

The German Corona Consensus Dataset (GECCO): a standardized dataset for COVID-19 research in university medicine and beyond

Thun, Sylvia; Sass, Julian; Bartschke, Alexander; Lehne, Moritz;
Essenwanger, Andrea; Rinaldi, Eugenia; Rudolph, Stefanie; Heitmann, Kai
U.; Vehreschild, Jörg J.; Kalle, Christof von

Veröffentlichungsversion / Published Version

Zeitschriftenartikel / journal article

Empfohlene Zitierung / Suggested Citation:

Thun, S., Sass, J., Bartschke, A., Lehne, M., Essenwanger, A., Rinaldi, E., ... Kalle, C. v. (2020). The German Corona Consensus Dataset (GECCO): a standardized dataset for COVID-19 research in university medicine and beyond. *BMC Medical Informatics and Decision Making*, 20(341), 1-7. <https://doi.org/10.1186/s12911-020-01374-w>

Nutzungsbedingungen:

Dieser Text wird unter einer CC BY Lizenz (Namensnennung) zur Verfügung gestellt. Nähere Auskünfte zu den CC-Lizenzen finden Sie hier:
<https://creativecommons.org/licenses/by/4.0/deed.de>

Terms of use:

This document is made available under a CC BY Licence (Attribution). For more information see:
<https://creativecommons.org/licenses/by/4.0>

TECHNICAL ADVANCE

Open Access



The German Corona Consensus Dataset (GECCO): a standardized dataset for COVID-19 research in university medicine and beyond

Julian Sass¹ , Alexander Bartschke², Moritz Lehne¹ , Andrea Essenwanger¹ , Eugenia Rinaldi², Stefanie Rudolph² , Kai U. Heitmann³, Jörg J. Vehreschild^{4,5,6} , Christof von Kalle^{1,2} and Sylvia Thun^{1,2,7*}

Abstract

Background: The current COVID-19 pandemic has led to a surge of research activity. While this research provides important insights, the multitude of studies results in an increasing fragmentation of information. To ensure comparability across projects and institutions, standard datasets are needed. Here, we introduce the “German Corona Consensus Dataset” (GECCO), a uniform dataset that uses international terminologies and health IT standards to improve interoperability of COVID-19 data, in particular for university medicine.

Methods: Based on previous work (e.g., the ISARIC-WHO COVID-19 case report form) and in coordination with experts from university hospitals, professional associations and research initiatives, data elements relevant for COVID-19 research were collected, prioritized and consolidated into a compact core dataset. The dataset was mapped to international terminologies, and the Fast Healthcare Interoperability Resources (FHIR) standard was used to define interoperable, machine-readable data formats.

Results: A core dataset consisting of 81 data elements with 281 response options was defined, including information about, for example, demography, medical history, symptoms, therapy, medications or laboratory values of COVID-19 patients. Data elements and response options were mapped to SNOMED CT, LOINC, UCUM, ICD-10-GM and ATC, and FHIR profiles for interoperable data exchange were defined.

Conclusion: GECCO provides a compact, interoperable dataset that can help to make COVID-19 research data more comparable across studies and institutions. The dataset will be further refined in the future by adding domain-specific extension modules for more specialized use cases.

Keywords: COVID-19, Interoperability, Standard dataset, FHIR

Background

In December 2019, first reports of a cluster of 41 patients infected by a novel coronavirus emerged from Wuhan, China [1]. Within a few months, the new virus, subsequently named “severe acute respiratory syndrome coronavirus 2” (SARS-CoV-2), has spread around the world

causing the global COVID-19 pandemic. Currently (as of November 12, 2020), SARS-CoV-2 has infected more than 50 million and killed more than a million patients worldwide [2].

The pandemic has spurred intensive scientific research, including numerous regional, national and international epidemiological surveys and studies [3–7]. While this research provides important new insights, the multitude of studies threatens to generate a dangerous fragmentation of information. This could delay or even prevent

*Correspondence: sylvia.thun@charite.de

¹ Berlin Institute of Health (BIH), Berlin, Germany

Full list of author information is available at the end of the article



© The Author(s) 2020. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

urgently needed scientific knowledge about SARS-CoV-2 and COVID-19. To avoid this fragmentation of information and make COVID-19 data more comparable and exchangeable across studies and institutions, interoperable datasets are needed.

