Title page

Title: Extending the German Corona Consensus Dataset (GECCO) to the immunization, pediatrics, and cardiology domains

Authors: Gregor Lichtner[1,2,3], Thomas Haese[1], Sally Brose[4], Larissa Röhrig[1,5], Liudmila Lysyakova[6,7], Stefanie Rudolph[6,7], Maria Uebe[6,7], Julian Sass[1], Alexander Bartschke[1], David Hillus[8], Florian Kurth[8,9], Leif Erik Sander[8], Falk Eckart[10], Nicole Toepfner[10], Reinhard Berner[10], Anna Frey[11], Marcus Dörr[12], Jörg Janne Vehreschild[13,14,15], Christof von Kalle[6,7], Sylvia Thun[1]

[1] Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Berlin, Germany

[2] Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Institute of Medical Informatics, Berlin, Germany

[3] Universitätsmedizin Greifswald, Department of Anesthesia, Critical Care, Emergency and Pain Medicine, Greifswald, Germany

[4] Robert Koch Institute, Department of Methodology and Research Infrastructure, Research Data Management, Berlin, Germany

[5] National Association of Statutory Health Insurance Physicians ("Kassenärztliche Bundesvereinigung"; KBV), Digitalization and IT, Department Interoperability, Berlin, Germany

[6] Charité - Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, joint Charité and BIH Clinical Study Center, Berlin, Germany

[7] Berlin Institute of Health at Charité – Universitätsmedizin Berlin, joint Charité and BIH Clinical Study Center, Berlin, Germany

[8] Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Department of Infectious Diseases and Respiratory Medicine, Berlin, Germany

[9] University Medical Centre Hamburg-Eppendorf, Department of Tropical Medicine, Bernhard Nocht Institute for Tropical Medicine and Department of Medicine I, Hamburg, Germany

[10] Department of Pediatrics, University Hospital Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany

[11] University Hospital of Würzburg, Medical Clinic and Policlinic I, Würzburg, Germany

[12] Universitätsmedizin Greifswald, Department of Internal Medicine B, Greifswald, Germany

[13] Department I of Internal Medicine, University Hospital of Cologne, Cologne, Germany

[14] German Centre for Infection Research (DZIF), Partner Site Bonn-Cologne, Cologne, Germany

[15] Department II of Internal Medicine, Hematology/Oncology, Goethe University, Frankfurt, Frankfurt am Main, Germany

Address of corresponding author:

Dr. Gregor Lichtner

Berlin Institute of Health at Charité – Universitätsmedizin Berlin

Anna-Louisa-Karsch-Str. 2

10178 Berlin, Germany

Abstract

Background

The COVID-19 pandemic has spurred large-scale, inter-institutional research efforts. To enable these efforts, the German Corona Consensus (GECCO) dataset has been developed previously as a harmonized, interoperable collection of the most relevant data elements for COVID-19-related patient research. As GECCO has been developed as a compact core dataset across all medical fields, the focused research within particular medical domains demanded the definition of extension modules that include those data elements that are most relevant to the research performed in these individual medical specialties.

Main body

We created GECCO extension modules for the *immunization*, *pediatrics,* and *cardiology* domains with respect to the pandemic requests. The data elements included in each of these modules were selected in a consensus-based process by working groups of medical experts from the respective specialty to ensure that the contents are aligned with the research needs of the specialty. The selected data elements were mapped to international standardized vocabularies and data exchange specifications were created using HL7 FHIR profiles on the appropriate resources. All steps were performed in close interdisciplinary collaboration between medical domain experts, medical information scientists and FHIR developers. The profiles and vocabulary mappings were syntactically and semantically validated in a two-stage process. In that way, we defined dataset specifications for a total number of 23 (*immunization*), 59 (*pediatrics*), and 50 (*cardiology*) data elements that augment the GECCO core dataset. We created and published implementation guides and example implementations as well as dataset annotations for each extension module.

Conclusions

We here present extension modules for the GECCO core dataset that contain data elements most relevant to COVID-19-related patient research in *immunization*, *pediatrics* and *cardiology*. These extension modules were defined in an interdisciplinary, iterative, consensus-based approach that may serve as a blueprint for the development of further dataset definitions and GECCO extension modules. The here developed GECCO extension modules provide a standardized and harmonized definition of specialty-related datasets that can help to enable inter-institutional and cross-country COVID-19 research in these specialties.

