Accepted Manuscript online: 2023-01-03 Article published online: 2023-01-30

Consistency as a Data Quality Measure for German Corona Consensus Items Mapped from National Pandemic Cohort Network Data Collections

Khalid O. Yusuf[1,*] Olga Miljukov[2,*] Anne Schoneberg[1] Sabine Hanß[1] Martin Wiesenfeldt[1] Melanie Stecher[3,4] Lazar Mitrov[5] Sina Marie Hopff[5] Sarah Steinbrecher[6] Florian Kurth[6] Thomas Bahmer[7,8] Stefan Schreiber[7] Daniel Pape[7] Anna-Lena Hofmann[2] Mirjam Kohls[2] Stefan Störk[9] Hans Christian Stubbe[10] Johannes J. Tebbe[11] Johannes C. Hellmuth[12,13] Johanna Erber[14] Lilian Krist[15] Siegbert Rieg[16] Lisa Pilgram[17,18] Jörg J. Vehreschild[3,4,18] Jens-Peter Reese[2] Dagmar Krefting[1,19]

[1] Department of Medical Informatics, University Medical Center Göttingen, Göttingen, Germany
[2] Institute for Clinical Epidemiology and Biometry (ICE-B), University of Würzburg, Würzburg, Germany
[3] Department I for Internal Medicine, University Hospital Cologne, Cologne, Germany
[4] German Centre for Infection Research, Partner Site Bonn-Cologne, Cologne, Germany
[5] Department I of Internal Medicine, Faculty of Medicine and University Hospital Cologne, Center for Integrated Oncology Aachen Bonn Cologne Duesseldorf, University of Cologne, Cologne, Germany
[6] Department of Infectious Diseases and Respiratory Medicine, Charité-Universitaetsmedizin Berlin, Berlin, Germany
[7] Internal Medicine Department I, University Medical Center Schleswig-Holstein Campus Kiel, Kiel, Germany
[8] Airway Research Center North (ARCN), German Center for Lung Research (DZL), Wöhrendamm Großhansdorf, Germany
[9] Department Clinical Research & Epidemiology, University Hospital Würzburg, Comprehensive Heart Failure Center, and Department Internal Medicine I, Würzburg, Germany
[10] Department of Medicine II, University Hospital, LMU Munich, Munich, Germany
[11] Department of Gastroenterology and Infectious Diseases, University Medical Center East Westphalia-Lippe, Klinikum Lippe, Lemgo, Germany
[12] Department of Medicine III, University Hospital, LMU Munich, Munich, Germany
[13] COVID-19 Registry of the LMU Munich (CORKUM), University Hospital, LMU Munich, Munich, Germany
[14] Department II of Internal Medicine, Technical University of Munich, School of Medicine, Germany
[15] Institute of Social Medicine, Epidemiology and Health Economics, Charité-Universitätsmedizin Berlin, Berlin, Germany
[16] Department of Medicine II, Division of Infectious Diseases, Medical Centre – University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany
[17] Department II of Internal Medicine, Hematology/Oncology, Goethe University, Frankfurt, Frankfurt am Main, Germany
[18] Department of Nephrology and Medical Intensive Care, Charité - Universitaetsmedizin Berlin, Berlin, Germany
[19] Campus Institute Data Science (CIDAS), Georg-August-University, Göttingen, Germany

Address for correspondence Khalid O. Yusuf, MSc., Department of Medical Informatics, University Medical Center Göttingen, Göttingen, Germany (e-mail: olusolakhalid.yusuf@med.uni-goettingen.de).

Methods Inf Med 2023;62:e47–e56.

Abstract

Background As a national effort to better understand the current pandemic, three cohorts collect sociodemographic and clinical data from coronavirus disease 2019 (COVID-19) patients from different target populations within the German National Pandemic Cohort Network (NAPKON). Furthermore, the German Corona Consensus Dataset (GECCO) was introduced as a harmonized basic information model for COVID-19 patients in clinical routine. To compare the cohort data with other GECCO-based

Keywords
- data quality
- COVID-19
- consistency
- metadata

*These authors contributed equally.

received
July 18, 2022
accepted after revision
October 31, 2022
accepted manuscript online
January 3, 2023
article published online
January 30, 2023

DOI https://doi.org/10.1055/a-2006-1086.
ISSN 0026-1270.

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

studies, data items are mapped to GECCO. As mapping from one information model to another is complex, an additional consistency evaluation of the mapped items is recommended to detect possible mapping issues or source data inconsistencies.

Objectives The goal of this work is to assure high consistency of research data mapped to the GECCO data model. In particular, it aims at identifying contradictions within interdependent GECCO data items of the German national COVID-19 cohorts to allow investigation of possible reasons for identified contradictions. We furthermore aim at enabling other researchers to easily perform data quality evaluation on GECCO-based datasets and adapt to similar data models.

Methods All suitable data items from each of the three NAPKON cohorts are mapped to the GECCO items. A consistency assessment tool (dqGecco) is implemented, following the design of an existing quality assessment framework, retaining their-defined consistency taxonomies, including logical and empirical contradictions. Results of the assessment are verified independently on the primary data source.

