrecht (UMCU) in the Netherlands. We then translate these requirements into relevant concepts and their relations and arrange them in a generic framework. The main merit of the framework we propose lies in providing clear concepts that need to be instantiated in a research infrastructure for reusing EHR data, thereby supporting learning healthcare organizations that aim to do so. We furthermore describe the implementation of our proposed infrastructure framework in the Psychiatry Department of the UMCU, and present three specific applications that were enabled by this infrastructure.

1.1 Related Work

In contrast to research into reusing EHR data within an organization, there are various examples of projects that aim to integrate all types of clinical data from different institutions. For example, the CER Hub (Hazlehurst *et al.*, 2015) which provides standardized access to the patient centric EHR across organizations, the SHARPn project (Rea *et al.*, 2012) which enables using the EHR for secondary purposes in multiple academic centers, and EHR4CR (De Moor *et al.*, 2015) which offers a scalable and efficient approach to interoperability between EHR systems. Additionally the eMERGE, PCORnet and SHRINE projects provide more research into the ethical, legal and social issues of combining data from multiple sites (Weber *et al.*, 2009; McCarty *et al.*, 2011; Fleurence *et al.*, 2014). All of these projects address topics such as semantic interoperability, data quality and data integration through a Trusted Third Party (TTP), which are only indirectly relevant when reusing EHR data within an organization.

Data management practices within an organization are usually designed for dealing with data from CS and RCT studies (Krishnankutty *et al.*, 2012), accompanied with infrastructure in the form of a Clinical Data Management System (CMDS) (Lu and Su, 2010). The data that is produced by CS and RCT studies contains measurements that are clearly defined in a study protocol, and that are often static after patient enrolment has ended. Challenges include data-entry and medical coding of data. This type of clinical data strongly differs from secondary EHR data, which is already present data that is updated live, and is often undocumented.

Research into infrastructure for reusing EHR, which is scarce in the first place, typically describes one or two requirements, and thereby only a small part of the solution that is needed. For example, they focus on the preprocessing and analysis pipeline (Peek, Holmes and Sun, 2014), analysing and storing large datasets (Youssef, 2014), integration of data sources (Bauer *et al.*, 2016) or composing

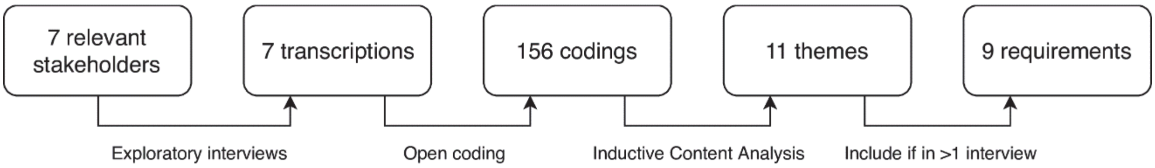

Figure 1: The process of identifying the nine requirements for the infrastructure.

research datasets from secondary data (Murphy *et al.*, 2010). Not one approach however provides a unifying data management practice, failing to provide the broad scope for infrastructure that we envision. The generic framework that our study will provide therefore has additional value for the field of clinical research data management.

2 METHODS

2.1 Identification of Requirements

Requirements for an infrastructure for reusing EHR data were identified in the University Medical Center Utrecht (UMCU) in the Netherlands. The process of identifying these requirements from expert interviews is depicted in Figure 1. First, semi-structured interviews were conducted with seven relevant stakeholders in the UMCU to explore the requirements for an infrastructure. A board level stakeholder and psychiatrist, a nurse researcher, and data and IT experts with several backgrounds were included in the interviews, ensuring representation of all relevant stakeholders. In this case, semi-structured interviews are the most appropriate method for eliciting new information (Gill *et al.*, 2008). Questions about the participants' views on current data management practices, possible improvements to these practices and their feasibility in the context of the UMCU, current issues experienced, and their ideal data management situation were asked.

The transcripts of these interviews were then processed using a grounded theory approach. Researchers first applied an open coding process to the transcripts by segmenting them and describing each segment in a word or short sequence of words (Strauss and Corbin, 1990). This resulted in 156 codings in the seven interviews combined. To be

Table 1: The number of codings in each interview, distributed over the nine themes and seven interviewees. The rightmost column shows the requirement that was formulated based on the theme in the leftmost column.