Keywords

- COVID-19
- Interoperability
- GECCO dataset
- FHIR

Background

The COVID-19 pandemic has led to unprecedented strong efforts in connecting nationwide and international research to help in managing the disease and its effects on public health. To enable research across different health care providers, institutions or even countries, interoperability between the medical data systems is essential [1]. Therefore, early in the pandemic, the German Corona Consensus Dataset (GECCO) has been developed in a collaborative effort to provide a standardized, unified core dataset for inter-institutional COVID-19-related patient research [2]. The GECCO dataset specifies a set of 81 essential clinical data items from 13 domains such as anamnesis & risk factors, symptoms, and vital signs, that have been selected by expert committees from university hospitals, professional associations, and research initiatives. The selected data items were mapped to international standard vocabularies such as the *Systematized Nomenclature of Medicine - Clinical Terms* (SNOMED CT) and *Logical Observation Identifiers Names and Codes* (LOINC). The *Fast Healthcare Interoperability Resources* (FHIR) standard was used to specify the interoperable data exchange format for the GECCO data items. Since its development, the GECCO dataset has been implemented in a large number of institutions, most notably in virtually every German university hospital, which now provides access to the GECCO dataset in the context of the German COVID-19 Research Network of University Medicine ("Netzwerk Universitätsmedizin") [3].

The GECCO dataset was developed to contain as many relevant data items as possible, but few enough to keep the effort of implementing the dataset manageable. Therefore, the dataset contains mostly data items of general research interest, excluding data items that are only of interest for particular medical specialties or use cases. These data items are considered part of domain-specific extensions to the GECCO dataset introduced in this article.

We here introduce three domain-specific extensions of the GECCO dataset that cover COVID-19-related research data items relevant to the medical specialties of *immunization*, *pediatrics,* and *cardiology*. These extensions use the same international health IT standards and terminologies as the GECCO dataset and were developed in close alignment with the GECCO dataset to ensure interoperability and compatibility with existing definitions. We describe the iterative, consensus-based data item selection and data format definition process that was performed in close collaboration between medical experts from *immunology*, *pediatrics,* and *cardiology* domains on the content definition side and medical information specialists and FHIR developers on the technical side. This process may serve as a blueprint for further development of consensus-based data set definitions.

Construction and Contents

Selection of data items

The content of the domain-specific GECCO extension modules was defined by medical domain experts in a transparent process (Figure 1). The involvement of the medical domain experts as the end-users of the data to be provided ensured that the contents of the extension modules are aligned to the actual research needs. In the context of the national pandemic cohort net ("Nationales Pandemie Kohorten Netz"; NAPKON) project of NUM [4], so-called subject- and organ-specific working groups were established by the voluntary association of medical experts from different medical fields. Each of the subject- and organ-specific working groups elected a board, and all communication between the GECCO extension module developers and the working groups was organized and carried out via the working groups' board. In preparation for the GECCO extension modules, the subject- and organ-specific working groups for *immunology*, *pediatrics,* and *cardiology* were invited to provide up to 50 data items with up to 10 response options each that were of particular interest to their field concerning COVID-19-related research. If necessary, more data items or response options could be provided in coordination with the GECCO extension module developers. The provided data items were then reviewed by the GECCO extension module developer team and a first definition of the contents of the extension module was returned to the respective subject- and organ-specific working group for approval or change requests. After approval by the subject- and organ-specific working group, the definition of the extension module was considered finalized. Note that for the *immunization* extension module, physicians from the "NUM-COVIM" study for the determination and use of SARS-CoV-2 immunity [5] assumed the role of the organ-specific working group, as no such working group had been established previously.

