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

doi:10.3233/SHTI230071

The Importance of Being FAIR and FAST – The Clinical Epidemiology and Study Platform of the German Network University Medicine (NUKLEUS)

Dagmar KREFTING[a,b,1], Gabi ANTON[c], Irina CHAPLINSKAYA-SOBOL[a], Sabine HANSS[a], Wolfgang HOFFMANN[d], Sina M. HOPFF[e,f], Monika KRAUS[c], Roberto LORBEER[g], Bettina LORENZ-DEPIEREUX[c], Thomas ILLIG[h], Christian SCHÄFER[i], Jens SCHALLER[g], Dana STAHL[j], Heike VALENTIN[j], Peter HEUSCHMANN[k] and Janne VEHRESCHILD[e,l]

[a] *Dpt. of Medical Informatics, University Medical Center Göttingen, German Center for Cardiovascular Research (DZHK) partner site Göttingen, Germany*
[b] *Campus Institute Data Science (CIDAS), Georg-August-University Göttingen,*
[c] *Institute of Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany*
[d] *Institute for Community Medicine, University Medicine Greifswald, Germany*
[e] *Faculty of Medicine, University of Cologne, Department I of Internal Medicine, University Hospital Cologne, Germany*
[f] *Center for Integrated Oncology Aachen Bonn Cologne Duesseldorf, Cologne, Germany*
[g] *Medical Heart Center and Institute of Computer-assisted Cardiovascular Medicine, Charité – Universitätsmedizin Berlin, Germany*
[h] *Hannover Unified Biobank, Hannover Medical School, Hannover, Germany*
[i] *Institute of Clinical Chemistry and Laboratory Medicine,University Medicine Greifswald, Germany*
[j] *Independent Trusted Third Party of the University Medicine Greifswald, Germany*
[k] *Institute of Clinical Epidemiology and Biometry, University of Würzburg; Clinical Trial Center, University Hospital Würzburg, Germany*
[l] *German Centre for Infection Research (DZIF), partner site Bonn- Cologne, Cologne, Department II for Internal Medicine, Hematology/Oncology, University Hospital Frankfurt, Frankfurt am Main, Germany.*

ORCID IDs: Dagmar Krefting 0000-0002-7238-5339, Gabi Anton 0000-0003-0483-892X, Irina Chaplinskaya-Sobol 0000-0003-4811-1028, Sabine Hanß 0000-0002-9133-8685, Wolfgang Hoffmann 0000-0002-6359-8797, Sina Hopff 0000-0002-8886-4596, Monika Kraus 0000-0003-4825-8724, Roberto Lorbeer 0000-0002-2224-9208, Bettina Lorenz-Depiereux 0000-0001-7215-9174, Thomas Illig 0000-0003-4284-5389, Christian Schäfer 0000-0001-8873-7231, Jens Schaller 0000-0002-9399-1653, Dana Stahl 0000-0002-4283-4543, Heike Valentin 0000-0001-6379-6534, Peter Heuschmann 0000-0002-2681-3515, Janne Vehreschild 0000-0002-5446-7170

[1] Corresponding Author: Dagmar Krefting, University Medical Center Göttingen, Germany, E-mail: dagmar.krefting@med.uni-goettingen.de

Abstract. The COVID-19 pandemic has urged the need to set up, conduct and analyze high-quality epidemiological studies within a very short time-scale to provide timely evidence on influential factors on the pandemic, e.g. COVID-19 severity and disease course. The comprehensive research infrastructure developed to run the German National Pandemic Cohort Network within the Network University Medicine is now maintained within a generic clinical epidemiology and study platform NUKLEUS. It is operated and subsequently extended to allow efficient joint planning, execution and evaluation of clinical and clinical-epidemiological studies. We aim to provide high-quality biomedical data and biospecimens and make its results widely available to the scientific community by implementing findability, accessibility, interoperability and reusability – i.e. following the FAIR guiding principles. Thus, NUKLEUS might serve as role model for FAIR and fast implementation of clinical epidemiological studies within the setting of University Medical Centers and beyond.

Keywords. clinical research infrastructure, FAIR principles, COVID-19, cohort study, Data reuse, biosamples, Images