Theme	Interviewee							Requirement
	Board level, psychiatrist	Data manager	Information Architect	Data Analyst 1	Data Analyst 2	Data Analyst 3	Nurse Researcher	
Data sources		1	3		1	1		(1) Integrate data sources
Data standardization and preparation	7	1	6	12	2	6		(2) Preprocess data
Data storage			4	7	1	9		(3) Store data
Software and tooling	2	2	12	9	1	3	1	(4) Support various software and tooling packages
Coding best practices and documentation	3	2	2	3	1	3		(5) Support collaboration and documentation
Repeatability	1		1	2			1	(6) Enhance repeatability
Privacy and security	1					1		(7) Enhance privacy and security
Data process automation	3	3	3	6		4	1	(8) Automate data process
Healthcare practice applications	4	1	3	6		4	2	(9) Support analysis applications

able to identify requirements out of these codings, the open codings were then processed into broader categories by grouping codings based on their similarities and differences using Inductive Content Analysis (Elo and Kyngäs, 2008). This iterative process resulted in eleven themes that spanned multiple codes and could not be further combined. In a final step, all themes that were mentioned by more than one interviewee were reformulated into requirements. Table 1 shows the roles of the interviewees within the UMCU, and the nine requirements that were identified based on the interviews, including the number of codings for each interview and theme.

2.2 Requirements

The nine requirements that were identified based on the expert interviews are discussed below.

1. Integrate data sources. In a healthcare organization, typically multiple data sources exist (e.g. different database systems for structured patient records, unstructured text notes, correspondence, lab values, financial administration, genetics), that store data in their own format and do not necessarily communicate and interoperate with each other (Coorevits *et al.*, 2013; Nair, Hsu and Celi, 2016). Further integration with open data sources and patient-gathered data (e.g. from wearables or social media) can offer even better insights.

2. Preprocess data. Data preparation is a crucial step in the data analysis process that is relatively easy to model yet time consuming, especially if its steps are repeated for each separate analysis (Priest *et al.*, 2014; Wickham, 2014). Applying preprocessing steps (e.g. tidying, standardizing, reshaping and integrating data) in a central, collaborative manner thus saves time and effort for all researchers involved. There is on the other hand a tradeoff with flexibility, and researchers should also be enabled to make their individual choices in data preparation steps where needed.

3. Store data. Data that is gathered from various sources and then preprocessed should be accessible in a uniform format, allowing researchers to load necessary data into their tools for analysis (Apte *et al.*, 2011; Jensen, Jensen and Brunak, 2012).

4. Support various software and tooling packages. Data analysis teams are typically multidisciplinary, consisting for example of data analysts, health researchers, practitioners and statisticians (Lokhandwala and Rush, 2016), each applying a wide range of different techniques such as classical statistics, machine learning and data visualization (Katal, Wazid and Goudar, 2013). This leads to a variety of different software and tooling packages being used, which all need to be able to interoperate with a central infrastructure if adoption is to be achieved.

5. Support collaboration and documentation. Collaboration among researchers within a data analysis project is vital for obtaining both high quality data and analysis (Cheruvelil *et al.*, 2014; Priest *et al.*, 2014). This is mainly achieved by documentation and code collaboration (Wilson *et al.*, 2014), adoption of which is currently low in health care research (Murphy *et al.*, 2012). Documenting data firstly improves shared knowledge about data, a lack of which is one of the largest barriers for performing analysis, especially in health care (Lee *et al.*, 2015). Code collaboration secondly reduces redundancy and errors.

6. Enhance repeatability. Reproducible research is slowly becoming the norm in data-intensive scientific research (Peng, 2011), yet it is still not uncommon for researchers to be unable to recover data associated with their own published works (Goodman *et al.*, 2014; Pollard *et al.*, 2016). Data analysis in healthcare requires a reproducible workflow, which has well-recognized benefits, both internally (e.g. traceability of data, better insights into data provenance) and externally (e.g. better substantiation of results, enabling reuse of methods and results for others) (Johnson *et al.*, 2014; Wang and Hajli, 2017).

7. Enhance privacy and security. Healthcare data that are made available for research comprise sensitive data, that should be handled securely and with respect for patient privacy by design (Gil *et al.*, 2007; Kupwade Patil and Seshadri, 2014). Security-wise, restrictions on who can access which part of research datasets help prevent data leaks and unnecessary risks of patient re-identification. Regarding privacy, de-identification techniques (e.g. pseudonymization, de-identification of free-text variables, k-anonymity measures) are needed to mitigate impact on patient privacy (Menger *et al.*, 2017).

8. Automate data process. By automating all data processing steps, up-to-date EHR data becomes available periodically, without the need to perform additional time intensive operations before analysis is started. This additionally leads to better speed to decision (Wang *et al.*, 2017) and even better model learning (Lin and Haug, 2006).