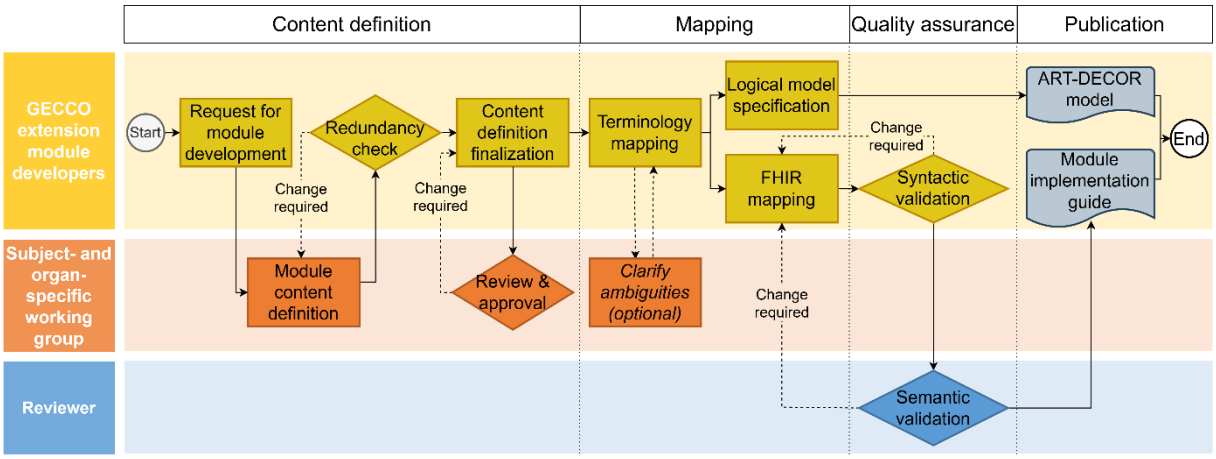

Figure 1 Flowchart of the consensus-based, interdisciplinary dataset definition and mapping process for the GECCO extension modules.

Development of the standardized data formats

To map the data items selected by the subject- and organ-specific working groups to international standard vocabularies, we performed a consensus-based mapping procedure, where every concept was mapped to appropriate vocabularies SNOMED CT for general concepts [6], LOINC for observations [7], *International Statistical Classification of Diseases and Related Health Problems, 10th revision, German modification* (ICD-10-GM) for diagnoses [8], *Anatomical Therapeutic Chemical Classification System* (ATC) for Germany for drugs and active ingredients [9], *Unified Code for Units of Measure* (UCUM) for measurement units [10]) by two medical information scientists independently. Ambiguities and non-matching mappings were then discussed within the GECCO extension module developer team and in close collaboration with the medical experts of the subject- and organ-specific working groups until consensus was achieved. The data item-to-concept mappings were annotated on ART-DECOR, an open-source collaboration platform for creating and maintaining dataset element descriptions [11].

As for the GECCO dataset, the format for data exchange was specified using HL7 FHIR resources. The mapping of data items to FHIR resources was performed in an iterative, consensus-based process among the GECCO extension module developer team. Wherever possible, published FHIR profiles from the GECCO dataset, from the Medical Informatics Initiative (MII) [12] or the National Association of Statutory Health Insurance Physicians ("Kassenärztliche Bundesvereinigung"; KBV) [13] – in this order of priority – served as the base definition for the future extension module profiles.

The profiles and value sets were specified using the FHIR Shorthand (FSH) language (version 1.2.0) and translated to Structure Definition JSON files using the HL7 FSH SUSHI software package (version 2.2.3) [14, 15]. We required that at least one exemplary instance be defined for every profile. Syntactic validation of the profiles and value sets definitions was performed using the error-free conversion of the FSH files to JSON using SUSHI and subsequent validation of each profile with their defined instances using the HL7 FHIR validator as implemented in the FHIR Shorthand Validator Python package (version 0.2.2) [16]. After successful syntactic validation of a set of profiles, the profiles were subjected to a two-stage review process as follows. First, the profiles and corresponding value sets and extensions were internally reviewed for semantic appropriateness with the GECCO core developer (JS). After all necessary changes and approval by the internal reviewer, the profiles were subjected to the second review round by an external FHIR development expert. Subsequent to necessary corrections and approval of the external reviewer, the respective profiles together with their value sets and optionally extensions and code systems were considered finalized and published to the main branch of the git repository.

The whole development process was performed collaboratively on GitHub. Syntactic validation of the profiles was performed by continuous integration/continuous development (CI/CD) workflows

implemented as GitHub actions. Semantic validation during the internal and external review rounds was performed using pull requests to two different git branches. After the final approval, profiles and value sets were merged into the main branch of the extension module's repository, which served as the publication branch of the respective module. Since then, maintenance requests and updates of the extension modules are handled via GitHub issues. All kinds of relevant changes become a subject of the internal review as defined above; major changes (e.g., non-technical corrections) are additionally exposed to the external review.