1. Introduction

Timely availability of research results is required in the context of a pandemic, as evidence about influential factors on the severity and course of infections may have a major impact on planning efficient and effective initial treatment options and improving long-term outcomes of affected patients. But such evidence requires large-scale studies - comprehensive characterization of the patients by clinical data items, biospecimens, standardized imaging, and patients being recruited in different research settings. Typically, such large-scale studies are planned and conducted on a mid- to long-term scale, as they need significant financial resources but in particular personnel to initiate, run and analyze the study. In Germany, the *Network University Medicine* (NUM) has been founded in April 2020 to understand and master the COVID-19 pandemic by joint forces of all university medical centers. Among the 13 projects initiated in the first funding period of NUM, the National Pandemic Cohort Network (NAPKON) as the largest initiative initiated three cohort studies [1]. They differ in study population, study protocol and phenotyping details. The cohorts where complemented by overarching methodological core units. In particular during the setup phase of such multi-center studies, a fast and timely recruitment start on one hand and the fulfillment of organizational, regulatory and technical requirements for a findable, accessible, interoperable and reusable (FAIR) data collection on the other hand have been experienced as conflicting requirements. The aim of this paper is to describe the measures taken and lessons learned to reconcile these conflicts during the first funding phase and how these aspects are addressed in the NUM clinical epidemiology and study platform (NUKLEUS), established within the second funding period as a permanent collaborative research infrastructure.

2. Methods

For the successful start of the NAPKON studies, several challenges had to be addressed in a timely manner, in particular: (a) Building and supporting a large network of participating sites covering different source populations and different levels of expertise, (b) establishing a legal framework including study and regulatory documents, and

contracts between all participating infrastructure partners, and (c) coordinating implementation of the study data into the data management systems, including consent data, clinical and imaging data, and biospecimen.

We had an explicit order to make the three cohorts interoperable with each other as well as with other national and international initiatives such as the national project to share routine clinical data of COVID-19 patients [2]; and to allow for wide reuse of the collected data and biospecimen. Therefore we made specific efforts to harmonize (i) study documents - in particular patient information and informed consents, (ii) biosample collections and standard operation procedures (SOPs), (iii) common data items including quality checks and the integration of the German Corona Consensus dataset [3]; and last but not least (iv) fine-grained case fees.

The NAPKON project set up four core units for a harmonized organizational and governmental implementation of the three cohorts. The *biosample core unit* defined a core set of biospecimen to be collected by all cohorts including SOPs. The *epidemiology core unit* provides methodological consultancy of clinical epidemiological studies, delivers regular reports on data quality, and supports use and access requests and statistical analyses. The *integration core unit* evaluates external and existing cohorts upon integration into NAPKON. Last but not least, the *interaction core unit* coordinates the overall project and is in particular engaged in enabling and enhancing communication within the cohorts, the overall recruitment network, and researchers interested in data usage – reaching out to the full scientific NUM community [1].

The German Center for Cardiovascular Research (DZHK) has made available its clinical study infrastructure to NAPKON. The DZHK infrastructure encompasses integrated data management systems for informed consents, electronic case report forms, imaging data and biospecimens, respectively, with record-linkage through a trusted third party and data exports for approved requests through a transfer office. An ethics coordination supports the study coordinators in creating the study documentation and patient information suited for the intended data handling and reuse.

In 2022, we integrated the clinical study infrastructure with the *biosample*, *epidemiology* and *interaction core units* to the joint infrastructure called NUKLEUS. Within NUKLEUS, we develop and operate the research infrastructure to support joint planning, execution and evaluation of clinical and clinical epidemiological studies. The vision of NUKLEUS is to make high-quality data, biospecimens, and analysis results widely available to the scientific community. While the continuous support of the NAPKON cohorts is the main short-term objective, we consider the platform to support different multicenter clinical and clinical epidemiological studies conducted within the NUM and beyond.

3. Results

3.1. Technical infrastructure components

The technical components of the research data infrastructure encompass the following systems operated at different sites: *secutrial*® for clinical data, *CentraXX*® for biospecimen management, and *TrialComplete*® for images and biosignals. Identifying patient data and informed consents are handled by an independent trusted third party employing the open source tools *gICS*® for consent management including management

of partial or complete withdrawal, *E-PIX*® for identity management, and *gPAS*® as pseudonymization service.

The transfer office provides services to request data from the different data management systems and compiles the digital data collection. If biospecimens are part of the request, it hands over the biospecimens list to the *biosample core unit*. Biosamples are stored locally at the study sites while the metadata is centrally documented - enabling fast and coordinated access for approved usage projects. The transfer office is also responsible to compile metadata and data collections for quality assurance, reporting and accounting. Originally adopted from the DZHK, all operating sites run dedicated instances for NUKLEUS. Specific technical extensions are an export interface in *secutrial* for the harmonized data items, and a digital consent solution with *gICS*. We employ the *ProSkive* application portal for the management of data requests, with internal communication on data request handling by *GitLab* issue tracking. Further internal and external communication uses the *NAPKON-Suite*, a growing collection of collaboration and project management solutions.