9. Support analysis applications. The various applications of reusing EHR data, such as decision support, dashboarding, fundamental research, data

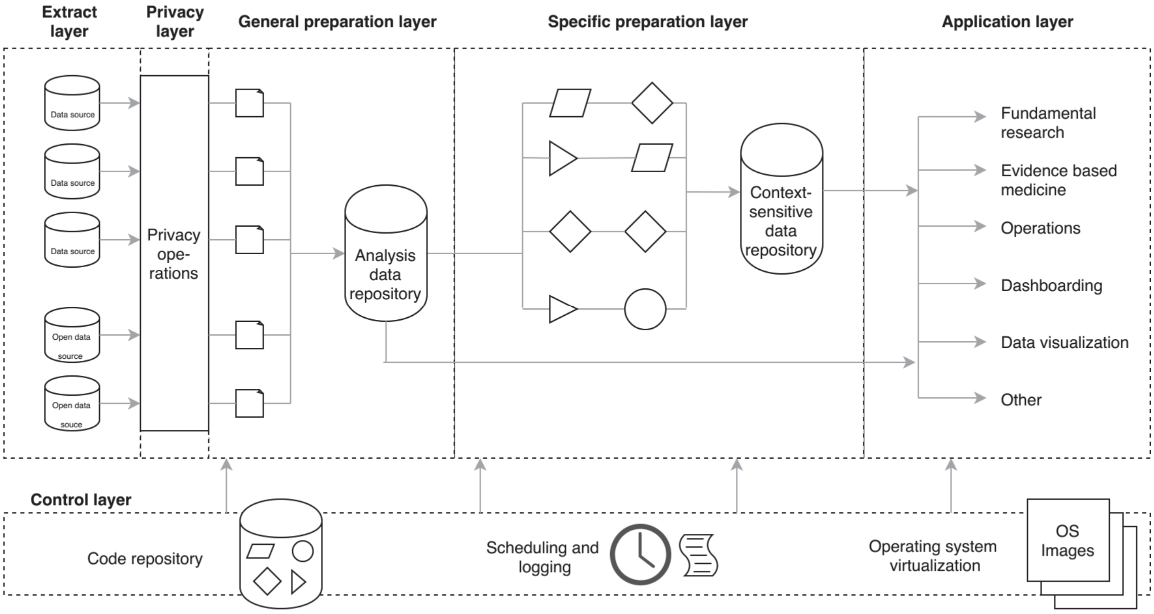

Figure 2: The CARED framework, consisting of five data processing layers from left to right, and an additional control layer that governs the data processing.

visualization, and several others (Chen, Chiang and Storey, 2012; Gandomi and Haider, 2015), should all be supported.

2.3 Framework Development

A conceptual framework for a data infrastructure was designed based on the nine requirements described above. As current data repositories are commonly based on Data Warehouse technology, the DWH model of Inmon et al. (Inmon, 2002) was used as a starting point. This model defines four data layers: the Data Source layer, the Staging layer, the Data Warehouse layer, and the Data Access layer. The layers in this model were iteratively refined, separated and combined, and new layers were added in order to meet all the nine requirements. Next, these layers were integrated in a single unifying framework. This design was presented and discussed in a focus group with the stakeholders in Table 1, aiming to demonstrate the framework and to evaluate it with respect to the nine requirements. A focus group is appropriate so that interaction between stakeholders is possible, so that all opinions about the framework can be explored, and so that all its potential issues are found (Gill *et al.*, 2008). One of the researchers facilitated the focus group, while the stakeholders were present to discuss the requirements and the framework. Comments mainly concerned the extent to which data preprocessing can be done in advance, the privacy steps that needed to be taken, and the viability of implementing infrastructure based on the framework in the UMCU. The participants attitude towards the framework was generally positive, and based on this focus group no major changes to the framework were introduced.

An infrastructure based on this framework was finally implemented in the Psychiatry Department of the UMCU. There, an initiative to bring data driven research to the daily practice resulted in some preliminary results (Menger *et al.*, 2016), but with no further supporting infrastructure in place, making it an ideal case for implementing the framework.

3 RESULTS

3.1 Framework

The CApable Reuse of EHR Data (CARED) framework we designed can be seen in Figure 2. It consists of five data layers (from left to right) in which data is processed, and a control layer for governing the data process.