Implementation guides were created for all three extension modules using the FHIR IG publisher tool and a customized template for the implementation guide's HTML pages [17]. The implementation guides are published to GitHub pages and remain automatically synchronized with the main branch of the respective repository via CI/CD workflows.

Contents

A total number of 23 for the *immunization* extension module, 59 data items for the *pediatrics* extension module, and 50 for the *cardiology* extension module have been collected via an iterative consensus-based approach from the respective subject- and organ-specific working groups, belonging to 10 of the 12 data categories of the GECCO dataset (Table 1).

GECCO data category	**Extension module**		
	Pediatrics	**Cardiology**	**Immunization**
Anamnesis & Risk factors	7	6	6
Complications	6	7	5
Demographics	3	-	-
Epidemiological factors	-	-	2
Imaging	1	33	-
Laboratory values	28	2	1
Medication & Therapy	13	1	3
Onset of illness & admission	-	-	3
Outcome at discharge	-	-	-
Study enrollment & Inclusion criteria	-	-	-
Symptoms	1	-	3
Vital signs	-	1	-
Total items	**59**	**50**	**23**

Table 1 Number of data items per GECCO dataset category for each extension module.

All data items were mapped to the appropriate FHIR based resources Observation, Condition, Procedure, MedicationStatement, Encounter, Questionnaire, QuestionnaireResponse, Immunization, ImagingStudy, List, and Specimen, and 26, 14, and 18 profiles (25, 17, and 12 value sets) were created for the *immunization*, *pediatrics*, and *cardiology* extension modules, respectively. The data items that

were already part of the GECCO dataset and that were not removed during the data selection step were taken over from GECCO and referenced as such in the implementation guides.

The implementation guides for the three extension modules have been published on GitHub pages [18–20]. The source FHIR ShortHand (FSH) files have been published on GitHub [21–23]. Dataset descriptions can be found on ART-DECOR [24–26]

Utility and Discussion

We here present three extension modules of the GECCO dataset that contain data items relevant for COVID-19-related patient research in the *immunization*, *pediatrics*, and *cardiology* fields. The data items are represented in HL7 FHIR profiles and use international terminologies, to ensure a harmonized, standardized, and interoperable dataset definition for these medical domains. The provision of data according to the dataset definitions introduced in this article will enable cross-institutional and cross-country data collection and collaborative research with a particular focus in *immunization*, *pediatrics*, and *cardiology*.

We have implemented an interdisciplinary, iterative, consensus-based process for the selection of data items and the development of the dataset definition. The close collaboration and the constant feedback loops with domain experts from the respective medical specialties right from the beginning of the project, as performed here, are key for the successful development of a useful dataset definition. Indeed, since the selection of relevant data items was driven by the end-users of the dataset, who are the researchers that later will be using the data for their specialized areas of research, the semantic usability of the datasets is guaranteed. Likewise, having medical information specialists develop the formal dataset specification ensures technical interoperability and usability of the dataset definition.

Next to the successful development of dataset definitions, several factors determine a successful deployment or use of the developed dataset definitions. First and most importantly, clear and concise documentation of how to implement and provide data using the dataset definition is required. For FHIR-based dataset definitions, so-called implementation guides are used to provide both a narrative overview as well as deep technical details on the dataset definition. Thus, we have created and published implementation guides for each of the here-developed GECCO extensions modules. Second, the example implementations of the dataset definitions serve as a blueprint for developers and data engineers who implement the dataset definitions for their clinical databases. From our experience with the implementation of the GECCO dataset, well-defined example data items may be of equal if not higher importance than the technical description of the dataset specification, as developers and engineers tend to use the examples as blueprints for their implementation. Thus, we equip every FHIR profile defined in the extension modules with at least one example. These examples are incorporated and issued within the implementation guides of the modules. Specifically, we aimed to provide one example for each different category of response option per profile. Thirdly, the actual implementation of the dataset definitions should be part of follow-up infrastructure projects to supply funding and resources for filling the dataset definition with actual data. For the GECCO dataset, this is ensured by follow-up projects of the German COVID-19 Research Network of University Medicine ("Netzwerk Universitätsmedizin"), such as CODEX+, which includes several implementation tasks that are actively

