MOSAIC – A Modular Approach to Data Management in Epidemiological Studies

M. Bialke[1]; T. Bahls[1]; C. Havemann[1]; J. Piegsa[1], K. Weitmann[1]; T. Wegner[2]; W. Hoffmann[1]

[1]Institute for Community Medicine, Section Epidemiology of Health Care and Community Health, University Medicine Greifswald, Greifswald, Germany;
[2]Institute of Applied Microelectronics and Computer Engineering, University of Rostock, Rostock, Germany

Keywords
Medical data management, data privacy protection, informed consent, pseudonyms, record linkage

Summary
Introduction: In the context of an increasing number of multi-centric studies providing data from different sites and sources the necessity for central data management (CDM) becomes undeniable. This is exacerbated by a multiplicity of featured data types, formats and interfaces. In relation to methodological medical research the definition of central data management needs to be broadened beyond the simple storage and archiving of research data.
Objectives: This paper highlights typical requirements of CDM for cohort studies and registries and illustrates how orientation for CDM can be provided by addressing selected data management challenges.
Methods: Therefore in the first part of this paper a short review summarises technical, organisational and legal challenges for CDM in cohort studies and registries. A deduced set of typical requirements of CDM in epidemiological research follows.
Results: In the second part the MOSAIC project is introduced (a modular systematic approach to implement CDM). The modular nature of MOSAIC contributes to manage both technical and organisational challenges efficiently by providing practical tools. A short presentation of a first set of tools, aiming for selected CDM requirements in cohort studies and registries, comprises a template for comprehensive documentation of data protection measures, an interactive reference portal for gaining insights and sharing experiences, supplemented by modular software tools for generation and management of generic pseudonyms, for participant management and for sophisticated consent management.
Conclusions: Altogether, work within MOSAIC addresses existing challenges in epidemiological research in the context of CDM and facilitates the standardized collection of data with pre-programmed modules and provided document templates. The necessary effort for in-house programming is reduced, which accelerates the start of data collection.

Correspondence to:
Martin Bialke
Institute for Community Medicine
Department Epidemiology of Health Care
and Community Health
University Medicine Greifswald
Ellernholzstr. 1–2
17487 Greifswald
Germany
E-mail: martin.bialke@uni-greifswald.de

Methods Inf Med 2015; 54: 364–371
http://dx.doi.org/10.3414/ME14-01-0133
received: December 5, 2014
accepted: June 3, 2015
epub ahead of print: July 21, 2015

1. Introduction

The collection and provision of medical data forms the basis for an analytical medical research. Therefore, cohort studies and registries mainly focus on quality-assured collection of primary research data and associated metadata as well as preservation of data interpretability.

When designing cohort studies and registries many research projects are confronted with considerable challenges (►Figure 1). Considering the phases of the data lifecycle [1], this concerns organisational and technical effort necessary for the realization of a comprehensive data management including collection, processing, long-term storage and provision of data pursuant to recommendations of accredited institutions (e.g. the DFG [2]). One particular challenge is the integration of data from heterogeneous source systems. This includes laboratory information and management systems (LIMS), clinical information systems (CIS), diagnostic equipment and electronic case report forms (eCRF) usually generating a wide variety of data types and formats (e.g. form or image data, CSV exports, GDT files, XML structures, HL7 messages). Consequently, suitable measures must be adopted in order to ensure data quality throughout processing activities, to enable automated validation and to support interactive data correction. In addition, compliance with legislation for data protection must be assured at both federal and state level. For this reason, the ethics model must be built around a central consent management system administrating individual items of consent and authorisation (i.e. informed consent) and permitting revocation cross-checking on a daily basis. Management of participants is required in order to aggregate personal medical data within a central data repository while avoiding mistakes due to homonyms or synonyms. Furthermore, the ability to process and provide research data in a pseudonymised form is mandatory.

Therefore, the introduction of a central data management (CDM) aids to simplify the process of design and implementation of data management for cohort studies and registries. This is accomplished by providing structured, uniform and reproducible processes throughout all phases of the lifecycle of research data. Simultaneously, preconditions for long-term data usability and the comparability of derived results are constituted. More precisely, the use of CDM allows for the integration of all core issues for data management into cohort studies and registries as early as the planning stage [3]. In order to reduce effort, the deployment of reusable, modular solutions is encouraged. Thereby, the high effort arising from the repeated procedure of developing concepts and implementing dedicated software solutions for individual projects is avoided.