In the data processing layers, first the *extract layer* connects to various internal and external data sources, and extracts the data in their own format. The subsequent *privacy layer* performs operations that help guarantee patient privacy, such as de-identification, pseudonymization and removal of non-consenting patients' data. Next, we propose to apply data preprocessing in two separate *general* and *specific* preparation steps. In the *general preparation layer*, the extracted and de-identified data from the several sources is transformed and tidying steps such as standardization, reformatting and reshaping are applied. The data remains semantically unchanged, meaning that only transformations concerning the format of data are performed. The cleaned data is stored in an analysis data repository, where it can be accessed by researchers through an Application Programmable Interface (API). In the *specific preparation layer*, advanced transformations that require domain knowledge, such as imputation, data integration and data enrichment are applied to data from the analysis data repository. After this phase, processed data are stored in a second, context-specific data repository that is also accessible through an API. In the final *application layer*, data from the analysis data repository or context-specific data repository are processed further by individual researchers according to their analysis purpose.

Additionally, the *control layer* drives the data through the data processing layers. This layer consists of three parts. Firstly, it contains a code repository where all code that transforms the data in or between layers is collaboratively written and maintained by all researchers involved. This creates a scientific workflow that traces data and code from source to application when guidelines for documenting code and data with version control are applied. Secondly, the scheduling and logging component makes sure that all clinical data is periodically extracted from the data sources and processed as specified in the code base. Reporting and notification ensure that errors in this process can be noticed and corrected. The execution of code thirdly is performed by using containerization on the operating system level. Images that provide all necessary software and libraries to execute code from the repository in a container are specified, so that the data process is not dependent on individual researchers software.

3.2 Evaluation

Below, we will describe how and in which parts of the framework the identified requirements are addressed. Between square brackets, the abbreviated layer(s) in which this requirement is satisfied is written (e=extract layer, p=privacy layer, gp=general preparation layer, sp=specific preparation layer, a=application layer, c=control layer).

1. Integrate data sources [e]. Multiple data sources are integrated in the extract layer, which allows flexibility with regard to adding or removing data sources.

2. Preprocess data [e, gp, sp]. We propose to divide the preprocessing of data into two steps: a general data preparation step and a subsequent specific data preparation step. In the first step, operations concerning the format of data are performed, and in the second step additional operations concerning the contents of data are performed. This distinction allows researchers to choose between two datasets with a different level of preparation, balancing between flexibility and time-efficiency. In the first case, all preprocessing operations are performed by the individual researcher, but can be tailored to specific needs, while in the second off-the-shelf preparation both the effort needed to start analysis and the likelihood of errors is reduced. In both cases, the researcher needs to perform final analysis-specific preparation in order to perform the analysis.

3. Store data [gp, sp]. Data are stored in an accessible location, in two analysis data and context-sensitive data repositories. An important requirement for data storage is that read- and write access can be provided for relevant software packages through an API. The data storage method can be subject to organizational and technical requirements, with options ranging from a shared drive to a database scheme (e.g. NoSQL) or distributed file systems. To ensure that previous versions of datasets are retrievable, snapshots of data can be stored using data differencing techniques.

4. Support various software and tooling [a, c]. The infrastructure is not dependent on specific software packages. This means that both running the data through the five data layers and performing analysis can be performed using any software package that can access data in the two repositories through the API.

5. Support collaboration and documentation [c]. Container images and documentation of code and data are shared in the code repository. Access to the central code repository for all researchers enables collaboration both in the data process and in specific research applications.

6. Enhance repeatability [e, gp, sp, a, c]. Firstly, the pipeline structure of the framework ensures that all data can be traced back to its source, providing data lineage for all eventual applications. Secondly, previous versions of data, code and operating system containers that are all stored together create a scientific workflow, which allows repeating analysis internally. By making the combination of these three items publicly available (e.g. along with a published result), analysis additionally becomes repeatable for