Results Our consistency assessment tool helped in correcting the mapping procedure and reveals remaining contradictory value combinations within COVID-19 symptoms, vital signs, and COVID-19 severity. Consistency rates differ between the different indicators and cohorts ranging from 95.84% up to 100%.

Conclusion An efficient and portable tool capable of discovering inconsistencies in the COVID-19 domain has been developed and applied to three different cohorts. As the GECCO dataset is employed in different platforms and studies, the tool can be directly applied there or adapted to similar information models.

Introduction

The consistency assessment of health research data could be necessitated by a number of usage scenarios including but not limited to: (1) assembling data from different sources; (2) extraction, transformation, and loading of source data to a secondary repository; and (3) mapping source data to an interoperable or uniform standard.[1–3] Throughout the article, we follow the definition of consistency dimension according to Schmidt et al: "consistency comprises indicators that use Boolean type checks to identify inadmissible, impossible, or uncertain data values or combinations of data values."[4] There are different varieties of consistency checks that can be evaluated within a dataset depending on the requirements that are specific to the domain where the data emanate. For instance, a dataset can be evaluated for its adherence to some predefined protocols (value or format violations). Data values within data items can be evaluated independently on the basis of their agreement with certain gold standards.[2–4] A subdomain of consistency are contradictions. Contradictions comprise indicators within dataset that address the possibility or certainty of combinations of interdependent data values within the dataset. Contradictions form a fundamental ground upon which a dataset might be rejected or subjected to a mandatory correction before declaring it fit for health research.[1] This is due to the fact that a dataset flawed with contradictions cannot produce results that can be trusted.[3] An example of contradiction is the case where there is an indication of fever as a coronavirus disease 2019 (COVID-19) symptom at baseline, but normal body temperature was reported at the same time-point.

The National Pandemic Cohort Network (NAPKON) is a joint project of the Network University Medicine (NUM) aimed at harmonizing the collection and use of COVID-19 data.[5] The three NAPKON cohorts collect sociodemographic and clinical data from COVID-19 patients from different target populations and across sectors. The data are collected within the acute course and longitudinally up to 12 months after initial diagnosis. Additionally, to provide COVID-19 researchers with a uniform dataset that uses international terminologies and interoperable health IT standards, the German Corona Consensus Dataset (GECCO) was introduced.[6] Hence, data items from the individual cohorts must be mapped to the GECCO items. The heterogeneity of data sources as well as the mapping process to the GECCO dataset presents the ground for the consistency assessment presented here. This is supported by the fact that recent research on the quality of COVID-19 datasets has also reported inconsistencies in COVID-19 surveillance data.[7]

Objectives

In this work, the goal is to develop a consistency assessment package that can be used in assessing the consistency of mapped GECCO items from the NAPKON cohorts. Of greater interest is the identification of contradictions within interdependent GECCO items to allow the investigation of the cause of such contradictions. We furthermore aim at enabling other researchers to easily perform data-quality

evaluation on GECCO-based datasets and adapt the framework to similar data models.

Methods

NAPKON Cohort Studies

NAPKON as currently constituted comprises three cohorts encompassing the intersectoral platform (SÜP), the population-based platform (POP), and the high-resolution platform (HAP). They all collect data from COVID-19 patients, but from different populations with varying sets of data items and acquisition times. The POP recruits patients with survived COVID-19 infections in specific regions in Germany and captures the course of the disease retrospectively as well as prospective follow-ups. The SÜP collects prospectively comprehensive information on clinically ill patients throughout the university clinics in Germany and other in- and out-patient health care providers. The HAP accomplishes deep phenotyping of clinically ill COVID-19 patients in selected university clinics. To aid a central management of the collected clinical, imaging, and biosample data, the NAPKON partnered with the German Centre for Cardiovascular Research (DZHK) to utilize its existing infrastructure (secuTrial) and expertise for capturing, storing, and retrieving data. As the mapping to GECCO has been anticipated already in the electronic data capture (EDC) design of SÜP and POP, most items to be mapped have a respective prefix "gec_". As of June 30, 2022, more than 10,000 unique items have been implemented in the EDC systems of the cohorts and over 5,500 patients have been recruited.[5]

NAPKON Quality Assurance

The NAPKON cohorts have established several measures for quality assurance of collected data. One of those measures comes into effect directly at the level of data entry into the EDC systems, where numerous rules have been implemented. This enabled implausible values to be automatically detected in real-time and to instruct the user to correct or confirm the entry.

Furthermore,

- The NAPKON project has established the Epidemiological Core Unit (ECU) to investigate and describe the data quality of the three NAPKON cohorts, among other methodological tasks.
- At the study center level, there is an approval procedure by which the study center confirms that the data entered had been reviewed and found to be correct (review level A). At the time of writing, records of 5,359 patients have passed review level A and have been included into the GECCO mapping and subsequent quality analysis procedure.
- The entered data are also continuously monitored by the individual study management constituted for each of the three cohorts. Although the cohorts use the same EDC system (secuTrial), data collection and monitoring are handled independently according to the schemas of each cohort. Implausible values, which can be more complex at this level and get undetected during data entry, are queried to and corrected by the study centers in a query process. After completion of monitoring, the corresponding data sections are given a further, final quality status: review level B.
- In a parallel process, the ECU frequently performs a centralized external independent quality control of the input data by applying statistical data quality measures using selected quality indicators from data quality dimensions such as completeness and consistency at individual item level. Analysis results are made available in regular reports.