using the GECCO dataset items. Fourth, once the dataset definitions are implemented and leveraged in use cases, additional demands to the dataset are likely raised or issues with existing definitions are revealed. The maintenance of existing definitions (e.g., performing technical corrections or even evolving the definitions or adding new items) is, therefore, necessary and must be organized and funded. Last, successful use of the dataset definitions is also highly dependent on the degree of interoperability of the definitions in the first place. For example, several data items of the here presented extension modules were at first specified in a questionnaire-like fashion and direct implementation of these as Questionnaire resources in FHIR would limit the applicability of such data elements, especially when aiming to map these items from an electronic health records (EHR) system. In such cases, we revised the data item specification, wherever possible, in close collaboration with the medical domain experts to be able to map them to interoperable concepts, and we recommend not to use Questionnaire/QuestionnaireResponse profiles in cases where the information to be represented can be modeled using more general and therefore interoperable concepts and FHIR resources.

Conclusion

We here introduce the development process and the resulting dataset definitions of extension modules for the German Corona Consensus (GECCO) dataset, expanding the original GECCO dataset into the *immunization*, *pediatrics*, and *cardiology* domains. We have defined and implemented a process in which interdisciplinary teams of medical domain experts, medical information scientists and FHIR developers closely collaborate in an iterative, consensus-based fashion for the successful development of useful and interoperable dataset definitions. This process may serve as a blueprint for further dataset definition projects, such as further GECCO extension modules. The dataset definitions described in this work have been validated and published. Their implementation and active use are anticipated in the context of current nationwide COVID-19 research networks in Germany.

Declarations

Ethics approval and consent to participate

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

Consent for publication

Not applicable.

Availability of data and materials

The implementation guides for the three extension modules have been published on GitHub pages [18–20]. The source FHIR ShortHand (FSH) files have been published on GitHub [21–23]. Dataset descriptions can be found on ART-DECOR [24–26]

Competing interests

The authors declare that they have no competing interests.

Funding

The NAPKON ("Nationales Pandemie Kohorten Netz", German National Pandemic Cohort Network) project is funded under a scheme issued by the Network of University Medicine (Nationales Forschungsnetzwerk der Universitätsmedizin (NUM)) by the Federal Ministry of Education and Research of Germany (Bundesministerium für Bildung und Forschung (BMBF)) grant number 01KX2021.

Authors' contributions

All authors contributed to the development of the datasets. GL, TH, SB, LR, JS, AB, ST performed terminology mapping, FHIR profiling and critical review of the concept and resource mappings. TH, SB, LR defined the datasets in ART-DECOR. DH, FK, LES, FB, FE, NT, RB, AF, MD developed and compiled the list of data items for the extension modules. SR, LL and MU coordinated the project and the consensus finding process within and between working groups. JJV, CvK, ST conceived the work. GL drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank Yannick Börner for his valuable contribution to the definition of the FHIR profiles. We thank all members of the subject- and organ-specific working groups.

Abbreviations

- ART-DECOR: Advanced Requirement Tooling - Data Elements, Codes, OIDs and Rules
- ATC: Anatomical Therapeutic Chemical Classification System
- BMBF: Bundesministerium für Bildung und Forschung (German Federal Ministry of Education and Research)
- CODEX: COVID-19 Data Exchange Platform
- COVID-19: Coronavirus Disease 2019
- FHIR: Fast Healthcare Interoperability Resources
- FSH: FHIR ShortHand
- 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
- JSON: JavaScript Object Notation
- KBV: Kassenärztliche Bundesvereinigung (National Association of Statutory Health Insurance Physicians)
- LOINC: Logical Observation Identifiers Names and Codes
- NAPKON: Nationales Pandemie Kohorten Netz (National Pandemic Cohort Network)
- NUM: Netzwerk Universitätsmedizin
- MII: Medizininformatik Initiative (Medical informatics initiative)
- SARS-CoV-2: Severe Acute Respiratory Syndrome Coronavirus 2
- SUSHI: SUSHI Unshortens Short Hand Inputs
- UCUM: Unified Code for Units of Measure

References

1. Lehne M, Sass J, Essenwanger A, Schepers J, Thun S. Why digital medicine depends on interoperability. Npj Digit Med. 2019;2:1–5.

2. Sass J, Bartschke A, Lehne M, Essenwanger A, Rinaldi E, Rudolph S, 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.