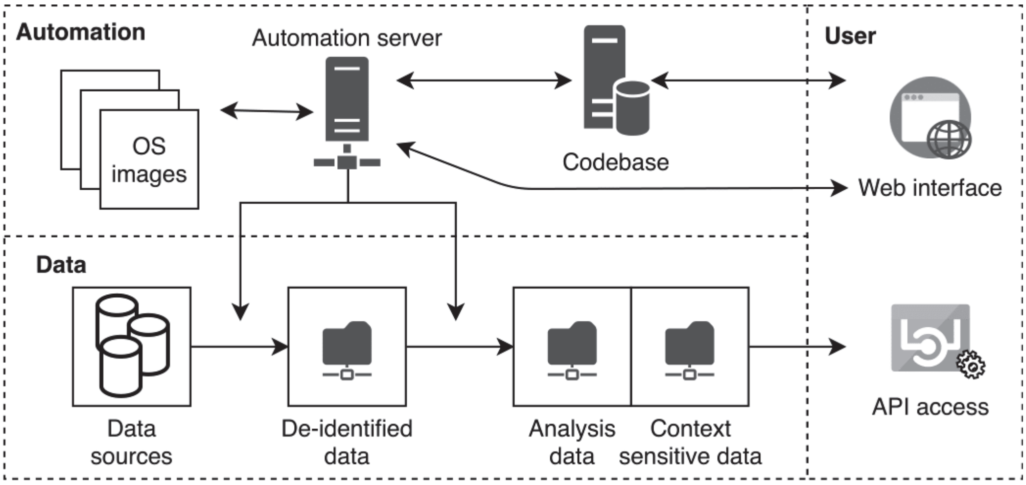

Figure 3: A schematic overview of infrastructure implemented in the Psychiatry Department of the UMCU based on the CARED framework. It consists of an automation server and codebase that can be accessed through a web interface, and several data components that can be accessed through an API.

the entire research commuity.

7. Enhance privacy and security [p]. Privacy of patients is enhanced by incorporating a separate privacy layer as one of the first data processing layers, ensuring that no datasets with identifying information proliferate in the analysis process. Security of data is enhanced by storing data in two central repositories that can be accessed by an API, preventing creation of copies of datasets on personal drives. Both the data access API and any other interfaces that allow interaction with the infrastructure can have restrictions on access for individual researchers, for example when only a subset of data is relevant for a specific analysis goal.

8. Automate data process [c]. All code that drives the data through the processing layers is available in the code repository, and can therefore be executed in its specified container at defined time intervals by an automation server. This ensures that data in both data repositories are periodically updated.

9. Support analysis applications [a]. The data in the two repositories can support a broad range of applications that is able to read data from one of the two data repositories through the API, such as data visualization, dashboards, (re)training of machine learning models, and decision support in the EHR.

3.3 Implementation

Based on this framework, an infrastructure was designed and implemented in the Psychiatry Department of the UMCU in a time period of 6 months. The subsequent programming of data processing pipelines took another approximate 6 months. Software components were implemented with open source packages such as GitLab, Docker, Jenkins, Python and R, supplemented with already present enterprise software to extract data from internal data sources. The hardware setup consists of a Linux and a Windows server, in order to allow interoperation with existing systems.

The implementation is schematically depicted in figure 3. The core of the implementation is the automation server, which updates data weekly by fetching code and OS containers and applying it to data. The data is extracted, de-identified and stored in a network folder. Then additional preprocessing steps are applied, and another network folder stores the data in analysis data and context sensitive data repositories. After mounting the network folder, researchers can access data in these repositories through an API based on their autorization. The code base and automation server are accessible through their web interfaces.

Our infrastructure has enabled several applications within the Psychiatry Department, three specific cases are highlighted below.

Personalized antipsychotics and antidepressants prescriptions. Patients with a psychotic or depressive disorder are often prescribed medication as part of their therapy. There are however various types of both antipsychotics and antidepressants, and choosing a drug and dose that improve an individual patients symptoms while minimizing side effects is mostly based on trial and error. During their lifetime, these patients typically switch medication and dose several times before the optimal combination is found. To smoothen this process, we developed visualizations of medication history based on information in the EHR, that is displayed to psychiatrists and patients. An overview of previous steps taken in finding an optimal prescription are comprehensively displayed, facilitating better decisions about next steps.

Prediction of inpatient violence incidents. During psychiatric admission, violence from patient directed at staff or other patients can occur. This topic has been thoroughly researched in psychiatry literature, yet a data driven approach had not been applied. We integrated admission data, textual data in the first 24 hours of admission, and violence incident reports, and then applied machine learning to train a classifier that is able to assess violence risk for individual patients, outperforming trained psychiatrists and other existing violence risk assessment tools (Menger, Scheepers and Spruit, 2018). This experiment is fully repeatable, while predictions can eventually be shown in the EHR.

Tracking patient enrolment status. Before a patient can start a planned admission or therapy, several administrative steps need to be taken that can take a span of multiple weeks or months. This includes for example obtaining referral documents from previous care organizations or a general practitioner, checking health insurance, and planning admission or therapy. Based on the current status of enrolment as written in text notes in the EHR, we therefore designed a status tracking system for patient enrolment. This leads to better insight into a current enrolment status, which benefits both patients and staff.