Various initiatives have started to define uniform datasets and Common Data Elements (CDEs) for the collection of information about COVID-19. For example, questionnaires and case report forms (CRFs) have been developed to collect data about COVID-19 patients in a standardized way [5, 8, 9]. While the CDEs defined in these projects are an important step, they are not enough to ensure interoperability. To make data syntactically and semantically interoperable, data elements have to be embedded in standard data structures that can be exchanged across IT systems; they also have to use common terminologies that unambiguously define the meaning of clinical concepts.

To improve interoperability of COVID-19 data, we developed the German Corona Consensus Dataset (GECCO), which uses international health IT standards and terminologies for interoperable data exchange. GECCO defines a compact set of data elements to be collected in COVID-19 studies and was developed within the German COVID-19 Research Network of University Medicine (“Netzwerk Universitätsmedizin”) funded by the German Federal Ministry of Education and Research (BMBF) [10]. This article provides an overview of the GECCO dataset and its development.

Methods

Selection of data elements

An initial dataset was compiled as a working basis by merging data elements and response options of the following projects: the ISARIC-WHO CRF [8]; the Pa-COVID-19 study [11], which investigates the pathophysiology of COVID-19 in a prospective patient cohort; the LEOSS case registry [3], a clinical patient registry for patients infected with SARS-CoV-2 initiated by the

ESCMID Emerging Infections Task Force (EITaF), the German Center for Infection Research (DZIF) and the German Society for Infectiology (DGI). This draft dataset was saved in a spreadsheet and sent to members of an expert board for comment and proposal of additional data elements. The expert board was composed of health professionals from 50 institutions, in particular departments from German university hospitals, professional associations and other relevant organizations (such as the Medical Informatics Initiative [12] or the National Association of Statutory Health Insurance Physicians [13]). New data elements proposed by the expert board were added to the dataset for subsequent prioritization. For the prioritization, the experts were asked to assign a priority value to each data element of the dataset. Priorities were indicated on a 5-level scale that was loosely based on the NIH model for CDEs [14] (Table 1).

From the data elements with the highest prioritizations, a preliminary core dataset with roughly 100 data elements was compiled (this size was chosen to include as many relevant data elements as possible, while keeping the dataset manageable and practical). This core dataset was then reviewed by an editorial team of seven experts from different disciplines. In consensual decisions, data elements not considered necessary for the core dataset were discarded (note that these data elements were retained for additional extension modules, see Results); conversely, data elements that were considered highly important but had not yet been included in the core dataset were added. The final data elements of the core dataset were grouped into meaningful categories (e.g., demographics, symptoms or medication). Figure 1 shows the workflow of consensus building and dataset definition.

Standardization

To ensure syntactic and semantic interoperability, elements and response options of the core dataset were mapped to international standards and terminologies.

Table 1 Prioritization of data elements

Scale value	Priority	NIH classification	Definition
5	Highly relevant	General core/disease core*	Data element with essential general or specific information relevant to COVID-19
4	Very relevant	Supplemental—highly recommended	Data element that is essential under certain conditions or for certain study types and is therefore strongly recommended
3	Relevant	Supplemental	Data element that is often collected in clinical studies, but whose relevance depends on the study design or type of research
2	Less relevant	Exploratory	Data element that requires further validation, but which can fill current gaps in the data elements and/or replace an existing data element
1	Not relevant	–	Data elements that are not considered relevant to the dataset

* Since this is a disease-specific (i.e. COVID-19) dataset, both the general and disease-specific core categories of the NIH were assigned to the highest priority level



Fig. 1 Workflow of consensus building and definition of data elements for the GECCO dataset

The following terminologies and code systems were used: the International Statistical Classification of Diseases and Related Health Problems, 10th revision, German modification (ICD-10-GM) [15] for diagnoses; Logical Observation Identifiers Names and Codes (LOINC) [16] for laboratory values and other measurements; the Unified Code for Units of Measure (UCUM) [17] for measurement units; the Anatomical Therapeutic Chemical Classification System (ATC) [18] for active ingredients of drugs and medications; SNOMED CT [19] for diagnoses and other medical concepts. We used two terminology systems—SNOMED CT and ICD-10-GM—for diagnoses because ICD-10-GM is the dominant classification system in German healthcare and is important for reimbursement purposes, whereas SNOMED CT allows for a more detailed coding of clinical terms and is therefore preferable for better medical accuracy. The annotation of data elements with international terminologies was done using ART-DECOR [20], an open source collaboration platform for experts from medical, terminological and technical domains aiming on creation and maintenance of datasets with data element descriptions, use case scenarios, value sets and Health Level 7 (HL7) templates and profiles.