3. Netzwerk Universitätsmedizin. https://www.netzwerk-universitaetsmedizin.de/. Accessed 16 Mar 2022.

4. NAPKON – Nationales Pandemie Kohorten Netz des Netzwerk Universitätsmedizin. https://napkon.de/. Accessed 16 Mar 2022.

5. NUM-COVIM study. https://covim-netzwerk.de/. Accessed 16 Mar 2022.

6. SNOMED CT. SNOMED. https://www.snomed.org/. Accessed 16 Mar 2022.

7. 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.

8. Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM). ICD-10-GM. https://www.bfarm.de/EN/Code-systems/Classifications/ICD/ICD-10-GM/_node.html. Accessed 16 Mar 2022.

9. Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM). ATC. https://www.bfarm.de/DE/Kodiersysteme/Klassifikationen/ATC/_node.html. Accessed 16 Mar 2022.

10. The Unified Code for Units of Measure. https://ucum.org/trac. Accessed 16 Mar 2022.

11. art-decor.org. https://www.art-decor.org/mediawiki/index.php/Main_Page. Accessed 16 Mar 2022.

12. Medizininformatik Initiative - SIMPLIFIER.NET. https://simplifier.net/organization/koordinationsstellemii. Accessed 15 Mar 2022.

13. Kassenärztliche Bundesvereinigung (KBV) - SIMPLIFIER.NET. https://simplifier.net/organization/kassenrztlichebundesvereinigungkbv. Accessed 15 Mar 2022.

14. FHIR Shorthand. http://hl7.org/fhir/uv/shorthand/. Accessed 25 Apr 2022.

15. SUSHI Unshortens Short Hand Inputs. 2022. https://github.com/FHIR/sushi. Accessed 25 Apr 2022.

16. Lichtner G. FHIR Shorthand Validator. 2021. https://github.com/glichtner/fsh-validator. Accessed 15 Mar 2022.

17. napkon-module-template. 2022. https://github.com/BIH-CEI/napkon-module-template. Accessed 15 Mar 2022.

18. NAPKON Cardiology Module Implementation Guide. https://bih-cei.github.io/napkon-cardiology/. Accessed 15 Mar 2022.

19. NAPKON Pediatrics Module Implementation Guide. https://bih-cei.github.io/napkon-pediatrics/. Accessed 15 Mar 2022.

20. NAPKON Vaccination Module Implementation Guide. https://bih-cei.github.io/napkon-vaccination/. Accessed 15 Mar 2022.

21. NAPKON Cardiology Module. 2022. https://github.com/BIH-CEI/napkon-cardiology. Accessed 15 Mar 2022.

22. NAPKON Pediatrics Module. 2022. https://github.com/BIH-CEI/napkon-pediatrics. Accessed 15 Mar 2022.

23. NAPKON Vaccination Module. 2022. https://github.com/BIH-CEI/napkon-vaccination. Accessed 15 Mar 2022.

24. ART-DECOR: NAPKON Cardiology Module. https://art-decor.org/art-decor/decor-datasets--covid19f-?id=2.16.840.1.113883.3.1937.777.53.1.2&effectiveDate=2020-08-12T00%3A00%3A00&conceptId=2.16.840.1.113883.3.1937.777.53.2.250&conceptEffectiveDate=2021-02-16T13%3A25%3A43&language=en-US. Accessed 16 Mar 2022.

25. ART-DECOR: NAPKON Pediatrics Module. https://art-decor.org/art-decor/decor-datasets--covid19f-?id=2.16.840.1.113883.3.1937.777.53.1.2&effectiveDate=2020-08-12T00%3A00%3A00&conceptId=2.16.840.1.113883.3.1937.777.53.2.453&conceptEffectiveDate=2021-08-25T12%3A45%3A26&language=en-US. Accessed 16 Mar 2022.

26. ART-DECOR: NAPKON Vaccination Module. https://art-decor.org/art-decor/decor-datasets--covid19f-?id=2.16.840.1.113883.3.1937.777.53.1.2&effectiveDate=2020-08-12T00%3A00%3A00&conceptId=2.16.840.1.113883.3.1937.777.53.2.453&conceptEffectiveDate=2021-08-25T12%3A45%3A26&language=en-US. Accessed 16 Mar 2022.