4 DISCUSSION

To fully realise the potential of analysing already existing EHR data, an infrastructure consisting of appropriate hardware and software components is needed, so that important technical, organizational and ethical challenges of reusing EHR data are mitigated. Current data repositories in healthcare organizations are often still based on DWH technologies, which fall short in addressing many of these challenges. Our CARED framework, designed based on requirements that were identified in the UMCU, provides a modern and unifying approach to infrastructure for EHR data reuse. It addresses important challenges, that are too often disregarded or solved in an ad hoc manner, such as analysing sensitive data with regard for patient privacy, repeatability of analysis, collaboration among researchers, and documentation of data and its analysis. Current research typically manages one or two of these challenges, while our research provides a framework that covers all the important aspects of reusing EHR data. We argue that adopting this framework improves quality of analysis, enhances patient privacy and data security, and aids efficient use of time, resources and skills. By providing a generic framework, we furthermore circumvent problems of interoperability with current IT systems, improving likelihood of its adoption. Adhering to the CARED framework when designing and implementing infrastructure in a healthcare organization will therefore be able to improve the state of data analytics research on secondary EHR data.

Implementing an infrastructure that is based on our proposed framework in the Psychiatry Department of the UMCU furthermore shows the feasibility of such a project. Although organizational factors caused some delays and practical difficulties, no fundamental setbacks were experienced. Additionally, we made use of existing open source software packages, leveraging knowledge and efforts from the extensive ecosystem of data analysis researchers. This is an important benefit that remains unaddressed in other software solutions. Modern data analysts are well versed in performing analysis using open source packages, typically implemented in the Python and/or R programming languages. Using such open software packages is a cost-effective measure that additionally lowers the threshold for data analysts from various domains to join the challenge of obtaining value from EHR data, which holds many promises for the future.

5 CONCLUSION

In this study, we used expert interviews to identify the most important requirements for an infrastructure for reusing EHR data, and subsequently designed the CApable Reuse of EHR Data (CARED) framework for infrastructure that addresses these challenges. The CARED framework we propose consists of five data processing layers: an *extract layer*, a *privacy* layer, two *preprocessing* layers, and an *application* layer. The framework is governed by a control layer, which consists of a code base where code and analysis is documented, a scheduler that automates the process, and containerization to make the analysis more robust and repeatable. We have elaborated upon the implementation of an infrastructure based on the proposed framework, showing its feasibility. Our study shows how an infrastructure based on the

CARED framework in place will improve the quality of analysis, enable types of analysis that are otherwise not possible, and aid efficient use of time, resources and skills.

REFERENCES

Apte, M. *et al.* (2011) 'Using Electronically Available Inpatient Hospital Data for Research', *Clinical and Translational Science*. Wiley-Blackwell, 4(5), pp. 338–345.

Badawi, O. *et al.* (2014) 'Making big data useful for health care: A summary of the inaugural MIT critical data conference', *Journal of Medical Internet Research*. JMIR Publications Inc., 16(8), p. e22.

Bauer, C. R. K. D. *et al.* (2016) 'Integrated data repository toolkit (IDRT): A suite of programs to facilitate health analytics on heterogeneous medical data', *Methods of Information in Medicine*, 55(2), pp. 125–135.

Chen, H., Chiang, R. H. L. and Storey, V. C. (2012) 'Business Intelligence and Analytics: From Big Data To Big Impact', *Mis Quarterly*, 36(4), pp. 1165–1188.

Cheruvelil, K. S. *et al.* (2014) 'Creating and maintaining high-performing collaborative research teams: The importance of diversity and interpersonal skills', *Frontiers in Ecology and the Environment*. Ecological Society of America, 12(1), pp. 31–38.

Coorevits, P. *et al.* (2013) 'Electronic health records: New opportunities for clinical research', *Journal of Internal Medicine*, 274(6), pp. 547–560.

Danciu, I. *et al.* (2014) 'Secondary use of clinical data: The Vanderbilt approach', *Journal of Biomedical Informatics*. NIH Public Access, 52, pp. 28–35.

Dean, B. B. *et al.* (2009) 'Review: Use of Electronic Medical Records for Health Outcomes Research', *Medical Care Research and Review*, 66(6), pp. 611–638.

Elo, S. and Kyngäs, H. (2008) 'The qualitative content analysis process', *Journal of Advanced Nursing*, 62(1), pp. 107–115.

Fleurence, R. L. *et al.* (2014) 'Launching PCORnet, a national patient-centered clinical research network', *Journal of the American Medical Informatics Association*, 21(4), pp. 578–582.

Friedman, C. *et al.* (2014) 'Toward a science of learning systems: a research agenda for the high-functioning Learning Health System', *Journal of the American Medical Informatics Association*. Oxford University Press, 22(1), pp. 43–50.

Gandomi, A. and Haider, M. (2015) 'Beyond the hype: Big data concepts, methods, and analytics', *International Journal of Information Management*. Pergamon, 35(2), pp. 137–144.

