Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230292

Conducting an Epidemiologic Study and Making It FAIR: Reusable Tools and Procedures from a Population-Based Cohort Study

Carsten Oliver SCHMIDT[a,1], Stephan STRUCKMANN[a], Maik SCHOLZ[a], Janka SCHÖSSOW[a], Dörte RADKE[a], Adrian RICHTER[a], Achim REINEKE[b], Elisa KASBOHM[a], Joany Mariño CORONADO[a], Birgit SCHAUER[a], Kristin HENSELIN[a], Susanne WESTPHAL[a], Darko BALKE[a], Torsten LEDDIG[a], Henry VÖLZKE[a] and Jörg HENKE[a]

[a] *University Medicine Greifswald, Institute for Community Medicine. Greifswald, Germany*

[b] *Leibniz Institute for Prevention Research and Epidemiology - BIPS. Bremen, Germany*

ORCiD ID: C Schmidt (0000-0001-5266-9396), M Scholz (0009-0000-5385-864X), S Struckmann (0000-0002-8565-7962), J Schössow (0009-0003-4598-3397), A Richter (0000-0002-3372-2021), D Radke (0009-0006-4926-8018), A Reineke (0000-0002-0092-4110), E Kasbohm (0000-0001-5261-538X), J Mariño (0000-0002-4657-3758), B Schauer (0000-0001-9847-130X), K Henselin (0009-0004-6290-8468), S Westphal (0009-0008-7580-8465), D Balke (0009-0003-1550-0470), T Leddig (0000-0001-8883-5480), H Völzke (0000-0001-7003-399X), J Henke (0009-0000-3583-4310)

Abstract. Conducting large-scale epidemiologic studies requires powerful software for electronic data capture, data management, data quality assessments, and participant management. There is also an increasing need to make studies and the data collected findable, accessible, interoperable, and reusable (FAIR). However, reusable software tools from major studies, underlying such needs, are not necessarily known to other researchers. Therefore, this work gives an overview on the main tools used to conduct the internationally highly networked population-based project Study of Health in Pomerania (SHIP), as well as approaches taken to improve its FAIRness. Deep phenotyping, formalizing processes from data capture to data transfer, with a strong emphasis on cooperation and data exchange have laid the foundation for a broad scientific impact with more than 1500 published papers to date.

Keywords. FAIR, cohort study, web applications, quality management, data quality

[1] Corresponding Author: Carsten Oliver Schmidt, University Medicine Greifswald, Institute for Community Medicine, SHIP-Clinical Epidemiological Research – Functional Division Quality in the Health Sciences, Walther Rathenau Str. 48, 17475 Greifswald, Germany; Email: Carsten.schmidt@uni-greifswald.de.

1. Introduction

Successfully running an epidemiologic study requires a wide range of IT-tools to cover aspects such as electronic data capture, data management, quality management, and participant management. There is a growing need to make studies FAIR to facilitate networked research and increase reproducibility [1]. Meeting all demands is a complex challenge in large-scale epidemiologic cohort studies [2]. Potentially powerful solutions are in place to meet demands from such studies. These solutions often remain unknown to other scientists and groups, despite their potential for reuse. This work therefore provides an overview of software tools and processes that have been put in place to conduct an internationally highly networked cohort-study and make it FAIR.

2. Methods

2.1. The Study of Health in Pomerania (SHIP)

The SHIP project studies the prevalence and incidence of risk factors, subclinical disorders, clinical diseases, and their inter-relations [3]. It consists of three cohorts (SHIP-START: N=4308; SHIP-TREND: N=4420; SHIP-NEXT: baseline ongoing, projected N=4000), the first of which started in 1997. Data collections are ongoing in all three cohorts. Major follow-up examinations are carried out in approximately five-year intervals. SHIP has implemented one of the widest scope of examinations worldwide in population-based research, including amongst others interviews, questionnaires, biomaterials (blood, urine, faeces, saliva), imaging (e.g., ultrasound; full-body magnetic resonance imaging (MRI)), cardiovascular, dental, and dermatological examination, polysomnography, body scanning, as well as examinations of animals. Not counting OMICS data and externally linked data sources, the SHIP database includes far more than 50,000 data elements. Attached to the main examination waves are hundreds of side projects (e.g., MRI readings) with independent but centrally managed data collections.