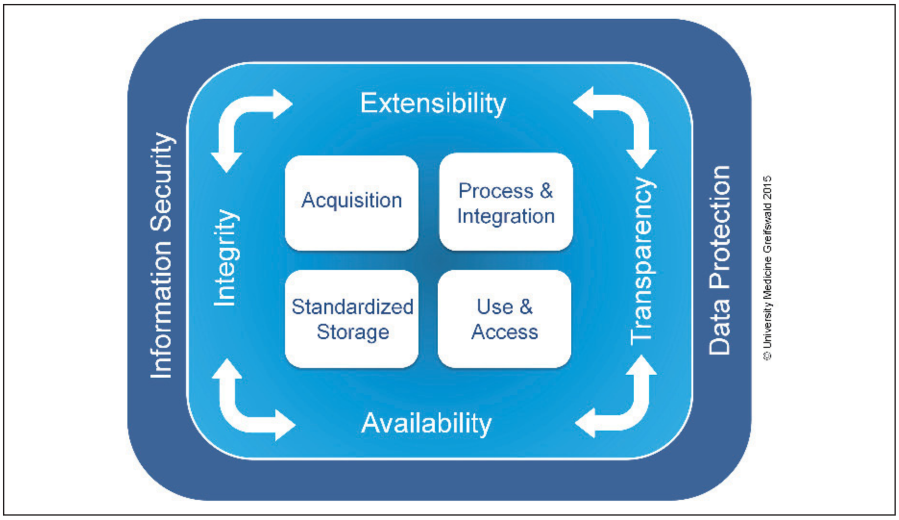

Figure 1 CDM in the context of epidemiological cohort studies and registries: CDM comprises all technical and organisational measures aiming for data acquisition, data processing and integration, the standardized storage as well as the use and access of data. At the same time CDM has to fullfil non-functional requirements and relating interdependencies, such as requirements for data integrity, data availability, process transparency and system extensibility. Also CDM has to ensure conformity to legal standards of information security and data protection.

2. Objectives

Aim of this paper is to highlight typical requirements of CDM for cohort studies and registries, deduced from legal aspects, organisational task areas and technical approaches of data management typically applied within the phases of epidemiological projects. Based on cohort studies and registries, the paper illustrates how MOSAIC provides orientation for CDM by addressing selected data management challenges with a first set of ready-to-use and open source tools. MOSAIC is introduced to support the implementation of CDM for cohort studies and registries and to facilitate reducing in-house programming effort as well as promoting sustainability of existing solutions.

3. Methods

3.1 Central Data Management in Epidemiological Research

In cohort studies and registries many organisational and technical tasks must be completed prior to actual operation. ►Figure 2 depicts both organisational and a technical measures and parameters that need to be considered during the preparation phase and the subsequent phases of acquisition and usage. Starting with the specification of research questions and variables to be surveyed, the ethical framework (including determination of patient details and informed consent (e.g. paper-based or as digital document) for the study must be established first (**preparatory phase**). Additionally, appropriate strategies for implementing authorisations and revocations, and the definition of workflows to support these activities are determined. In order to guarantee the consistent participant management and the necessary privacy an unambiguous identification of persons and the pseudonymisation of personal data must be ensured [4]. Subsequently, the necessary techniques for data collection, processing, archiving and provisioning must be defined. Furthermore, the technical environment necessary to ensure protection of data privacy and information security must be specified. Specifications must also be drawn up for separation and storage of personal data and medical data. This comprises a data model capable of mapping the collected information, a data repository for information storage and a role-based management system for access authorisation, to name but a few. Moreover, a data protection concept must be specified considering organisational, technical and personnel-related issues. This concept requires review and approval by the responsible data protection officer.

For the collection of research data, appropriate electronic case report forms (eCRFs) must be designed reverting to standard data collection tools including the data validation and possible data correction (automated and interactive). Amongst others, required interfaces, expected data formats, necessary metadata and data transfer protocols must be defined for data extraction from diagnostic equipment and other sources. In addition, measures to support record linkage, in order to minimize synonym and homonym errors during the data merging process [5], as well as to secure the data transfer between the study centres and the target storage system require specification.

During the subsequent **acquisition phase**, the focus is on ensuring high quality of the collected research data and on guaranteeing its security and long-term provisioning. Related tasks include the unobstructed operation of all systems required for error-free data collection as well as the definition and implementation of quality assurance measures (e.g. instructing study personnel or source data checking and correction). At the same time, a high level of data integrity needs to be secured (i.e. by matching procedures for aggregation of personal data from multiple sources). Mechanisms for historisation and version control – aided by continuous monitoring and documentation of all data-handling processes – guarantee the necessary auditability of the data processing systems. In addition, backup strategies and rules regulating data access guarantee data security during this phase. To facilitate follow-up data collection, CDM also supports the recontacting of study participants.

During the **usage phase** the collected medical data is provided to researchers while complying with stated requirements for data privacy and protection. This comprises well-defined procedures for use and access and study-specific pseudonymisation. Furthermore, free specification of variable sets and export formats is provided to researchers. A preferably automated procedure is applied to determine and request the associated pseudonymised data, following prior verification of the necessary consent. Subsequently, the requested data is transmitted via a dedicated transfer unit. On completion of a study, a similar approach must be used to re-integrate the research results into the data pool. This includes derived variables, generated scores or variable coding work, for example. In the case of subsequent consent revocation or incidental findings subject to reporting obligations, the system must be capable of re-establishing the link from the data to individuals by a defined de-pseudonymisation procedure.