Another measure implemented by the interaction core unit is the project-internal billing process of study activities and the amount of entered data. The study centers receive disbursements if the data were entered timely and passed the quality checks described above. This has an additional direct influence on the data quality and represents a further monitoring instance of the data quality.

GECCO and Mapping from NAPKON

The GECCO dataset has been designed as a minimal dataset that should be reported for all COVID-19 patients in the different studies throughout the NUM to allow for data federation and big data analytics.[8] The GECCO information model encompasses 13 so-called medical concepts with about 83 relevant data items that use international terminologies and the Fast Healthcare Interoperability Resources standard.[6] A medical concept is a structured way of categorizing clinical data items such that each set of related data items are grouped within a nomenclature that best describes what they represent. All suitable data items for the 83 GECCO items from each of the 3 NAPKON cohorts captured in secuTrial were mapped according to the 13 concepts presented in GECCO using an R-Script.[9] While one-to-one mapping was suited for data items with only baseline records, items with multiple timepoints and/or multiple variants were mapped in a many-to-many relationship. GECCO and NAPKON experts have also been consulted to resolve unclear situations. Consequently, we achieved a one-to-one mapping mostly for the POP and a larger part of SÜP. However, as all cohorts define at least three follow-up examinations, some items were mapped in a many-to-many relationship to four different timepoints in GECCO. Items with two variants (e.g., chronic and acute heart failure) were mapped to two values in GECCO as two instances of the same item. The mapping of the HAP items was similar to SÜP. A difficulty in the mapping was, in particular, that items such as specific therapy methods may refer to any timepoint and not just to the examination moment. We found that such information was reported several times in the cohorts. Such items were reduced to one or a set of entries according to a reduction rule. As an illustration, the dialysis item in GECCO refers to the question about the application of dialysis in the whole treatment phase. Therefore, once there was a "Yes" in at least one of the entries in the EDC, "Yes" was set for the respective GECCO item. In the absence of a "Yes" in the whole of the treatment phase, "No" was the next value to be

selected and so on. For other items with explicit nominal values (e.g., respiratory-therapy-type with predefined values "invasive" and "noninvasive"), we retained all possible unique values across the treatment phase. We would like to emphasize that intermediate results from the consistency assessment were used to subsequently adjust the mapping in an iterative process. The results reported in the Results sections refer to the remaining contradictions that could not be resolved through corrected data mapping.

Review of Consistency Assessment Frameworks

The consistency dimension has been well researched in the past and is part of the recent harmonized data quality frameworks.[1–4,10] In the past, several studies have already singled out consistency as a data quality dimension for the purpose of removing data errors in a database.[2] Embury et al designed a method based on domain-specific consistency rules for the reconciliation of contradictions resulting from data integration.[11] Also, Mezzanzanica et al[2] introduced a formal approach to the identification of inconsistencies in database by mapping its records to a set of events and then, mapping the attributes of these records to the defined events thereby, enhancing consistency checks of the records against specified protocols.

In recent times, consistency like other data quality dimensions has witnessed a significant improvement through harmonization.[1,4,10] For instance, Nonnemacher et al[1] distinguished between safe and severe contradictions. The severity of the contradictions identified in a dataset would determine the trustworthiness of the dataset.[1,3] In some instances, certain contradictions would require mandatory corrections, and the acceptability of the dataset would depend on effecting those corrections. On the other hand, some contradictions would be revealed through analysis but are indeed possible. A typical example is the case of male breast cancer.[1]

In the Healthcare Data Quality Framework (HDQF) by Johnson et al,[3] "Domain Consistency" was introduced to measure the rate at which a dataset fulfills domain-specific rules. This approach identifies the relevant concepts of a domain of interest within a dataset and evaluates the plausibility of the data values based on the semantics of the domain. The consistency rate is measured as a ratio of satisfactory constraints within each item to the total count of data items in a dataset. This framework was developed into a Python software program. As already mentioned in the Introduction, Schmidt et al,[4] upon which our implementation in this study is based, recognizes two sub-domains within the consistency dimension as follows: (1) *range and value violations* and (2) *contradictions*. While in the former, data values are tested for compliance to data types and format restrictions in a pattern similar to the Domain constraints by Johnson et al,[3] in the latter, contradictory values are evaluated on the basis of their interdependent relationships among data items. The framework further distinguished between *logical* and *empirical* contradictions. While domain knowledge is required to establish empirically contradictory value combinations, common knowledge can help determine value combinations that are logically implausible.

The consistency framework by Schmidt et al[4] is available as an R package (*con_contradictions)* and capable of assessing contradictions that exist between two interdependent data items using built-in generic logical rules. While the implementation of *con_contradictions* module in the *DataquieR* package (a comprehensive data quality assessment workflow) is open source thereby, giving room for possible extensions by other scientists, the HDQF[3] is not open source and its Domain consistency framework is not as robust as the *con_contradictions*. In a recent study, a data quality assessment workflow specific to the Biobank domain introduced another dimension of interdependency among data items where three-way and four-way multi-item contradictions were demonstrated.[12] This framework was implemented as a SmartR plugin for the open source analysis platform tranSMART.[13] In the current work, we have adapted our design to the existing *con_contradictions* framework by Schmidt et al[4] to ease the use of the tool by data managers and transfer sites.