To define interoperable formats for data exchange, the HL7 standard "Fast Healthcare Interoperability Resources" (FHIR) [21] was used. FHIR builds on a set of "resources", which provide generic data structures for common healthcare concepts, such as Patient, Practitioner, Observation, Medication or Condition. From

these resources more specific data structure definitions, so-called "profiles," can be defined, which allow for interoperable data exchange across health IT systems. To ensure interoperability, care was taken to build on previous work where possible, in particular the FHIR profiles of the German Medical Informatics Initiative [22], the International Patient Summary (IPS) [23], the Logical COVID-19 profiles [24] and the FHIR base profiles of HL7 Germany [25]. FHIR profiles were defined using Forge [26] and published on the Simplifier platform [27].

Results

Combining the initial draft dataset and the additional proposals from the expert board, 702 potentially relevant data elements were collected. From these data elements and based on the prioritization of the expert board, the editorial team compiled a core dataset consisting of 81 elements with 281 response options. These data elements were grouped into the following categories: anamnesis/risk factors (n=16); imaging (n=2); demographics (n=7); epidemiological factors (n=1); complications (n=1); onset of illness/admission (n=1); laboratory values (n=25); medication (n=4); outcome at discharge (n=3); study enrollment/inclusion criteria (n=2); symptoms (n=2); therapy (n=6); vital signs (n=11) (Fig. 2).

For all data elements and their corresponding response options, value sets were created using codes from SNOMED CT, LOINC, UCUM, ICD-10-GM and ATC. Data elements, response options and associated value