In summary, the study lifecycle requires a wide range of processes and measures that are not limited to separate phases within an individual study. On account of the considerable technical effort required to realize the necessary systems and functionalities, implementation should commence at the earliest possible juncture. However, the system architecture must exhibit sufficient versatility to satisfy supplementary requirements, changes and extensions in the study design.

Epidemiological studies and registries show that individual research projects exhibit particular requirements and settings [6–9]. Nevertheless, implementing a data management system as part of a study generates a recurrent set of similar issues. Such issues can be categorized to functional aspects, non-functional requirements and establishing compliance with legal frameworks. Fundamentally, CDM is targeted on satisfying each of these requirements.

Meyer et al. [10] define non-functional requirements for a CDM system. The authors are particularly focussing on data management processes. Reverting to this, their implementation is described by suitable technical measures with focus on quality assurance concentrating on the processed data.

In the context of epidemiological cohort studies these technical, organisational and staffing measures required to implement data management can be grouped into primary task areas. Combined with the non-functional requirements, they constitute the most important core elements of CDM. Details on the individual measures that each core element typically involves are provided in ▶Table 1.

Since the primary task of CDM is to aggregate data from heterogeneous and federated sources and to transform these data into suitable structures, the focus is on ensuring that the collection, processing, storage and provisioning of data is standardized and homogenized and coordinated from a single central point.

Therefore, the successful implementation of CDM in an epidemiological context is conditional on expert support (i.e. by com-

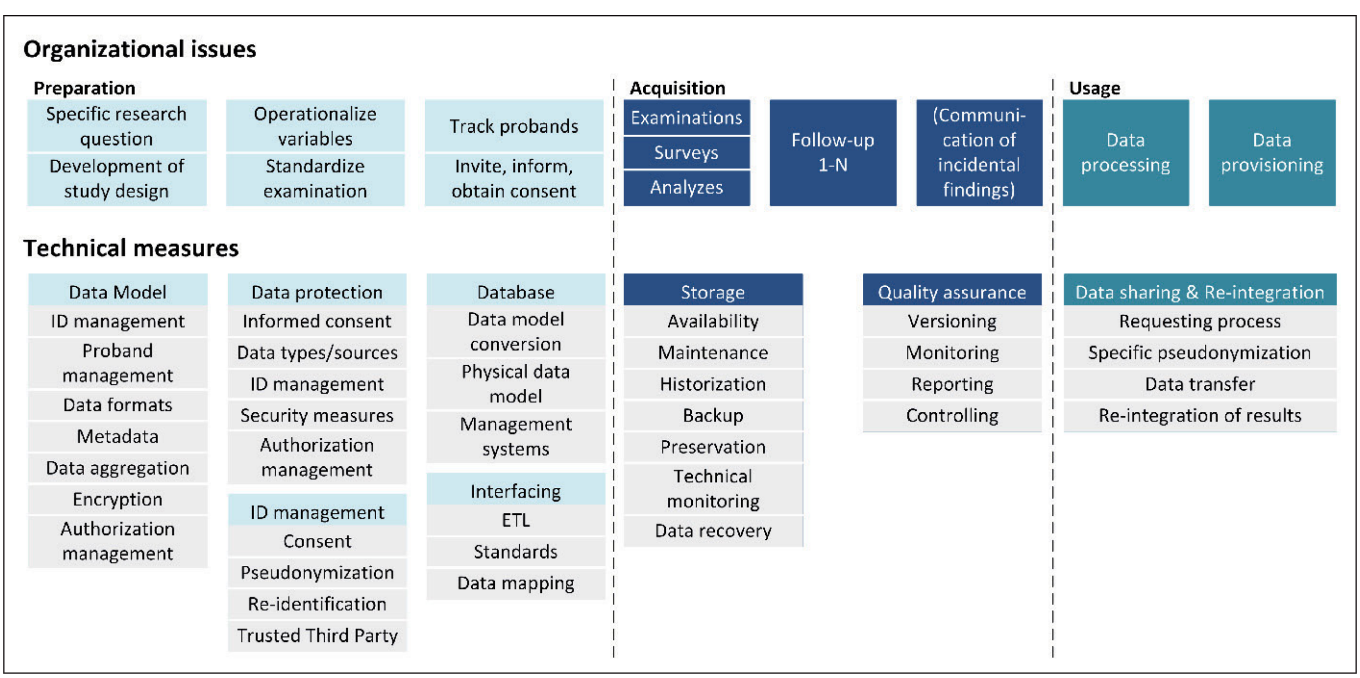

Figure 2 Typical phases of a cohort study: Organisational issues and technical measures concerning CDM regarding the phases of a cohort study

puter scientists and IT personnel) for the research. In addition, data management must be included in study planning at an early stage. Ideally, implementation of the technical and organisational requirements should be completed before data collection starts.

4. Results

4.1 A Modular Approach to Support the Implementation of CDM

In the context of the MOSAIC project (DFG program "Information infrastructure for research data") a set of mutually independent tools is developed, each of them addressing a practical requirement of data management in cohort studies and registries (▶Figure 2).