Consistency in GECCO

To support quality assurance under the GECCO initiative, we evaluated the consistency of possible dependencies that exist within GECCO data items with a focus on the subdomain of contradictions. The interdependencies examined in the present work according to the design of GECCO[14] are: (1) history of Diabetes compared with Insulin medication; (2) fever reported as a symptom compared against the body temperature as a vital sign at the same time-point; and (3) COVID-19 severity at the point of diagnosis compared with the relevant indicators (joined with a Boolean "OR") that cut across several concepts in the GECCO dataset. Examples are increased levels of specific laboratory values and oxygen supplement in the therapy. The full list of considered items is reported in ►Table 1.

Consistency Assessment Using dqGecco R-package

We implemented the *dqGecco* consistency assessment package based on the design of the existing *con_contradictions* module by Schmidt et al.[4] The *dqGecco* R package was built based on the following modifications to the *con_contradictions* module to meet the task- and domain-related requirements in the GECCO use case: (1) implement suitable logical rules to determine the consistency of value combinations within intra- and inter-concept-dependent items; (2) render a detailed report to trace the identified findings back to the data source through their unique keys; and (3) visualize the identified findings within the relevant data items to aid comparison with other consistent values. The defined taxonomies within the *con_contradictions* framework including logical and empirical contradiction metrics were retained. The interdependent items within the mapped GECCO dataset described in section "Consistency in GECCO" were assessed using logical rules implemented in the *dqGecco* tool. The contradiction detection performance of the tool was

Table 1 Classes of COVID-19 infection severity during admission with their indicators (formulated from Art-Decor)[14]

Covid-19 severity	GECCO-concept	Indicators
Uncomplicated	Symptoms	Asymptomatic OR upper respiratory tract infections OR nausea OR emesis OR diarrhea OR fever
Complicated	Anamnesis	Heart failure with pulmonary edema OR cardiac arrhythmia
	Vital signs	PaO_2 <70 mmHg OR SO_2 at room air <90 %
	Therapy	Need for new oxygen supplement
	Laboratory	AST or ALT > 5 × ULN
Critical	Anamnesis	Life-threatening cardiac arrhythmia OR acute renal failure
	Vital signs	qSOFA ≥ 2
	Therapy	Unplanned mechanical ventilation, dialysis
	Laboratory	INR > 3.5

Abbreviations: ALT, alanine aminotransferase; AST, aspartate aminotransferase; INR, international normalized ratio; qSOFA, quick sequential organ failure assessment; ULN, upper limit of normal.
Note: "OR" signifies a Boolean OR.

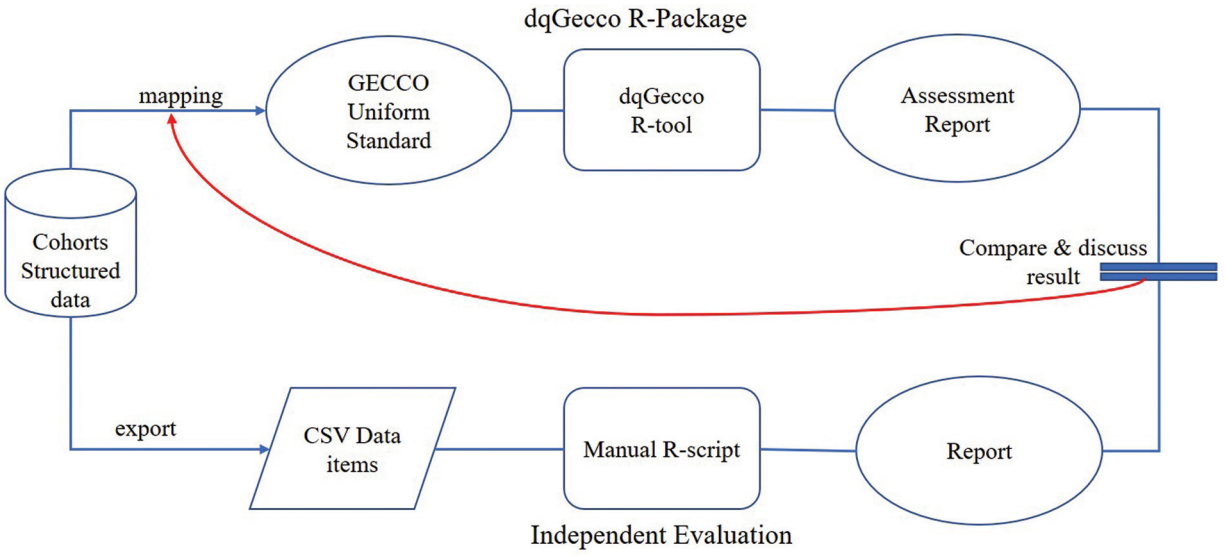

Fig. 1 Workflow of the consistency assessment using dqGecco R-package and independent evaluation of the assessment results.

evaluated independently as described in ►Fig. 1 by an ECU member in the following way: selected rules were implemented in R without any insight into the source code nor usage of the *dqGecco* tool and executed on directly as csv-exported EDC data. The results from both assessments are compared to reproduce the set of identified contradictions. The positive detection rates were compared and concordance was assessed. Any disagreement was discussed iteratively and modifications of the rules or the mapping were made or further clarification by the cohort in question was sought until resolved.