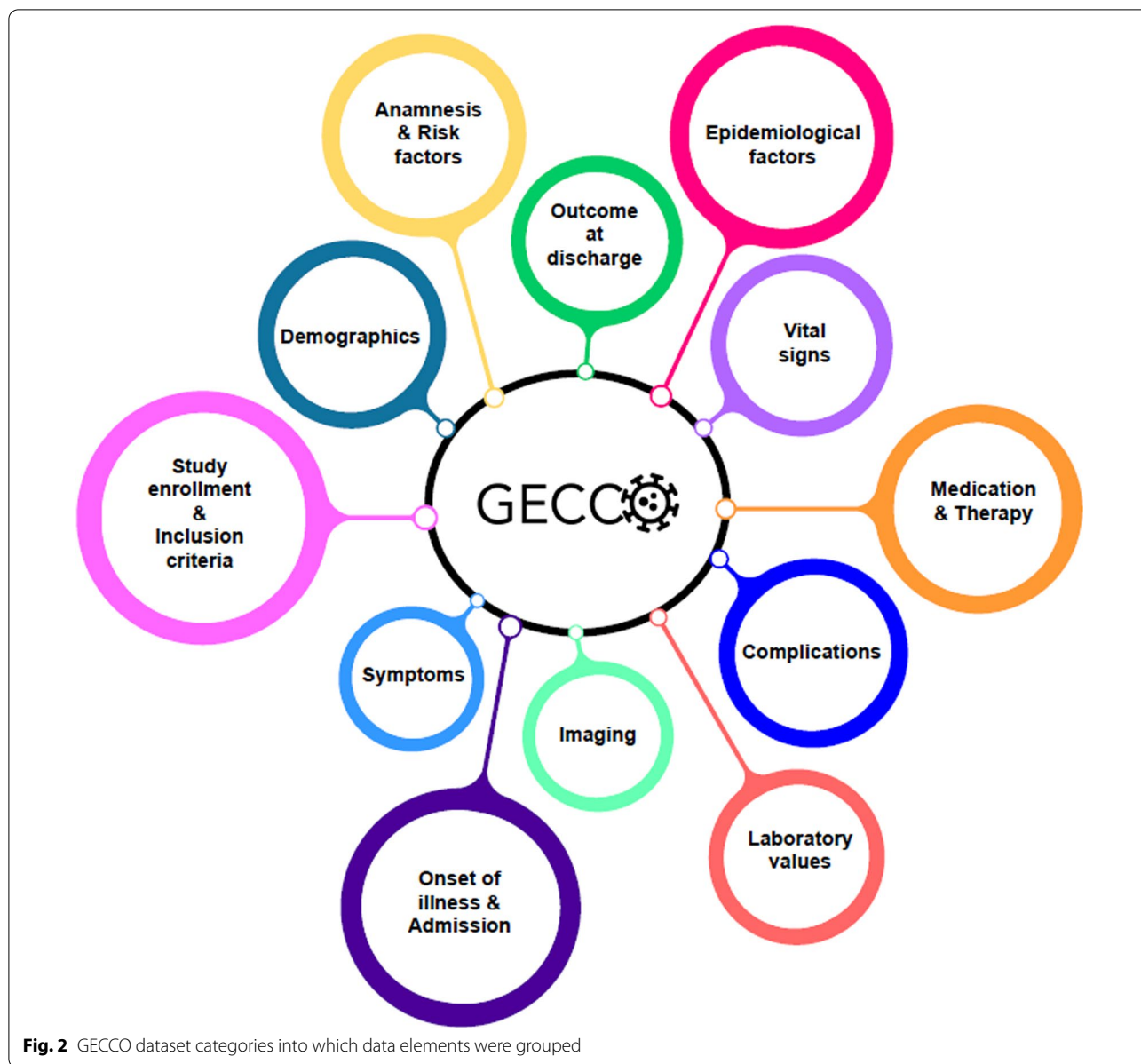


Fig. 2 GECCO dataset categories into which data elements were grouped

sets of the GECCO dataset can be accessed on the ART-DECOR platform [28].

Subsequently, FHIR profiles were created for the data elements. The following FHIR resources were used to model the data elements: Patient, Consent, Observation, Condition, Procedure, Encounter, Medication and MedicationStatement. The FHIR profiles can be accessed on Simplifier [29].

During the consolidation process, it became clear that some data elements are important for certain disciplines but irrelevant for others. These elements were not included in the GECCO core dataset as they would have inflated the size of the dataset. The editorial team decided

to include these data elements in domain-specific extension modules, which will be specified in more detail at later stages of the project.

Discussion

In this report, we presented the GECCO dataset, a core collection of data elements for acquiring and exchanging information about COVID-19 patients. By using standardized data structures (HL7 FHIR profiles) and international terminologies, the GECCO dataset is an important step towards interoperability of COVID-19 research data. It can facilitate harmonized data collection and analysis

across institutions and IT systems, for example in clinical studies, registries or digital health applications.

A key factor to the successful application of standard datasets like GECCO is a close collaboration with the scientific community. To ensure a high acceptance of the dataset, the development of GECCO therefore included clinicians from a wide variety of medical disciplines and professional associations as well as experts in digital health, standardization and clinical terminologies. GECCO also collaborates closely with standards developing organizations such as HL7 and Integrating the Healthcare Enterprise (IHE) as well as other initiatives aiming to improve health data interoperability, such as the Medical Informatics Initiative [12], NFDI4Health [30] and the Corona Component Standards (cocos) [31].

For the successful application of standard datasets like GECCO, it is also important that these datasets are embedded in larger infrastructures for secure and interoperable data sharing across institutions. Initiatives like, for example, the National COVID Cohort Collaborative (N3C) in the US [32], OpenSAFELY in the UK [33] or the international project Secure Collective Research (SCOR) [34] are developing platforms for a secure, cross-institutional analysis of COVID-19 data. Similarly, GECCO is part of the German COVID-19 Research Network of University Medicine [10], which aims to bundle the resources of German university hospitals to improve diagnostics and treatment of COVID-19 patients. The network also includes a research data infrastructure for the secure and interoperable data exchange across university hospitals [35], for which GECCO provides a standard data structure. For example, projects such as NAPKON, a national project for collecting research data during pandemics [36], will collect their data according to the specifications of the GECCO dataset.

Note that the GECCO dataset was originally developed for German university medicine. However, we used international standards and terminologies to ensure that the GECCO dataset can be extended to use cases beyond university medicine and also be applied in international contexts. When defining the FHIR profiles, we also took into account international work (for example the IPS [23]) to make sure that the GECCO dataset can provide interoperability when used internationally.