2.2. IT-background

SHIP web-applications have been developed using, amongst others, the technology of Java Server Faces. They are partly in a refactoring process to migrate to the Spring Boot architecture. All of them are provided as Tomcat (https://tomcat.apache.org) applications. GitLab is used as a development infrastructure that includes a central source code repository for distributed version control, a wiki for documentation, and a ticket system. From the beginning, internationalization was implemented to enable the handling of different languages. As part of continuous integration, some SHIP applications are build automatically for SHIP's own repositories. Auto deployment can be done easily this way. Keycloak is currently being established to allow for a single sign-in. All SHIP data are stored in a central PostgreSQL database.

3. Results

3.1. Tools for conducting the study and quality management

Table 1 provides a brief description of the major tools used to run SHIP data collections, accompanying processes such as data and quality management, as well as participant management. Figure 1 sketches their interrelation. Many tools have already been reused in other major studies. Examples are SHIPPIE and SHIPdesigner in the polish Bialystok PLUS study, WebMODYS and Square² in the German National Cohort (NAKO Gesundheitsstudie), the largest German epidemiologic cohort study to date, dataquieR in Eurocrine. The following paragraphs describe approaches to improve the FAIRness of SHIP.

3.2. Improving Findability

To make content better findable rather than relying on study specific web pages, several collaborations have been established, e.g. with Maelstrom Research [4] (https://www.maelstrom-research.org/), the Portal of Medical Data Models [5] (https://medical-data-models.org/), NFDI4Health (https://www.nfdi4 health.de/), and euCanSHare (https://mica.eucanshare.bsc.es/) to facilitate the search for SHIP data. A key feature of these collaborations has been the semantic annotation of data elements by cooperation partners to describe the content of study variables, using either the Maelstrom taxonomy (Maelstrom, euCanSHare, NFDI4Health) or Unified Medical Language System (UMLS, MDM). This greatly enhances comparison options with other studies, such as UK-Biobank or Rotterdam Study, amongst others.

3.3. Improving Accessibility

For legal and data protection issues, access to SHIP data cannot directly be made available publicly or from a central portal. However, a dedicated public web-site (https://www.fvcm.med.uni-greifswald.de/) and formal Use & Access process enables sustainable access to SHIP data [3]. More than 200 applications have been processed annually in the previous years. The transfer-site is open internationally for scientists to register for data requests.

3.4. Improving Interoperability

The SHIP PostgreSQL database has a proprietary data model but the need for exchange with other Central Metadata Repositories has led to the creation of export routines to formats such as OPAL, CDISC, as well as common formats used in statistical software such as R, SAS, Stata, and SPSS. Availability of SHIP forms via MDM [5] also enables their direct export and use in applications like REDCap.

3.5. Improving Reusability

Since study data and metadata are stored in a central database, they can be easily managed in terms of data protection, availability and long-term archiving. The processes are supported by a fully automated pipeline that covers all steps of data management

from data capture, through data curation to data extraction for scientists with accepted data applications [6]. Reusability also requires high quality data and metadata. Therefore, extensive data quality assessment pipelines have been implemented to find and fix issues.