Primarily targeting newly initiated cohort studies and registries, the MOSAIC web-portal (mosaic-greifswald.de) presents guidance and insights to legal, organisational and technical aspects of CDM. Scientists' attention is drawn to relevant literature, existing solutions (provided and recommended e.g. by the TMF [12]) and current issues, assisting and encouraging them to specify the individual requirements for their study setting. Addressing these requirements and in accordance to the rules the TMF consented upon with the German protection officer [12], MOSAIC provides a portfolio of independent software modules, templates, checklists and recommendations in order to facilitate re-use of existing solutions. For already established cohort studies and registries, some MOSAIC tools might be of interest as well, e.g. if previously defined requirements have changed and new technical ways for participant management or pseudonymisation of existent medical data have to be identified.

The aim of MOSAIC is to provide support for the responsibilities of CDM through modular solutions (▶Figure 3). This approach allows for focusing on selected challenges instead of establishing complete solutions for a data management lacking adaptability, interoperability and flexibility.

A particular benefit of modularity is enhanced re-usability. The tools might be utilized separately in one epidemiological project to address a specific issue or used in combination in another project to satisfy a set of requirements. Nevertheless, MOSAIC does not intend to provide a complete software suite serving as a stand-alone solution for CDM. Each tool is or will be developed based on long-time experience in data management and in close cooperation with potential users. For this purpose, MOSAIC seeks to cooperate with newly initiated and already existing cohort studies and registries to acquire knowledge about specific needs and perceptions resulting from individual settings. The first set of tools, already available from the MOSAIC portal, is derived from solutions developed within existing projects [6–9]. Taking usability improvements for individual software tools as one example, the target was to reduce the number of configuration steps and to simplify the integration into existing infrastructures. Ease-of-use is improved by deploying web-based graphical user interfaces and by adding documentation (e.g. quick start guides, developer documentation and brochures).

To feature a high-level overview necessary for planning, designing and implementing CDM, a **web-based reference platform** [13] is offered. Depicting the necessary steps and typical issues occurring during the individual phases, the interactive reference platform aims to concentrate existing topical knowledge from the research community. For this purpose, based

Table 1 Functional and non-functional elements of CDM

	Core Element	Description
Functional requirements	Acquisition	Specification of data sources, data formats, interfaces and data transfer methods
	Processing and integration	Implementation of data integration processes, metadata enrichment and data quality assurance
	Standardized storage	Planning and implementation of a generic data model and provisioning of the necessary IT infrastructure
	Use and access	Options for data exploration, coordination of the data request process, and the import/export of data
Non-functional requirements	Integrity	Comprises measures for guaranteeing data consistency, security and protection
	Transparency	Surveillance plus continuous monitoring of processes and measures ensuring end-to-end traceability and reproducibility
	Extensibility	Hardware/Software must be readily extensible to accommodate expansion of the study and a higher level of requirements
	Availability	High level of long-term reliability for collection, storage, archiving and provisioning of data
	Data protection	Separation of identifying and medical data at the earliest possible juncture in accordance with data protection legislation. If participants must be uniquely identifiable, the informed consent and revocation documentation must be effectively managed. Pseudonymisation and anonymisation of the data must be possible.
	IT and information security	As defined by the BSI's Baseline Protection Catalogue [11]. Comprises measures for authentication, authorisation, secure data transmission, encrypted data storage and the hardening of network infrastructure against unauthorized access.