George, J., Kumar, B. V. and Kumar, V. S. (2015) 'Data Warehouse Design Considerations for a Healthcare Business Intelligence System', in *Proceedings of the World Congress on Engineering 2015*, pp. 4–7.

Gil, Y. *et al.* (2007) 'Privacy enforcement in data analysis workflows', in *CEUR Workshop Proceedings*. CEUR-WS.org, pp. 46–53.

Gill, P. *et al.* (2008) 'Methods of data collection in qualitative research: interviews and focus groups', *Bdj*. Nature Publishing Group, 204(6), pp. 291–295.

Goodman, A. *et al.* (2014) 'Ten Simple Rules for the Care and Feeding of Scientific Data', *PLoS Computational Biology*. Public Library of Science, 10(4), p. e1003542.

Groves, P. *et al.* (2013) 'The "big data" revolution in healthcare: accelerating value and innovation', *McKinsey Global Institute*. Center for US Health System Reform Business Technology Office, (January), pp. 1–22.

Hammami, R., Bellaaj, H. and Kacem, A. H. (2014) 'Interoperability of healthcare information systems', in *2014 International Symposium on Networks, Computers and Communications, ISNCC 2014*.

Hazlehurst, B. L. *et al.* (2015) 'CER Hub: An informatics platform for conducting comparative effectiveness research using multi-institutional, heterogeneous, electronic clinical data', *International Journal of Medical Informatics*. Elsevier, 84(10), pp. 763–773.

Hersh, W. R. *et al.* (2013) 'Recommendations for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research', *eGEMs (Generating Evidence & Methods to improve patient outcomes)*. Academy Health, 1(1), p. 1018.

Hersh, W. R. *et al.* (2014) 'Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research', *Medical care*. NIH Public Access, 51(August), pp. S30–S37.

Inmon, W. H. (2002) *Building the data warehouse*, *John Wiley & Sons, Inc.*

Jensen, P. B., Jensen, L. J. and Brunak, S. (2012) 'Mining electronic health records: towards better research applications and clinical care', *Nature Reviews Genetics*. Nature Publishing Group, 13(6), pp. 395–405.

Johnson, K. E. *et al.* (2014) 'How the provenance of electronic health record data matters for research: a case example using system mapping.', *EGEMS (Washington, DC)*. Academy Health, 2(1), p. 1058.

Kasthurirathne, S. N. *et al.* (2015) 'Enabling Better Interoperability for HealthCare: Lessons in Developing a Standards Based Application Programing Interface for Electronic Medical Record Systems', *Journal of Medical Systems*, 39(11).

Katal, A., Wazid, M. and Goudar, R. H. (2013) 'Big data: Issues, challenges, tools and Good practices', in *2013 6th International Conference on Contemporary Computing, IC3 2013*. IEEE, pp. 404–409.

Krishnankutty, B. *et al.* (2012) 'Data management in clinical research: An overview.', *Indian journal of pharmacology*. Wolters Kluwer -- Medknow Publications, 44(2), pp. 168–72.

Kupwade Patil, H. and Seshadri, R. (2014) 'Big Data Security and Privacy Issues in Healthcare', in *2014 IEEE International Congress on Big Data*. IEEE, pp. 762–765.

Lee, E. S. *et al.* (2015) 'Characterizing Secondary Use of Clinical Data', *AMIA Summits on Translational Science*

Proceedings. American Medical Informatics Association, 2015, pp. 92–96.

Lin, J. and Haug, P. J. (2006) ‘Data Preparation Framework for Preprocessing Clinical Data in Data Mining’, *AMIA Annual Symposium proceedings*.

Lokhandwala, S. and Rush, B. (2016) ‘Objectives of the secondary analysis of electronic health record data’, in *Secondary Analysis of Electronic Health Records*. Cham: Springer International Publishing, pp. 3–7.

Lu, Z. and Su, J. (2010) ‘Clinical data management: Current status, challenges, and future directions from industry perspectives’, *Open Access Journal of Clinical Trials*.

McCarty, C. A. *et al.* (2011) ‘The eMERGE Network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies’, *BMC Medical Genomics*, 4(1), p. 13.

Menger, V. *et al.* (2016) ‘Transitioning to a Data Driven Mental Health Practice: Collaborative Expert Sessions for Knowledge and Hypothesis Finding’, *Computational and Mathematical Methods in Medicine*, 2016.

Menger, V. *et al.* (2017) ‘DEDUCE: A pattern matching method for automatic de-identification of Dutch medical text’, *Telematics and Informatics*.