Although the GECCO dataset was designed to be as compact and manageable as possible, acquiring and recording the information for all data elements still requires time (for example, when entering the information in an electronic case report form). Moreover, manual documentation is prone to transcription errors. Conversely, manually abstracted and structured information from unstructured health records may provide relevant insights for care-providers and improve their

understanding of risk and outcome. For some of the data items, it is therefore desirable to automatically exchange data between a GECCO-based study database and existing IT systems, such as hospital information systems or clinical trial software. This requires standard interfaces between these systems. The FHIR profiles of the GECCO dataset provide an interoperable, machine-readable data structure that can facilitate this data exchange across IT systems. For example, LOINC-coded information about patients' laboratory values could be directly transferred from the hospital information system to the GECCO dataset. For electronic data capture (EDC) systems used in clinical studies, converters are currently being developed that transform the underlying software formats into the GECCO HL7 FHIR format for interoperable data exchange. Independently of the work presented here, the GECCO dataset has also been converted to the CDISC Operational Data Model (ODM) and is published on the Portal of Medical Data Models (MDM) of the University of Münster [37].

The aim of the GECCO dataset was to define a compact set of core data elements for which most COVID-19 studies (particularly the studies conducted at German university hospitals) can provide full information. However, if not explicitly required, studies that want to use the GECCO dataset may use subsets of the dataset if they are unable to provide full information for all data elements.

Scientific knowledge about COVID-19 and SARS-CoV-2 is changing fast, which may necessitate modifications to the GECCO dataset in the future. Furthermore, the use of the GECCO dataset in clinical research projects will provide practical experience that may also motivate changes to the dataset. To incorporate new knowledge into the dataset, the COVID-19 Research Network of University Medicine [10] will put a governance framework in place that will coordinate revisions and extensions to the dataset. Domain-specific extension modules are already in preparation, which include many of the data elements that were not considered essential for the core dataset. Extension modules currently planned are: laboratory, diagnostics, immunology, gynecology and pregnancy, epidemiology, pediatrics, intensive care, oncology, radiology, virology, psychiatry and neurology (these extension modules are also made accessible on the ART-DECOR platform [38]).

Conclusion

The GECCO dataset provides researchers and healthcare professionals with a compact, interoperable dataset for collecting, exchanging and analyzing COVID-19 data across institutions and software systems. Developed by a multidisciplinary group of experts, GECCO builds heavily on international terminologies and IT standards.

GECCO can thus help to improve the harmonization and coordination of research efforts to successfully fight the COVID-19 pandemic. Future inclusion of domain-specific extension modules will further expand the use of the GECCO dataset.

Abbreviations

ATC: Anatomical Therapeutic Chemical Classification System; BMBF: Bundesministerium für Bildung und Forschung (German Federal Ministry of Education and Research); CDE: Common Data Element; CDISC: Clinical Data Interchange Standards Consortium; cocos: Corona Component Standards; COVID-19: Coronavirus Disease 2019; CRF: Case Report Form; DGI: Deutsche Gesellschaft für Infektiologie (German Society for Infectiology); DZIF: Deutsches Zentrum für Infektionsforschung (German Center for Infection Research); EITaF: Emerging Infections Task Force; ESCMID: European Society of Clinical Microbiology and Infectious Diseases; FHIR: Fast Healthcare Interoperability Resources; GECCO: German Corona Consensus Dataset; HL7: Health Level 7; ICD-10-GM: International Statistical Classification of Diseases and Related Health Problems, 10th revision, German modification; IHE: Integrating the Healthcare Enterprise; IPS: International Patient Summary; ISARIC: International Severe Acute Respiratory and Emerging Infection Consortium; IT: Information Technology; LEOSS: Lean European Open Survey on SARS-CoV-2 Infected Patients; LOINC: Logical Observation Identifiers Names and Codes; MDM: Medical Data Models; N3C: National COVID Cohort Collaborative; NAPKON: Nationales Pandemie Kohorten Netz (National Pandemic Cohort Network); NFDI: Nationale Forschungsdateninfrastruktur (National Research Data Infrastructure); NIH: National Institutes of Health; ODM: Operational Data Model; SARS-CoV-2: Severe Acute Respiratory Syndrome Coronavirus 2; SCOR: Secure Collective Research; UCU: Unified Code for Units of Measure; WHO: World Health Organization.