NAPKON Use and Access Approval

We made a formal proposal to the NAPKON Use and Access Committee, which was approved. The presented work is covered by the ethics committee votes for the three cohorts as a quality assurance measure.

Results

Consistency of Fever Diagnosis and Body Temperature

The body temperature (T_b) is one of the basic vital signs, in particular elevated T_b, identified as fever.[15,16] Fever is one of the indicators to look out for in clinical diagnosis of COVID-19 and is reported as a value-set ("no," "yes," "not set") within the GECCO diagnosis concept.[17] T_b is also one of the vital parameters captured in the respective GECCO concept. Depending on the method of measurement, the T_b that is considered as the threshold for fever differs. The T_b indicators for fever for patients according to the EDC design of the cohorts was $T_b \geq 38.3$°C for the rectal measurement and $T_b \geq 37.8$°C for other methods. For T_b consistency evaluation in this study, we used those thresholds to define the range for no indication of fever as T_b <37.8°C, the range for definite fever as $T_b \geq 38.3$°C. For T_b in between these thresholds, both

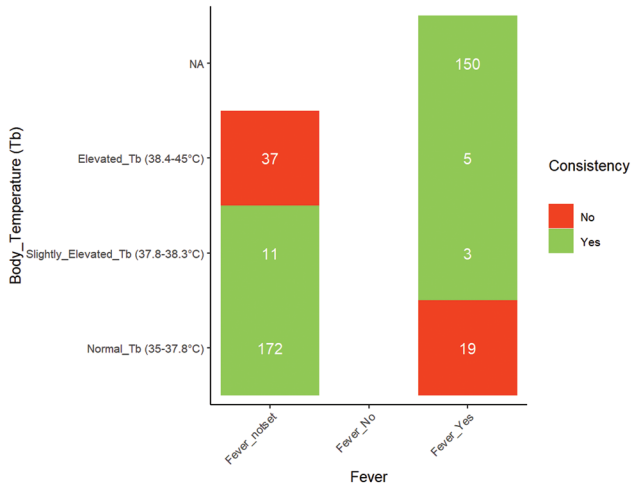

Fig. 2 Comparison of body temperature (T_b) against fever in the SÜP cohort. T_b values not captured at the same time-point as fever were excluded.

presence and absence of fever are consistent. While T_b stores continuous values, which are measured at different time-points (baseline was considered here), fever has nominal values which was captured only during screening of the study participants. For the evaluation of synchronicity, the date of vital sign assessment must be the same as the screening date in case fever is not indicated, or must be identical with the fever onset date in case of indication of fever. ►**Fig. 2** shows the distribution of combinations within the evaluated data records. The sets of contradictions are indicated in red in the chart. The evaluation of the EDC system to identify the cause of the contradictions showed that there were no association rules that linked the T_b input to fever indication. An only partly restricted field was provided for the T_b input. The field was only controlled against the entry of negative values and as a follow-up, a confirmation of values that exceeds a set range (20–45°C) was required. So, while consistency in terms of range violations has been established as automatic data quality check, contradictions to the fever indication have not been considered in the EDC layout. We would like to emphasize that due to the required synchroneity of fever appearance and vital sign assessment, many data records could not be evaluated, because they missed the reference time. It is also important to note that the methods used in obtaining the T_b of individual patients were not specifically captured.

Consistency of COVID-19 Severity and Its Indicators

Classifying COVID-19 cases according to the severity of disease indicators is an important task that requires due diligence. Reason for this is that researchers who request COVID-19 data might indicate inclusion criteria for certain severity classes and if there are contradictions between the severity and its respective indicators, there is a risk of using data with wrong severity classes. Researchers have identified the most significant indicators that can be used to stratify COVID-19 cases into different severity groups. However, the consistency of the severity classification against respective indicators has not been reported. Therefore, we examined all the categories of the COVID-19 severity in comparison with their respective indicators which were captured at the same time-point where a positive COVID-19 diagnosis was established. We defined consistency rules based on the rules formulated in ►**Table 1** derived from the EDC and GECCO,[14] then visualized the value combinations to detect contradictions. While COVID-19 infections with mild to moderate symptoms are expected to be categorized as uncomplicated, those who in addition to known symptoms required oxygen supplements or suffered from other risk factors including heart failure and cardiac arrhythmia qualify for the complicated category. An extreme layer is the critical phase which requires the patient to undergo invasive or noninvasive mechanical ventilation, or dialysis after suffering renal failure, or an increased international normalized ratio above 3.5 to mention a few. As depicted in ►**Table 1**, there are different combinations that could lead to contradictions. In ►**Table 2**, we present a scenario where COVID-19 cases were classified as uncomplicated, while ventilation indicated a complicated or even critical. The ventilation method can further contradict the assigned severity class. We would like to emphasize that the full evaluation rule matrix encompasses several items from different concepts, resulting in a multidimensional comparison. From the assessment of the EDC system, it was observed that the fields designed to capture the values of the severity classes were disjointed from the fields provided for the entry of the different severity indicators. The field for the severity classes was only supported with graphics of their respective indicators without any association rules to automatically cross-validate the interdependent value-sets at the point of data capture. The values of the indicators were captured independently across several tables, which made them prone to contradictions with the dependent item.