Menger, V., Scheepers, F. and Spruit, M. (2018) ‘Comparing Deep Learning and Classical Machine Learning Approaches for Predicting Inpatient Violence Incidents from Clinical Text’, *Applied Sciences*. Multidisciplinary Digital Publishing Institute, 8(6), p. 981.

Meystre, S. M. *et al.* (2017) ‘Clinical Data Reuse or Secondary Use: Current Status and Potential Future Progress’, *Yearbook of Medical Informatics*. Georg Thieme Verlag KG, 26(01), pp. 38–52.

De Moor, G. *et al.* (2015) ‘Using electronic health records for clinical research: The case of the EHR4CR project’, *Journal of Biomedical Informatics*, 53, pp. 162–173.

Murdoch, T. B. and Detsky, A. S. (2013) ‘The Inevitable Application of Big Data to Health Care’, *Jama*. American Medical Association, 309(13), p. 1351.

Murphy, S. N. *et al.* (2010) ‘Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2)’, *Journal of the American Medical Informatics Association*. American Medical Informatics Association, 17(2), pp. 124–130.

Murphy, S. N. *et al.* (2012) ‘Current State of Information Technologies for the Clinical Research Enterprise across Academic Medical Centers’, *Clinical and Translational Science*. Wiley-Blackwell, 5(3), pp. 281–284.

Nair, S., Hsu, D. and Celi, L. A. (2016) ‘Challenges and opportunities in secondary analyses of electronic health record data’, in *Secondary Analysis of Electronic Health Records*. Cham: Springer International Publishing, pp. 17–26.

Obermeyer, Z. and Lee, T. H. (2017) ‘Lost in Thought — The Limits of the Human Mind and the Future of Medicine’, *New England Journal of Medicine*. Massachusetts Medical Society, 377(13), pp. 1209–1211.

Peek, N., Holmes, J. H. and Sun, J. (2014) ‘Technical challenges for big data in biomedicine and health: data sources, infrastructure, and analytics’, *Yearbook of medical informatics*.

Peng, R. D. (2011) ‘Reproducible Research in Computational Science’, *Science*. American Association for the Advancement of Science, 334(6060), pp. 1226–1227.

Pollard, T. *et al.* (2016) ‘Data preparation’, in *Secondary Analysis of Electronic Health Records*. Cham: Springer International Publishing, pp. 101–114.

Priest, E. L. *et al.* (2014) ‘Developing electronic data methods infrastructure to participate in collaborative research networks.’, *EGEMS (Washington, DC)*. Academy Health, 2(1), p. 1126.

Priyanka, K. and Kulennavar, N. (2014) ‘A survey on big data analytics in health care’, *IJCSIT*, 5(4), pp. 5865–5868.

Raghupathi, W. and Raghupathi, V. (2014) ‘Big data analytics in healthcare: promise and potential’, *Health Information Science and Systems*. BioMed Central, 2(1), p. 3.

Rea, S. *et al.* (2012) ‘Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: The SHARPn project’, *Journal of Biomedical Informatics*. NIH Public Access, 45(4), pp. 763–771.

Roski, J., Bo-Linn, G. W. and Andrews, T. A. (2014) ‘Creating value in health care through big data: Opportunities and policy implications’, *Health Affairs*. Project HOPE - The People-to-People Health Foundation, Inc., 33(7), pp. 1115–1122.

Safran, C. (2014) ‘Reuse Of Clinical Data’, *IMIA Yearbook*. Schattauer Publishers, 9(1), pp. 52–54.

Strauss, A. and Corbin, J. (1990) *Basics of Qualitative Research*, *Newbury Park, CA: Sage*.

Wang, Y. *et al.* (2017) ‘An integrated big data analytics-enabled transformation model: Application to health care’, *Information & Management*.

Wang, Y. and Hajli, N. (2017) ‘Exploring the path to big data analytics success in healthcare’, *Journal of Business Research*. Elsevier, 70, pp. 287–299.

Weber, G. M. *et al.* (2009) ‘The Shared Health Research Information Network (SHRINE): A Prototype Federated Query Tool for Clinical Data Repositories’, *Journal of the American Medical Informatics Association*. American Medical Informatics Association, 16(5), pp. 624–630.

Wickham, H. (2014) ‘Tidy Data’, *Journal of Statistical Software*, 59(10), pp. 1–23.

Wilson, G. *et al.* (2014) ‘Best Practices for Scientific Computing’, *PLoS Biology*. Edited by J. A. Eisen. Public Library of Science, 12(1), p. e1001745.

Youssef, A. E. (2014) ‘A Framework for secure healthcare systems based on big data analytics in mobile cloud computing environments’, *International Journal of Ambient Systems and Applications*.