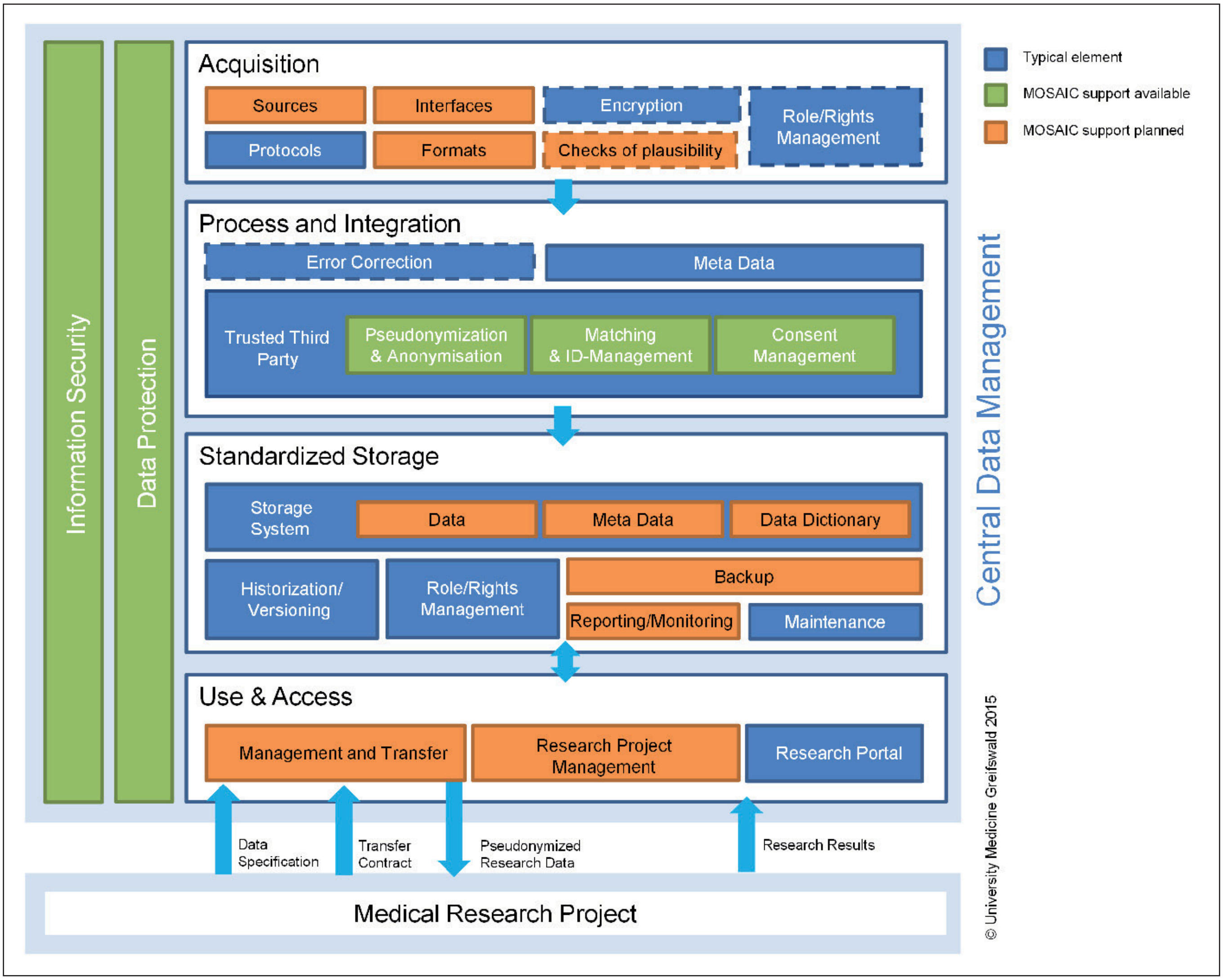

Figure 3 Functional requirements for CDM in an epidemiological context: typical responsibilities and specialized MOSAIC tools covering selected issues

on structured questions and answers, users are encouraged to rate existing solutions and to share their issues and experiences. The reference platform is based on the open source question-and-answer framework Question2Answer [14] and is completely integrated into the MOSAIC web-portal.

Setting up a new epidemiological project typically includes drafting a **data protection concept** note. To provide a starting point and to prevent missing essential data protection issues, MOSAIC shares a **document template** (DOCX-Format) to simplify the process of writing. Structured questions and recommendations guide potential authors towards the integration of study-specific attributes into a pre-defined, pre-formatted concept paper. Alongside a conventional and directly deployable version of the document, an additional "generator" version enables interactive modification of the document template structure regarding mandatory, selectable and optional components beforehand. The document-generator utilizes several preformatted Microsoft Word building blocks, macros and configuration scripts optimized for the use with Microsoft Office 2010.

To avoid duplicate participant entries, the **ID Management solution E-PIX** (Enterprise Identifier Cross Referencing) applies the Fellegi-Sunter algorithm [15] and the Levenshtein distance. The independent software module allows for unambiguous participant management and efficient aggregation of medical research data from federated study centres. Additionally, the correction of potential synonym errors is supported (i.e. false-negative record linkage). The E-PIX, as well as the subsequent tools, applies a service-oriented architecture to provide all functionalities via web services. It was developed using Java EE and several development frameworks, including PrimeFaces [16] for a web-based user-interface.

Before research data can be collected or provided within a cohort study or registry, legal conditions require checking for available consents or revocations from the specific participant. To manage both digital and paper-based informed consents MOSAIC offers the **Consent Management**

solution gICS (generic Informed Consent Administration Service) [17], which allows to check for various policies and modules of a consent automatically in real time. Comprehensive drafting of the consents and their validation through an ethics committee is assumed beforehand.

In cohort studies and registries, data storage requires pseudonymisation of each data record. Mostly the use of different pseudonyms in separate study centres or for different data categories (e.g. specimen, image data and medical data) is mandatory. The software module **gPAS** (generic Pseudonym Administration Service) [18] **generates and administers appropriate pseudonyms** using non-deterministic pseudonyms for arbitrary alphanumeric sequences. Additionally it allows defining domain-specific alphabets and generator algorithms as required and offers functions for de-pseudonymisation and anonymisation.