Table 2 Example of contradictory data value combinations, here for the therapy concept

Severity class	Ventilation	Ventilation class	Consistent severity class
Uncomplicated	yes	%	Complicated
Complicated	%	Noninvasive-mechanical	Critical
	%	Invasive-mechanical	Critical
Critical	%	Conventional oxygen	Complicated
	%	High-flow oxygen	Complicated

Note: % indicates that the item is not relevant for the evaluation.

Table 3 Summary of consistency assessment across 3 cohorts (POP, SÜP, HAP)

	Features	Passed items	Failed items	Pass rate (%)
POP				
	Diabetes_vs_Insulin	2,541	0	100
	Fever_vs_BodyTemp	2,532	9	99.65
	Covid_Severity_vs_Indicators	2,539	2	99.92
SÜP				
	Diabetes_vs_Insulin	2,214	0	100
	Fever_vs_BodyTemp	2,158	56	97.47
	Covid_Severity_vs_Indicators	2,122	92	95.84
HAP				
	Diabetes_vs_Insulin	604	0	100
	Fever_vs_BodyTemp	599	5	99.17
	Covid_Severity_vs_Indicators	582	22	96.36

Abbreviations: HAP, high-resolution platform; POP, population-based platform; SÜP, intersectoral platform.

Summary Report across the Three Cohorts

We applied our tool with the same set of rules to the same set of interdependent items from the three cohorts, which produced a comparable summary report described in ►Table 3. The metric for the estimation of the pass rate was the ratio of the absolute frequency of passed values to the total number of data records.[1,3,4] From these results, the consistency rates within the POP cohort were very high with a minimal consistency rate of 99.92%. The scores for the interdependent features within the SÜP cohort were also high with a minimal consistency rate of 95.84%. The features within the HAP cohort were similarly consistent with a minimal consistency rate of 96.32%. No contradictions were found in the indication of Diabetes and Insulin medication. While it can be inferred that the consistency rates across the cohorts in our assessment were relatively high considering that we estimated the rates with respect to the total sample collection of each cohort, the impact of the contradictions on individual research analysis will depend on the research question, type of analysis as well as the corresponding sample sizes since researchers usually specify sample size when requesting data from the cohorts. Our assessment was initially blinded to the existence of system-enforced measures designed to control contradictions at the point of data entry; therefore, we included the selected interdependent items in our logical rules. A follow-up evaluation of the EDC systems to determine the cause of the identified contradictions and the reason for the absence of contradictions in other features revealed the existence of measures that were put in place to control contradictory value combinations. In the case of Diabetes and Insulin medication, the EDC system ensured the insulin medication field was only displayed to the users when the value of the Diabetes was set to "Yes." This automatically voided the existence of "No:Yes" contradictions as well as "missing: Yes" contradictions in the Diabetes:Insulin value combinations. Such measures were missing for other sets of contradictions discovered in our assessment. Also, we observed that the interdependency dimension in diabetes and insulin medication checks can already be handled by one of the logical rules (e.g. missing(A) & B is present) defined in the tool of Schmidt et al.[4] Hence, our decision to reflect more on new dimensions not yet reported as obtained in the other results, i.e., fever against T_b input where all possible T_b value ranges were compared to valid fever outcomes (Fever_Yes, Fever_No) and secondly, a set of COVID-19 severity indicators compared to a dependent item (COVID-19 severity class) through a loop of each set of corresponding indicators.

Results of the Independent Evaluation

The final rules for detecting contradictions were applied to the corresponding original EDC items for:

- Fever_vs_BodyTemp.
- Covid_Severity_vs_Indicators of the POP cohort.

An important finding in the iterative process was that—as mentioned in the Methods section—the reference time for a reported information was not always well defined, for example, the presence of the symptom fever was asked during the acute phase retrospectively as well as whether the symptom persists (every now and then) until present without a clear relationship to the actual body temperature measurement reported in vital signs. In addition, it became apparent that the classification of disease severity at the time of diagnosis can only be assessed if the corresponding indicators have a documented starting date. In the final iteration, all disagreements were resolved, and the same data records failed. The concordance rate defined as the number of data records that are concordant over the total number of items assessed was therefore 100% for those features and in total (►Table 4). Relying on the result of this evaluation, it can be inferred that the mapping process

Table 4 Concordance per feature and in total for contradictions found by the *dqGecco* tool (in ►Table 3) in the research dataset and by the evaluation in the source data of POP

Feature	Pairs of failed items from dqGecco and verification (*n*)	Concordant pairs (*n*)	Concordance rate
Fever_vs_BodyTem	9	9	100 %
Covid_Severity_vs_Indicators	2	2	100 %
Total	11	11	100 %

Abbreviations: POP, population-based platform.

did not influence the identified contradictions, rather, these were rare contradictions that could not be resolved after correcting the mapping process. Consequently, the precautionary measures within the EDC system have to be fortified to prevent such contradictions at the point of data entry for subsequent cases. For already concluded cases, special efforts would be dedicated to ensuring that the identified contradictions in this work are corrected during review level B.