Acknowledgements

We thank the members of the expert board as well as Thomas Bahmer, Michael von Bergwelt, Marie von Lilienfeld-Toal, Patrick Meybohm and Ulrich Sax from the editorial team for their help with the development of the dataset.

Authors' contributions

All authors (JS, AB, ML, AE, ER, SR, KH, JV, CK, ST) contributed to the development of the GECCO dataset. JS, AB, AE and ER mapped the data elements to international terminologies. JS, AB and AE defined the dataset in ART-DECOR. JS developed the FHIR profiles. ML wrote the manuscript. All authors read and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials

Data elements, response options and value sets of the GECCO dataset (including future developments and extension modules) can be accessed on the ART-DECOR platform [28]. FHIR profiles are available on Simplifier [29].

Ethics approval and consent to participate

No permissions were required to access any data used in this study.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Author details

¹ Berlin Institute of Health (BIH), Berlin, Germany. ² Charité – Universitätsmedizin Berlin, Berlin, Germany. ³ HIH – Health Innovation Hub of the Federal Ministry of Health, Berlin, Germany. ⁴ Medical Department 2, Hematology/Oncology, University Hospital of Frankfurt, Frankfurt, Germany. ⁵ Department I for Internal Medicine, University Hospital Cologne, Cologne, Germany. ⁶ German Center for Infection Research, Partner Site Bonn-Cologne, Cologne,

Germany. ⁷ Hochschule Niederrhein – University of Applied Sciences, Krefeld, Germany.

Received: 29 July 2020 Accepted: 16 December 2020

Published online: 21 December 2020

References

1. WHO. Novel Coronavirus-China, Disease outbreak news: Update 12 January 2020. <http://www.who.int/csr/don/12-january-2020-novel-coronavirus-china/en>. Accessed 16 Nov 2020.
2. Johns Hopkins Coronavirus Resource Center. COVID-19 Map. <https://coronavirus.jhu.edu/map.html>. Accessed 16 Nov 2020.
3. Lean European Open Survey on SARS-CoV-2 Infected Patients. <https://leoss.net>. Accessed 16 Nov 2020.
4. GESIS Panel Team. GESIS Panel Special Survey on the Coronavirus SARS-CoV-2 Outbreak in Germany. GESIS Datenarchiv, Köln. ZA5667 Datenfile Version 1.1.0. 2020. <https://doi.org/10.4232/1.13520>.
5. WHO tool for behavioural insights on COVID-19. <https://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/technical-guidance/who-tool-for-behavioural-insights-on-covid-19>. Accessed 16 Nov 2020.
6. Timpson N, Haworth S, Angelantonio ED, Herbst K, Packer R, Steves C, et al. UK Covid-19 Questionnaire. 2020. https://www.nlm.nih.gov/dr2/UK_COVID19_Final_Questionnaire_23_April.pdf#. Accessed 16 Nov 2020.
7. Docherty AB, Harrison EM, Green CA, Hardwick HE, Pius R, Norman L, et al. Features of 20 133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: prospective observational cohort study. *BMJ*. 2020. <https://doi.org/10.1136/bmj.m1985>.
8. ISARIC. COVID-19 CRF. <https://isaric.tghn.org/COVID-19-CRF>. Accessed 16 Nov 2020.
9. Center for Disease Control and Prevention (CDC). Human Infection with 2019 Novel Coronavirus Person Under Investigation (PUI) and Case Report Form. https://www.phenxtoolkit.org/toolkit_content/PDF/CDC_PUI.pdf. Accessed 16 Nov 2020.
10. Netzwerk Universitätsmedizin. <https://www.netzwerk-universitaetsmedizin.de>. Accessed 16 Nov 2020.
11. Kurth F, Roennefarth M, Thibeault C, Corman VM, Mueller-Redetzky H, Mittermaier M, et al. Studying the pathophysiology of coronavirus disease 2019—a protocol for the Berlin prospective COVID-19 patient cohort (Pa-COVID-19). *medRxiv*. 2020. <https://doi.org/10.1101/2020.05.06.20092833>.
12. Medical Informatics Initiative. <https://www.medizininformatik-initiative.de/en>. Accessed 16 Nov 2020.
13. Kassenärztliche Bundesvereinigung (KBV). <https://www.kbv.de>. Accessed 16 Nov 2020.
14. National Institutes of Health (NIH). Classifications of Data Elements for a Particular Disease. <https://www.commondataelements.ninds.nih.gov/glossary>. Accessed 16 Nov 2020.
15. Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM). ICD-10-GM. <https://www.dimdi.de/dynamic/en/classifications/icd/icd-10-gm>. Accessed 16 Nov 2020.
16. McDonald CJ, Huff SM, Suico JG, Hill G, Leavelle D, Aller R, et al. LOINC, a universal standard for identifying laboratory observations: a 5-year update. *Clin Chem*. 2003;49:624–33.
17. The Unified Code for Units of Measure. <http://unitsofmeasure.org>. Accessed 16 Nov 2020.
18. WHO Collaborating Centre for Drug Statistics Methodology. International language for drug utilization research. <https://www.whocc.no>. Accessed 16 Nov 2020.
19. SNOMED International. <https://www.snomed.org>. Accessed 16 Nov 2020.
20. ART-DECOR. <https://www.art-decor.org>. Accessed 16 Nov 2020.
21. HL7 FHIR. <https://hl7.org/FHIR>. Accessed 16 Nov 2020.
22. Medical Informatics Initiative. FHIR profiles. <https://simplifier.net/organization/koordinationsstelleml/~home>. Accessed 16 Nov 2020.
23. International Patient Summary Implementation Guide. <http://hl7.org/fhir/uv/ips/2018Sep>. Accessed 16 Nov 2020.
24. Logica Implementation Guide: Covid-19. <https://covid-19-ig.logicahealth.org/index.html>. Accessed 16 Nov 2020.
25. HL7 Deutschland e.V. Basisprofil DE (R4). <https://simplifier.net/basisprofil-de-r4>. Accessed 16 Nov 2020.