Depicting only selected disciplines in CDM for epidemiological research yet, the presented tools primarily facilitate ensuring conformity to legal data protection regulations. Current and future work within MOSAIC focusses on remaining aspects of CDM (▶Figure 3, "MOSAIC support planned"). This comprises a solution for standardised storage, including an EAV-based metadata repository allowing for free definition of hierarchies between study items and the suitable research data repository. Supplementary ready-to-use examples will be provided to demonstrate how to use the storage solution with the open source EDC-Software OpenClinica [19], how to integrate research data from external devices (e.g. laboratory devices) and recommendations as well as checklists for an enhanced data protection strategy. Furthermore example reports for a basic quality assurance are elaborated using the open source statistical computing library R [20] allowing for an automated evaluation of metric and categorical study items. Also the preparation of document templates, in order to support the procedure of data provision, is intended.

All MOSAIC tools are made available on the project portal (mosaic-greifswald.de) and the mosaic subversion repository [21] under open source licensing. This applies to templates and documentation (CC BY 4.0 [22]) as well as software (AGPLv3 [23]).

5. Discussion

The growing relevance of large-scale research networks in the scientific community makes the deployment of uniform methodologies and the generation of homogeneous data essential, since this allows for efficient data pooling and facilitates comparability. Consequently, work on the MOSAIC project complies with the recommendations of the German Research Foundation (DFG) for secure storage and provisioning of digital research data [2].

Previous work by Fraser et al. [24] has shown that the re-use of specially developed tools is capable of cutting costs and effort involved for data collection in separate studies. Moreover, utilizing the proposed set of MOSAIC tools increases the efficiency and standardisation of individual work within cohort studies and registries resulting from methodology streamlining. This standardisation facilitates an increased data quality [25].

The MOSAIC project's approach to provide modular solutions for specific requirements addresses the need for practical tools, in terms of readily deployable templates and software, in order to support planning, design and implementation of CDM for cohort studies and registries. Especially in all cases where the scientific environment offers only marginal experience in data management, lacks access to IT personnel or suffers from insufficient resources in terms of software development.

Implementation of the respective tools demands precise knowledge of developments in the research community in order to avoid conducting primary development work in parallel to existing solutions. In case proven solutions existed for well-defined issues, the extent to which these solutions meet the requirements of epidemiological cohort studies and registries and the degree of applicability was assessed.

The TMF already provides a Guideline for Data Protection in Medical Research Projects (published in 2006, updated in 2014) [12]. Though it presents an introduction, guidance and recommendations to all aspects of data protection and ethical issues, it lacks an easy to use document template, which actually supports the respective author to write a data protection concept note. This aspect is well addressed by the MOSAIC template (published 2013), providing a structured starting point for essential typical data protection issues to be answered in the latter process. Using a question-based approach supports the application in various study or registry scenarios. Applying pre-worded text-blocks would narrow the scope of potential users, e.g. the data protection template of the Open source Registry System for Rare Diseases in the EU (published 2014 within OSSE, [26]) fits only for registries using the OSSE Software, but actually accelerates drafting a data protection concept note.

Several tools for the management of participants exist in the scientific community. For example the TMF PID-Generator (published 2005) [27] allows for merging participant identifying data from federated study sites even if the data sets are incomplete or faulty. The Java-based Mainzel List (published 2013) [28] aims for a more contemporary approach. It comes with a likely set of functionality, but with an easy to use REST-interface, which facilitates simplified system integration. However, unlike the E-PIX (published 2014), both systems are not yet capable of managing multiple local identifiers and identities for each participant and lack a graphical user interface to support the respective user in detecting and solving possible synonym errors.

Also for pseudonymisation and de-pseudonymisation a well-established tool exists. The TMF Pseudonymisation Service (PSD, published 2010, [29]) generates pseudonyms with a fixed length and alphabet (depending on the selected encryption method) using a synchronous algorithm. Thus the storage of associated value-pairs (original value and associated pseudonym) is not necessary. As a consequence the anonymisation of a medical dataset by simply deleting the association, which connects original value and pseudonym, is not possible. However gPAS (published 2013) [18] provides this mechanism and additionally facilitates the creation of pseudonym hierarchies. The generation of multiple pseudonyms for one participant allows a context-related pseudonymisation e.g. to use specific pseudonyms for different data sources (eCRF, specimen, MRI) or for data provision during the use and access process.

Unlike the TMF Informed Consent Wizard (published 2007) [30], gICS does not provide textual support for drafting a document. Neither gICS focusses clinical practise nor it is limited to a simplified file-based approach of HL7-documents like the Consent Management Suite (COMS, published 2011) [31]. gICS (published 2014) focusses on the management of informed consent documents using re-usable policies to build modules, providing the necessary service functionalities to support automatic checks within seconds, e.g. whether a participants allows to share his specimen or has revoked his consent.