3.2. Organizational and methodological infrastructure components

NUKLEUS defines and implements processes and provides methods to ensure high quality study designs, data collection and data analysis as well as efficient information and data flow within the project and with the different stakeholders. We aim at supporting both study coordinators and data users through the full research process from project definition to data reuse. For new projects, we defined an onboarding SOP and reimbursement schemes to allow a consistent calculation of the project costs. We offer methodological support to applicants for study design, biosample-related topics as well as for the statistical analysis. Informed consents are semantically annotated with a NUM-wide coding scheme to ease cross-platform data reuse. We also coordinate the comprehensive community of potentially participating sites, the so-called domain and organ specific working groups that form an independent organization body considered the specialist's backbone of NUM.

During the active recruitment phase of a study, we provide automated reporting on recruitment and data quality, as part of an accounting pipeline. Biobanking SOPs and audits at the study sites support high quality biospecimen collection.

To ease the data request process, we offer researchers who are interested in data reuse help to understand the available data and to define appropriate analysis methods. We established data request pipelines encompassing coordination of the Use & Access process, feasibility analysis, check for current consent status, data export, and finally data transfer.

3.3. Experiences from the so-far implemented studies

The three NAPKON cohorts were encouraged to start operations based on favorable reviews in August 2020 and received formal approval for funding in February 2021. The study protocols have partly been adopted from already running studies, in particular COVIDOM, LEOSS and Pa-COVID [4–6], but additional harmonization efforts and adoption of the common infrastructure were needed for all studies. First preparation steps were already ongoing since submission of the concept and funding application on July 17, 2020, and the first patient consented to the cross-sectoral cohort in Frankfurt on Nov 4, 2020, on the same day as receiving a final positive ethics vote for the study and one

day after setting the study documentation to production status. The two other cohorts recruited the first patients in the same month. All cohorts are still ongoing, with applications for data use being accepted from April 2021. To date, data from 6651 patients are available for reuse with about 32.000 reviewed visits. 83.300 primary biosamples from 5430 patients are available. Until now 104 usage applications has been approved, of which so far 64 received data. Until end of 2022, 12 sample usage applications received a total of 36.600 samples.

In April 2022, the project NU(M)KRAINE – a screening program on infectious diseases for refugees of the Ukraine was proposed. It received approval in June and started recruiting in October with almost 1800 study participants until December.

4. Discussion

The NUKLEUS study platform represents a promising infrastructure for the collection, analysis and provision of data and biosamples in the context of the ongoing COVID-19 pandemic and beyond. However, the timely implementation posed several challenges with regards to communication, workflow management, legal, ethical, data protection, and administrative challenges – in particular at study preparation. As main bottlenecks in the timely data provision for reuse we identified unclear specifications by the applicants and data not yet available. Efficient and streamlined workflow management is vital for the smooth operation of the platform, especially in a context where the integration of large amounts of data from various sources, platforms and stakeholders is a key aspect. Clear and consistent communication channels, both internally and with all stakeholders from study centers to interested researchers are essential to ensure that all participants are aware of the project's goals, objectives and their roles, as well as to facilitate the flow of information and reporting of issues in a timely manner. We are currently in discussion with further projects and will continuously explore how NUKLEUS can help collaborative research projects to be both, FAIR and FAST.

Acknowledgement: The work presented is funded by the German Ministry of Research and Education (NUM – 01KX2121).

References

[1] Schons M, Pilgram L, Reese J-P, et al. The German National Pandemic Cohort Network (NAPKON): rationale, study design and baseline characteristics. *Eur J Epidemiol*. Epub ahead of print 29 July 2022. DOI: 10.1007/s10654-022-00896-z.

[2] Prokosch H-U, Bahls T, Bialke M, et al. The COVID-19 Data Exchange Platform of the German University Medicine. *Stud Health Technol Inform* 2022; 294: 674–678. DOI: 10.3233/SHTI220554

[3] Sass J, Bartschke A, Lehne M, et al. The German Corona Consensus Dataset (GECCO): a standardized dataset for COVID-19 research in university medicine and beyond. *BMC Med Inform Decis Mak* 2020; 20: 341. DOI: 10.1186/s12911-020-01374-w

[4] Horn A, Krist L, Lieb W, et al. Long-term health sequelae and quality of life at least 6 months after infection with SARS-CoV-2: design and rationale of the COVIDOM-study as part of the NAPKON population-based cohort platform (POP). *Infection* 2021; 49: 1277–1287.

[5] Jakob CEM, Borgmann S, Duygu F, et al. First results of the 'Lean European Open Survey on SARS-CoV-2-Infected Patients (LEOSS)'. *Infection* 2021; 49: 63–73. DOI: 10.1007/s15010-020-01499-0

[6] Kurth F, Roennefarth M, Thibeault C, et al. Studying the pathophysiology of coronavirus disease 2019: a protocol for the Berlin prospective COVID-19 patient cohort (Pa-COVID-19). *Infection* 2020; 48: 619–626. DOI: 10.1007/s15010-020-01464-x