The German Corona Consensus Dataset (GECCO): A standardized dataset for COVID-19 research

Julian Sass[1], Alexander Bartschke[2], Moritz Lehne[1], Andrea Essenwanger[1], Eugenia Rinaldi[2], Stefanie Rudolph[2], Kai U. Heitmann[3], Jörg J. Vehreschild[4,5,6], Christof von Kalle[1,2], Sylvia Thun[1,2,7]

[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

* Correspondence concerning this article should be addressed to Sylvia Thun, E-mail: sylvia.thun@bihealth.de

Keywords: COVID-19, interoperability, standard dataset, FHIR

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 segmentation 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.

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, anamnesis, 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.

INTRODUCTION

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 July 1, 2020), SARS-CoV-2 has infected more than 10 million and killed more than half 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 segmentation of information. This could delay or even prevent urgently needed scientific knowledge about SARS-CoV-2 and COVID-19. To avoid this segmentation 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 also have to be embedded in standard data structures that can be exchanged across IT systems, and they 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 National Research Network of University

Medicine on COVID-19 ("Nationales Forschungsnetzwerk [NFN] der Universitätsmedizin zu COVID-19") funded by the German Federal Ministry of Education and Research (BMBF). The following paper 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[10], 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 German university hospitals, professional associations and other relevant organizations. 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[11] (Table 1).

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.*

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; 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.

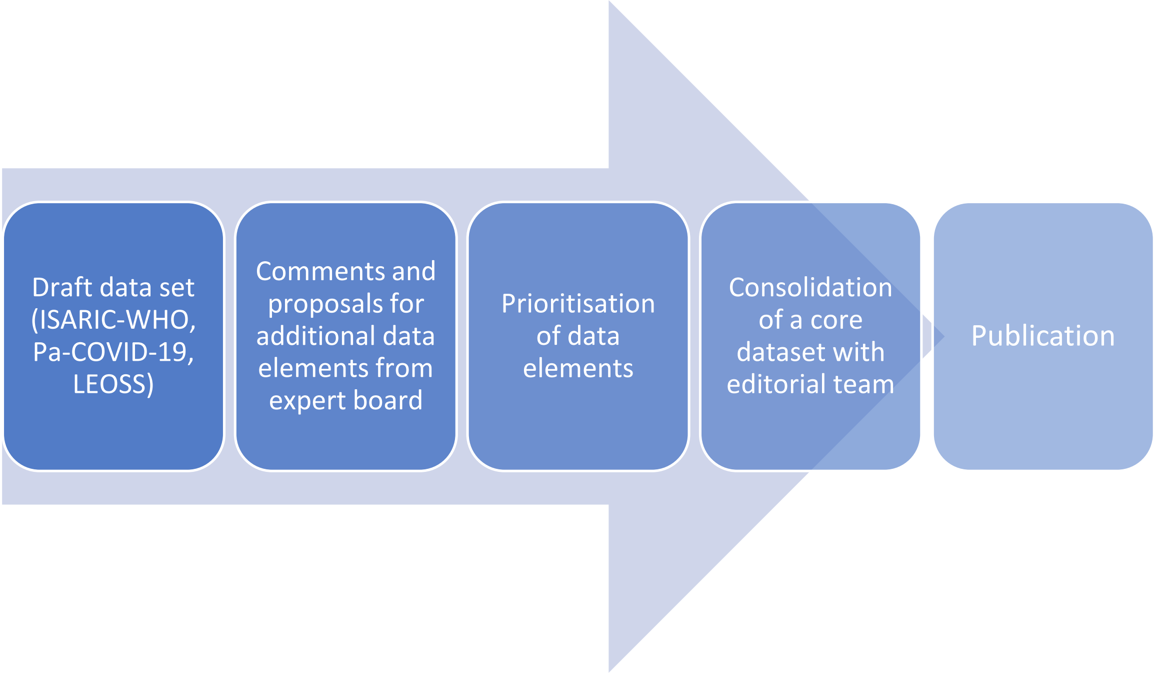
Draft data set (ISARIC-WHO, Pa-COVID-19, LEOSS)
Comments and proposals for additional data elements from expert board
Prioritisation of data elements
Consolidation of a core dataset with editorial team
Publication

Figure 1: Workflow of consensus building and definition of data elements for the GECCO core dataset.

Standardization

To ensure syntactic and semantic interoperability, elements and response options of the core dataset were mapped to international standards and terminologies. 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)[12] for diagnoses; Logical Observation Identifiers Names and Codes (LOINC)[13] for laboratory values and other measurements; the Unified Code for Units of Measure (UCUM)[14] for measurement units; the Anatomical Therapeutic Chemical Classification System (ATC)[15] for active ingredients of drugs and medications; SNOMED CT[16] for diagnoses and other medical concepts. The annotation of data elements with international terminologies was done using ART-DECOR[17], 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)[18] 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[19], the International Patient Summary (IPS)[20], the Logica COVID-19 profiles[21] and the FHIR base profiles of HL7 Germany.[22] FHIR profiles were defined using Forge[23] and published on the Simplifier platform.[24]

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) (Figure 2).