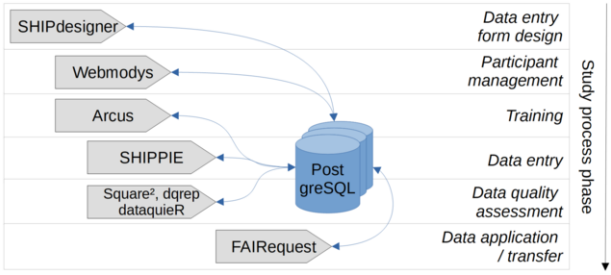

Figure 1. Position of SHIP tools in different phases of the study life cycle

Table 1. Overview of SHIP tools for conducting studies and quality management

SHIPPIE is used for electronic data capture, using Java Server Faces, Jooq, and a PostgreSQL database. SHIPPIE generates web forms in a fully dynamic manner based on specifications defined using SHIPdesigner. SHIPPIE has been designed with demands to handle multiple examinations and visits per person. A more extensive set of metadata information than usual is in place to enable automated data quality assessments. Not only simple limit violations, but also conditional violations can be assessed by entering Boolean expressions in the disjunctive normal form. Measurements from previous observations can be displayed as needed to prevent input errors. The web application offers two modes: the recording mode and the test mode. In the latter, recorded data are stored in a test database table. This way, no second test installation is needed for testing purposes. A powerful rights-and-role system allows to specify users' access rights according to their role in the study.
SHIPdesigner supports SHIPPIE users in creating and editing web forms. It uses primefaces (https://www.primefaces.org/) to render the GUI rather than standard jsf (plain html) components. Therefore, neither new database tables nor updates of the SHIPPIE application are needed, if forms are added. This increases the implementation speed of changes in the web forms and decouples the operational use of the applications from the development process.
WebMODYS is a web-application to support, control, and document participant management in population-based studies. It has been developed as a Java web application in a cooperation with BIPS, Bremen, Germany. A prominent feature lies in its free configurability for study-specific recruitment processes, which is achieved by storing the recruitment processes as state diagrams in the database. WebMODYS provides functions for daily participant management tasks and their documentation like editing and updating basic participant data, generating letters, scheduling appointments, and logging all contacts between participants and recruitment personnel as well as other participant-related recruitment tasks. It enables a full protocol of any participant contact and provides data to calculate response proportions or to analyze non-response.
ARCUS is a web application to train and certify readers of imaging data and is used for example in the context of ultrasound examinations. It was programmed using Java Server Faces, and PostgreSQL, and can read DICOM images in 16-bit color depth and allows measurements directly on those images. Arcus issues training-certificates for readers who successfully completed their training.
dataquieR, **dqrep** and **Square²** were designed to enable an automated data quality monitoring. dataquieR is available as an R package [7], which can be downloaded from CRAN. dqrep is a Stata package with functionalities for multi-report generation. Square² is a JAVA web application [8]. Square² offers a graphical user interface (GUI) to target all steps in the data quality assessment workflow: implementing the study structure, managing needed metadata and finally creating quality reports.
FAIRequest refers to a web application in Spring Boot that has replaced an older php data-application tool to search and apply for SHIP data using standardized data application access forms. Variable browsing is possible through direct links to the SHIP data dictionary.

There are further tools in place such as a daily run fully automated modularized SAS data-cleaning pipeline. It's output is used by an Access application to collect feedback on detected issues [6]. The **JoinUs4Health** platform (https://platform.joinus4health.eu) is a Wordpress application that allows anybody from the age of 16 years to submit own suggestions for SHIP topics, or vote on contributions of others.

4. Discussion

This paper provides an overview of key tools to conduct SHIP and improve its FAIRness. Their use has helped make SHIP one of the most interconnected epidemiologic projects in the world, with hundreds of cooperation partners worldwide and more than 1500 peer-reviewed articles. Yet, there are also structural limitations to FAIRness, for example related to the public storage of highly sensitive personal health data in compliance with the EU-General Data Protection Regulation.
The tools mentioned here are publicly available and can be downloaded directly or used free of charge for academic purposes through collaborations. Some tools, such as SHIPPIE, have been tailored to the specific needs of complex cohort studies. Others are more general in nature, like the data quality and participant management software. Future challenges will be their gradual update, and a stronger integration with standards and developments such as FHIR or OMOP. We also aim to increase FAIRness of SHIP towards non-scientists such as participants, the general population, and stakeholders to promote citizen science through means such as the JoinUs4Health platform.

Declarations

Ethical vote: not applicable

Conflict of Interest: The authors declare that there is no conflict of interest.

Acknowledgement: Content of this work was funded by the Ministry for Education, Science and Culture of the State of Mecklenburg-Vorpommern, European Union's funds (No UG 11 035A , 825903, No. 101006518), and by the German Research Foundation (DFG, SCHM 2744/3-1; SCHM 2744/3-4; NFDI 13/1).

References

[1] Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data 2016;3:160018.
[2] Wichmann HE, Kuhn KA, Waldenberger M, et al. Comprehensive catalog of European biobanks. Nat Biotechnol 2011;29:795-7.
[3] Volzke H, Schossow J, Schmidt CO, et al. Cohort Profile Update: The Study of Health in Pomerania (SHIP). Int J Epidemiol 2022.
[4] Bergeron J, Doiron D, Marcon Y, Ferretti V, Fortier I. Fostering population-based cohort data discovery:. PLoS ONE 2018;13:e0200926.
[5] Hegselmann S, Gessner S, Neuhaus P, Henke J, Schmidt CO, Dugas M. Automatic conversion of metadata from the Study of Health in Pomerania to ODM. Stud Health Technol Inform 2017;236:88-96.
[6] Werner A, Maiwald S, Henselin K, et al. [Modular automated data cleaning in a large population-based cohort]. In: Chenot JF, Minkenberg R, eds. Proceedings of the 20 conference of SAS®-Users in research and development: Shaker; 2016:279-84.
[7] Richter A, Schmidt CO, Krüger M, Struckmann S. dataquieR: assessment of data quality in epidemiological research. JOSS 2021;6:3039.
[8] Schmidt CO, Krabbe C, Schössow J, Albers M, Radke D, Henke J. Square² - A web application for data monitoring in epidemiological and clinical studies. Stud Health Technol Inform 2017;235:549-53.