Discussion

The logical rules implemented in this study revealed another interdependency dimension in the consistency framework which differs from the three- and four-dimensional checks from our own previous experience in the Biobank domain[12] or the item-wise (one-dimensional) checks of Schmidt et al and Johnson et al.[3,4] In one instance, this interdependency dimension requires the comparison of multiple sets of data items (severity indicators) and their dependent items (e.g. each class of COVID-19 severity has different sets of indicators as demonstrated in the section "Consistency of COVID-19 Severity and Its Indicators"). This is an indication that consistency assessments will be as dynamic as the task obtainable in different domains. The definition of consistency rules will rely largely on the amount of knowledge that can be derived from complex interdependent items within a domain. From the experience in this study and previous studies, it is assumed that contradiction as a data quality indicator will benefit more from the assembly of several interdependency dimensions that could be harnessed from different domains and will therefore require the community to report the discovery of unique dimensions applicable in different use cases.

Since our tool implementation is modular following the Schmidt et al[4] design, other use cases that find the consistency rules adaptable will only be required to create metadata that fits the arguments defined in the *dqGecco* package. In this study, we have not only contributed to the possible extension of the *con_contradictions* module by Schmidt et al[4] through the implementation of more complex rules supported with an enhanced consistency visualization revealing all interdependent datapoints, we have equally applied the framework in a multi-disciplinary domain (i.e., COVID-19). This tool will be used by the NAPKON data transfer site to automatically detect inconsistent items within dataset before data export. While Costa-Santos et al[7] only reported inconsistencies in COVID-19 surveillance data, we go beyond that to carry out consistency checks that cut across different research facets of COVID-19 for the benefit of COVID-19 researchers.

To demonstrate the efficacy of our assessment tool, we carried out an independent evaluation (without the *dqGecco* package) to reproduce some of the results directly from the database. This approach helps achieve two goals: (1) ascertain that the contradictions indeed emanate from the data source and are not introduced as a result of the mapping process or preprocessing of the data before handing over to researchers and (2) build confidence in the use of the sanity-checked tool with high concordance rates. It is not a coincidence that across all cohorts, the diabetes_vs_insulin feature returned no failed items. This is a result of the EDC consistency rule which displays the insulin medication field only if the presence of diabetes is confirmed in the anamnesis concept. However, more complex interdependent relationships among data items have to be envisaged and incorporated into the EDC design to prevent cases of false-positive and false-negative value combinations demonstrated in this study. In general terms, it is crucial to make sure the mapping from the source database to the GECCO items is unambiguous, and when in doubt, clarifications should be sought from suitable study personnel. More so, to assess the performance of the tool more thoroughly, e.g., a gold-standard data subset judged by clinical experts or simulated contradictions could be used for more robust performance metrics. But this is out of scope in this work and does not compromise the conclusion. In the future, we intend to extend this tool to support the consistency assessment of real-world dataset from a similar domain that operates interoperable health information technology standard.

Conclusion

We demonstrated the efficacy and portability of our methodology and tool with the discovery of inconsistencies in the COVID-19 domain by applying the same tool with the same set of rules to three different cohorts. As the GECCO data set is also employed in different platforms and studies, the tool can be directly applied there or adapted to similar data models. While the EDC system already implemented some consistency rules which for example resulted in contradiction-free diabetes and insulin medication items in the anamnesis concept across all cohorts, it is highly encouraged that a holistic interdependency relationship among all items of interest is established during study design and suitable consistency rules incorporated in the EDC system in order to prevent the nature of contradictions identified in this work.

Availability of Materials and Data

The GECCO83 dataset used for the study can be accessed through the normal use and access procedure of the NAPKON.[1] Also, the source code of the implementation is available in the gitlab repository of the project.[2]

Funding

The study was carried out using the clinical-scientific infrastructure and data of NUKLEUS, NAPKON, and CODEX of the Network University Medicine (NUM, grant number 01KX2121), with support from the German Center for Cardiovascular Research (DZHK, grant number 81Z0300108) both funded by the Federal Ministry of Education and Research (BMBF) .

[1] NAPKON-Proskive: https://proskive.napkon.de/

[2] dqGecco Project: https://gitlab.gwdg.de/medinfpub/dqgecco.git

Conflict of Interest
None declared.

Acknowledgements
We gratefully thank the three best recruiting NAPKON sites of each platform that contributed data to this analysis. The representatives of these NAPKON sites are (alphabetical order): Charité-Universitaetsmedizin Berlin, Berlin (Fricke J, Keil T, Kretzler L, Krist L, Schmidt S, Steinbeis F, Treue D, Triller P, Witzenrath M, Zoller T), Charité - Universitaetsmedizin Berlin, Berlin (Helbig T, Hummel M, Lippert L, Mittermaier M, Müller-Plathe M, Rönnefarth M, Sander LE, Steinbeis F, Steinbrecher S, Zvorc S), LMU Clinic of the University of Munich, Munich (Frank S, Hellmuth JC, Huber M, Kaeaeb S, Keppler OT, Khatamzas E, Mandel C, Mueller S, Muenchhoff M, Reeh L, Scherer C, Stubbe H, von Bergwelt M, Weiß L, Zwißler B), University Hospital Schleswig-Holstein, Kiel (Bahmer T, Enderle J, Hermes A, Krawczak M, Lehmann I, Lieb W, Maetzler C, Reinke L, Schreiber S, Tittman L), University Hospital Technical University Munich, Munich (Barkey W, Erber J, Fricke L, Lieb J, Michler T, Mueller L, Schneider J, Spinner C, Voit F, Winter C), University Hospital Wuerzburg, Wuerzburg (Frantz S, Haeusler KG, Hein G, Horn A, Isberner N, Jahns R, Kohls M, Morbach C, Stoerk S, Weissbrich B), University Medical Center East Westphalia-Lippe, Bielefeld (Alsaad K, Berger B, Hamelmann E, Heidenreich H, Hornberg C, Kulamadayil-Heidenreich NSA, Maasjosthusmann P, Muna A, Olariu C, Ruprecht B, Schmidt J, Stellbrink C, Tebbe J), University Medical Center Freiburg, Freiburg (August D, Barrera M, Goetz V, Imhof A, Koch S, Nieters A, Peyerl-Hoffmann G, Rieg SR).