The developed MOSAIC tools will be evaluated in cooperation with the users in order to assess their efficacy and acceptance. One particular evaluation element is a dynamic review from a functional perspective. This review is conducted by deploying the tools to current and future external and internal projects enabling direct tool modifications and enhancements. Another element is tool testing conducted by external partners. The template for drafting data protection concepts has already been used in several research projects such as the TORCH registry of the DZHK or the MonDAFIS Study of the Centre for Stroke Research Berlin at the Charité (CSB). The modules for ID-Management, pseudonym administration and informed consent administration have recently been used to set up trusted third parties as a substantial part of CDM in compound projects such as the German Centre for Cardiovascular Research [9] or the German National Cohort [8]. In addition the MonDAFIS Study of the CSB uses the ID-Management solution E-PIX altogether with the open source EDC-Solution REDCap [32].

6. Conclusions

Applying common phases of a cohort study, this paper has deduced core elements of CDM for cohort studies and registries in terms of functional and non-functional requirements. Associated challenges were highlighted, including requirements for IT security and legal data protection regulations.

In summary, CDM comprises the implementation of extensive technical and organisational means for the collection, processing, storage and provisioning of medical research data in accordance with the individual requirements specified by an epidemiological cohort study. Early incorporation of CDM features in study planning permits successful implementation before data collection commences. Furthermore CDM facilitates a process of systematic improvements in terms of data quality and availability while compliance with statutory legal frameworks is ensured.

This paper has introduced the MOSAIC project as a potential resource to support the implementation of CDM in epidemiological research with a first set of open source tools. Initial responses from the scientific community document how the concept of tool re-usability can be implemented while preserving associated benefits (cf. [24]). An important lesson learned is that the simplest solutions (e.g. a document template) awake great public interest.

Nonetheless, the MOSAIC tools will be unable to cover each requirement for CDM from an epidemiological perspective. Initially, no support for specimen management or for the linkage, import and processing of secondary data will be offered. In addition, no provision of servers or facilities for long-term data preservation is intended at present. MOSAIC also does not act as the commissioning instance for the services of a Trusted Third Party. On the other hand, technical support requests related to the tools as provided can be submitted via the project portal at any time.

For the remainder of its term, the MOSAIC project will focus on developing additional tools, which will continuously be made available to the research community via the MOSAIC project portal (mosaic-greifswald.de).

Acknowledgments

This research is funded by the German Research Foundation (DFG) as a part of the research grant programme "Information infrastructure for research data" (grant number HO 1937/2-1)

References

1. Higgins S. The DCC Curation Lifecycle Model. The International Journal of Digital Curation 2008; 3 (1): 134–140. doi: 10.2218/ijdc.v3i1.48.
2. Committee on Scientific Library Services:Subcommittee on Information Management. Recommendations for Secure Storage and Availability of Digital Primary Research Data (Empfehlungen zur gesicherten Aufbewahrung und Bereit-stellung digitaler Forschungsprimärdaten). [Online]. Bonn 2009 [cited 2015 02 10]. Available from: http://www.dfg.de/download/pdf/foerderung/programme/lis/ua_inf_empfehlungen_200901_en.pdf.
3. Jensen U. Guidelines for the management of research data (Leitlinien zum Management von Forschungsdaten). Technical Report. Köln: GESIS – Leibniz Institute for Social Sciences, Social Sciences; 2012.
4. Winter A, Funkat G, Haeber A, Mauz-Koerholz C, Pommerening K, Smers S, et al. Integrated Information Systems for Translational Medicine. Methods Inf Med 2007; 46 (5): 601–607. doi: 10.1160/ME9063.
5. Sariyar M, Borg A, Pommerening K. Evaluation of Record Linkage Methods for Iterative Insertions. Methods Inf Med 2009; 48 (5): 429–437. doi: 10.3414/ME9238.
6. Völzke V, Alte D, Schmidt CO, Radke D, Lorbeer R. Cohort Profile: The Study of Health in Pomerania. International Journal of Epidemiology 2011; 40 (2): 294–307. doi: 10.1093/ije/dyp394.
7. Grabe H, Assel H, Bahls T, Dörr M, Endlich K, Endlich N, et al. Cohort profile: Greifswald approach to individualized medicine (GANI_MED). Journal of Translational Medicine 2014; 12 (144). doi: 10.1186/1479-5876-12-144.
8. The German National Cohort (Nationale Kohorte e.V.). The German National Cohort Website. [Online]. 2014 [cited 2015 02 10]. Available from: http://www.nationale-kohorte.de/content/Datenschutzkonzept_130314.pdf.
9. German Centre for Cardiovascular Research (DZHK). dzhk.de. [Online]. 2015 [cited 2015 02 10]. Available from: http://dzhk.de/.
10. Meyer J, Ostrzinski S, Fredrich D, Havemann C, Krafczyk J, Hoffmann W. Efficient data management in a large-scale epidemiology research project. Comput Methods Programs Biomed 2012; 107 (3): 425–435. doi: 10.1016/j.cmpb.2010.12.016.
11. German Federal Office for Information Security (Bundesamt für Sicherheit in der Informationstechnik). BSI-Standard 100-2 IT-Grundschutz-Vorgehensweise. Bonn: German Federal Office for Information Security (Bundesamt für Sicherheit in der Informationstechnik); 2008.
12. Pommerening K, Drepper J, Helbing K, Ganslandt T. Guideline for Data Protection in Medical Research Projects: TMF's generic solutions 2.0. 1st ed. Berlin; 2014.