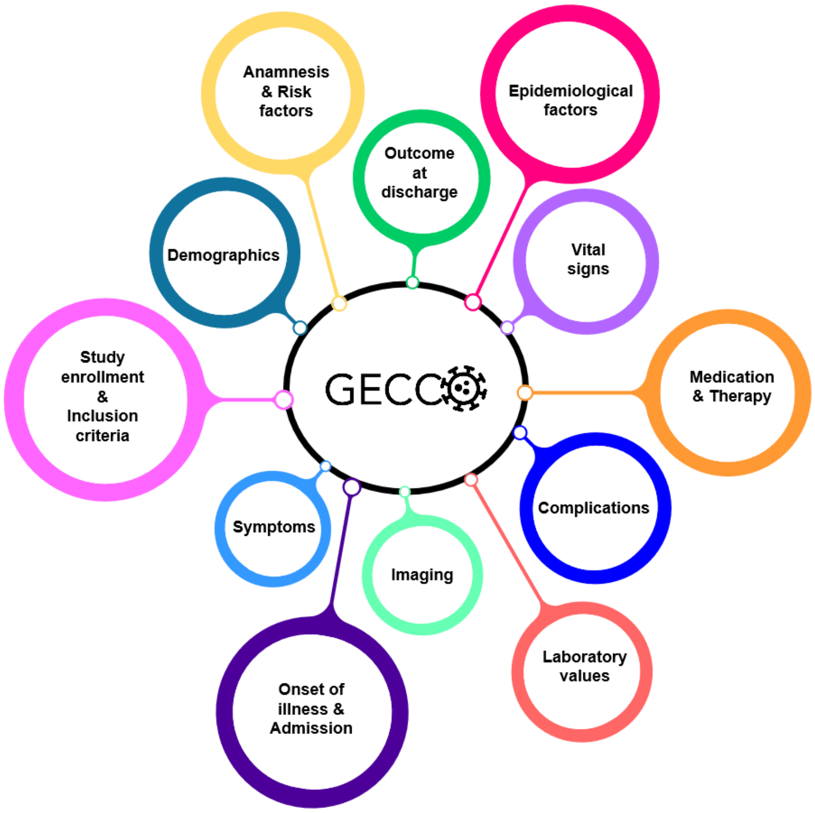

Figure 2: GECCO dataset categories into which data elements were grouped.

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 sets can be accessed at https://art-decor.org/art-decor/decor-datasets--covid19f-.

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 at https://simplifier.net/ForschungsnetzCovid-19.

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 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[25], NFDI4Health[26] and the Corona Component Standards (cocos)[27]. Building on this strong consensus, GECCO-based data collection now has become a requirement for projects funded by the National Research Network of University Medicine on COVID-19 (NFN).

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.

Scientific knowledge about COVID-19 and SARS-CoV-2 is changing fast, which may necessitate modifications to the GECCO dataset in the future. To incorporate new knowledge into the dataset, the NFN will put a governance framework in place that will coordinate revisions and extensions to the dataset. Domain-specific extension modules are already in preparation. 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 accessible on the ART-DECOR platform).

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.

ACKNOWLEDGMENTS

We thank the members of the expert board and editorial team for their help with the development of the GECCO dataset.

DATA AVAILABILITY

Data elements, response options and value sets of the GECCO dataset can be accessed at https://art-decor.org/art-decor/decor-datasets--covid19f-. FHIR profiles are available at https://simplifier.net/ForschungsnetzCovid-19.

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/
2. Johns Hopkins Coronavirus Resource Center | COVID-19 Map, https://coronavirus.jhu.edu/map.html
3. Lean European Open Survey on SARS-CoV-2 Infected Patients, https://leoss.net
4. GESIS Panel Team (2020). GESIS Panel Special Survey on the Coronavirus SARS-CoV-2 Outbreak in Germany. *GESIS Datenarchiv, Köln. ZA5667 Datenfile Version 1.1.0,* 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
6. Timpson, N. et al. (2020). UK Covid-19 Questionnaire, https://www.nlm.nih.gov/dr2/UK_COVID19_Final_Questionnaire_23_April.pdf#
7. Docherty, A. B. et al. (2020). Features of 20 133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: prospective observational cohort study. *BMJ* 369:m1985, https://doi.org/10.1136/bmj.m1985
8. ISARIC | Clinical Data Collection – The COVID-19 Case Report Forms (CRFs), https://isaric.tghn.org/COVID-19-CRF
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
10. Kurth, F. et al. (2020). Studying the pathophysiology of coronavirus disease 2019 - a protocol for the Berlin prospective COVID-19 patient cohort (Pa- COVID-19). *medRxiv,* https://doi.org/10.1101/2020.05.06.20092833

11. National Institutes of Health (NIH) | Classifications of Data Elements for a Particular Disease, https://www.commondataelements.ninds.nih.gov/glossary
12. Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM) | ICD-10-GM, https://www.dimdi.de/dynamic/en/classifications/icd/icd-10-gm
13. McDonald, C. J. et al. (2003). LOINC, a Universal Standard for Identifying Laboratory Observations: A 5-Year Update. *Clin Chem* 49, 624–633
14. The Unified Code for Units of Measure, http://unitsofmeasure.org
15. WHO Collaborating Centre for Drug Statistics Methodology | International language for drug utilization research, https://www.whocc.no
16. SNOMED International, https://www.snomed.org
17. ART-DECOR, https://www.art-decor.org
18. HL7 FHIR, https://hl7.org/FHIR
19. Medical Informatics Initiative | FHIR profiles, https://simplifier.net/organization/koordinationsstellemii/~home
20. International Patient Summary Implementation Guide, http://hl7.org/fhir/uv/ips/2018Sep
21. Logica Implementation Guide: Covid-19, https://covid-19-ig.logicahealth.org/index.html
22. HL7 Deutschland e.V. | Basisprofil DE (R4), https://simplifier.net/basisprofil-de-r4
23. Forge, https://fire.ly/products/forge
24. SIMPLIFIER.NET - The FHIR collaboration platform, https://simplifier.net
25. Medical Informatics Initiative, https://www.medizininformatik-initiative.de/en
26. nfdi4health, https://www.nfdi4health.de
27. cocos – Corona Component Standards, http://cocos.team