26. Forge. <https://fire.ly/products/forge>. Accessed 16 Nov 2020.
27. SIMPLIFIER.NET. <https://simplifier.net>. Accessed 16 Nov 2020.
28. GECCO Dataset. <https://art-decor.org/art-decor/decor-datasets--covid-19/?id=2.16.840.1.113883.3.1937.777.53.1.1&effectiveDate=2020-04-08T13%3A04%3A13&language=de-DE>. Accessed 16 Nov 2020.
29. GECCO FHIR profiles. <https://simplifier.net/ForschungsnetzCovid-19>. Accessed 16 Nov 2020.
30. NFDI4Health. <https://www.nfdi4health.de>. Accessed 16 Nov 2020.
31. cocos-Corona Component Standards. <http://cocos.team>. Accessed 16 Nov 2020.
32. Haendel MA, Chute CG, Gersing K. The National COVID cohort collaborative (N3C): rationale, design, infrastructure, and deployment. *J Am Med Inform Assoc*. 2020. <https://doi.org/10.1093/jamia/ocaa196>.
33. OpenSAFELY. <https://opensafely.org>. Accessed 16 Nov 2020.
34. Raisaro JL, Marino F, Troncoso-Pastoriza J, Beau-Lejdstrom R, Bellazzi R, Murphy R, et al. SCOR: a secure international informatics infrastructure to investigate COVID-19. *J Am Med Inform Assoc*. 2020. <https://doi.org/10.1093/jamia/ocaa172>.
35. CODEX|Netzwerk Universitätsmedizin. <https://www.netzwerk-universitaetsmedizin.de/projekte/codex>. Accessed 16 Nov 2020.
36. NAPKON|Netzwerk Universitätsmedizin. <https://www.netzwerk-universitaetsmedizin.de/projekte/napkon>. Accessed 16 Nov 2020.
37. Dugas M, Neuhaus P, Meidt A, Doods J, Storck M, Bruland P, et al. Portal of medical data models: information infrastructure for medical research and healthcare. *Database J Biol Databases Curation*. 2016. <https://doi.org/10.1093/database/bav121>.
38. GECCO Extension Modules. <https://art-decor.org/art-decor/decor-datasets--covid19/?id=2.16.840.1.113883.3.1937.777.53.1.2&effectiveDate=2020-08-12T00%3A00%3A00&language=de-DE>. Accessed 16 Nov 2020.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