13. MOSAIC. The Reference Portal plan.Tau. [Online]. 2014 [cited 2015 02 24]. Available from: https://mosaic-greifswald.de/qa/info.
14. Greenspan G. Question2Answer. [Online]. [cited 2015 2 24]. Available from: http://www.question2answer.org/.
15. Fellegi I, Sunter A. A theory for record linkage. Journal of the American Statistical Association 1969; 64 (328): 1183–1210. doi: 10.1080/01621459.1969.10501049.
16. PrimeTek. PrimeFaces. [Online]. 2014 [cited 2015 2 24]. Available from: http://primefaces.org/.
17. Bahls T, Liedtke W, Geidel L, Langanke M. Ethics Meets IT: Aspects and elements of Computer-based informed consent processing. In Fischer T, Langanke M, Marschall P, et al., editors. Individualized medicine, ethical, economical and historical perspectives. Springer; 2015. pp 209–229.
18. Geidel L, Bahls T, Hoffmann W. Ein generisches Pseudonymisierungswerkzeug als Modul des Zentralen Datenmanagements medizinischer Forschungsdaten. In: Löffler M, Riedel-Heller S, editors. Abstractband 8th Annual Conference of the German Society for Epidemiology (DGEpi) e.V. and 1st International LIFE Symposium (Abstractband 8. Jahrestagung der Deutschen Gesellschaft für Epidemiologie und 1. Internationales LIFE Symposium). Leipzig; 2013. pP 245–246.
19. OpenClinica, LLC. Open Clinica – Open Source for Clinical Research. [Online]. 2015 [cited 2015 2 24]. Available from: https://community.openclinica.com/.
20. The R Foundation. The R Project for Statistical Computing. [Online]. 2015 [cited 2015 2 24]. Available from: http://www.r-project.org/.
21. The MOSAIC Project. MOSAIC Subversion Repository. [Online]. 2015 [cited 2015 2 24]. Available from: http://www.mosaic-greifswald.de/submin.
22. Creative Commons. Creative Commons Attribution 4.0 International License. [Online]. 2013 [cited 2015 02 10]. Available from: http://creativecommons.org/licenses/by/4.0/.
23. Free Software Foundation. GNU Affero General Public License. [Online]. 2007 [cited 2015 02 10]. Available from: http://www.gnu.org/licenses/agpl-3.0.html.
24. Fraser H, Thomas D, Tomaylla J, Garcia N, Lecca L, Murray M, et al. Adaptation of a web-based, open source electronic medical record system platform to support a large study of tuberculosis epidemiology. BMC Medical Informatics and Decision Making 2012; 12 (125). doi: 10.1186/1472-6947-12-125.
25. Sariyar M, Borg A, Heidinger O, Pommerening K. A practical framework for data management processes and their evaluation in population based medical registries. Informatics for Health and Social care 2013; 38 (2): 104–119. doi: 10.3109/17538157.2012.735731.
26. Muscholl M, Lablans M, Wagner T, Ückert F. OSSE – open source registry software solution. Orphanet Journal of Rare Diseases 2014; 9 (Suppl 1). doi: 10.1186/1750-1172-9-S1-O9.
27. TMF e.V. The TMF PID-Generator. [Online]. 2014 [cited 2015 02 10]. Available from: http://www.tmf-ev.de/Themen/Projekte/V015_01_PID_Generator.aspx.
28. Muscholl M, Lablans M, Borg A, Ückert F. Integration des Identitätsmanagements für Forschungsdatenbanken in ETL-Prozesse am Beispiel der Mainzer Patientenliste. In: GMDS 2013. 58. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e.V. (GMDS); 2013; Lübeck. doi: 10.3205/13gmds052.
29. TMF e.V. The TMF pseudonymization service (PSD), a tool for the reversible pseudonymization of medical research data. [Online]. 2014 [cited 2015 02 16]. Available from: http://www.tmf-ev.de/Projekte/TMFProjekte/V000_01_PSD.aspx.
30. TMF e.V. Informed Consent – Software wizard of the TMF provides practical support. [Online]. 2007 [cited 2015 01 29]. Available from: http://www.tmf-ev.de/News/articleType/ArticleView/articleId/223.aspx.
31. Heinze O, Birkle M, Köster L, Bergh B. Architecture of a consent management suite and integration into IHE-based regional health information networks. BMC Medical Informatics and Decision Making 2011; 11 (58). doi: 10.1186/1472-6947-11-58.
32. Harris P, Taylor R, Thielke R, Payne J, Gonzalez N, Conde J. Research electronic data capture (REDCap) – A metadata-driven methodology and workflow process for providing translational research informatics support. Journal of Biomedical Informatics 2009; 42 (2): 377–381. doi: 10.1016/j.jbi.2008.08.010.