We gratefully thank all participating NAPKON infrastructures that contributed to this analysis. The representatives of these NAPKON infrastructures are (alphabetical order): Charité - Universitaetsmedizin Berlin, Berlin (Balzuweit B, Berger S, Krannich A, Kurth F, Lienau J, Lorbeer R, Pley C, Schaller J, Schmidt S, Thibeault C, Witzenrath M, Zoller T), University Hospital Cologne and University Hospital Frankfurt, Cologne and Frankfurt (Vehreschild J), University Hospital Cologne, Cologne (Brechtel M, Fuhrmann S, Hopff SM, Jakob CEM, Lee C, Mitrov L, Nunes de Miranda S, Nunnendorf M, Sauer G, Schmidt Ibanez C, Seibel K, Stecher M), University Hospital Frankfurt, Frankfurt (Appel K, Geisler R, Hagen M, Scherer M, Schneider J), University of Wuerzburg, Wuerzburg (Bauer C, Fiessler C, Goester M, Grau A, Hellmuth C, Heuschmann P, Hofmann AL, Hummel M, Jiru-Hillmann S, Kammerer K, Kohls M, Krawczak M, Miljukov O, Pley C, Reese JP, Ungethuem K), University Medicine Greifswald, Greifswald (Bahls T, Hoffmann W, Nauck M, Schäfer C, Schattschneider M, Stahl D, Valtentin H), University Medicine Goettingen, Goettingen (Chaplinskaya I, Hanß S, Krefting D, Pape C), German Center for Cardiovascular Diseases (DZHK), Berlin (Hoffmann J), Helmholtz Center Munich, Munich (Kraus M).

We gratefully thank the NAPKON Steering Committee: University Hospital Giessen and Marburg, Giessen (Herold S), University of Wuerzburg, Wuerzburg (Heuschmann P), Charité - Universitaetsmedizin Berlin, Berlin (Heyder R), University Medicine Greifswald, Greifswald (Hoffmann W), Hannover Unified Biobank, Hannover Medical School, Hannover (Illig T), Robert Koch Institute, Department of Epidemiology and Health Monitoring, Berlin (Neuhauser H), University Hospital Schleswig-Holstein, Kiel (Schreiber S), University Hospital Cologne and University Hospital Frankfurt, Cologne and Frankfurt (Vehreschild J), Jena University Hospital, Jena (von Lilienfeld-Toal M), Charité - Universitaetsmedizin Berlin, Berlin (Witzenrath M).

We gratefully thank the NAPKON Use and Access Committee: University Medicine Goettingen, Goettingen (Blaschke S), Lahn-Dill-Clinics, Wetzlar (Ellert C), LMU clinic of the University of Munich, Munich (Frank S), University Hospital Schleswig-Holstein, Kiel (Friedrichs A), University of Wuerzburg, Wuerzburg (Heuschmann P), Hannover Unified Biobank, Hannover Medical School, Hannover (Illig T), University Hospital Wuerzburg, Wuerzburg (Meybohm P), LMU Clinic of the University of Munich, Munich (Milger K), University Medicine Oldenburg, Oldenburg (Petersmann A), University Hospital Technical University Munich, Munich (Schmidt G), University Hospital Schleswig-Holstein, Kiel (Schreiber S), University Hospital Cologne and University Hospital Frankfurt, Cologne and Frankfurt (Vehreschild J), Jena University Hospital, Jena (von Lilienfeld-Toal M), Charité - Universitaetsmedizin Berlin, Berlin (Witzenrath M), University Hospital Essen, Essen (Witzke O).

References
1 Nonnemacher M, Nasseh D, Stausberg J, Bauer U Datenqualität in der medizinischen Forschung: Leitlinie zum adaptiven Management von Datenqualität in Kohortenstudien und Registern. 2., aktualisierte und erw. Aufl. Med. Wiss. Verl.- Ges; 2014
2 Mezzanzanica M, Boselli R, Cesarini M, Mercorio F Data quality sensitivity analysis on aggregate indicators: In: Proceedings of the International Conference on Data Technologies and Applications. Setúbal: SciTePress - Science and Technology Publications; 2012: 97–108
3 Johnson SG, Pruinelli L, Hoff A, et al. A framework for visualizing data quality for predictive models and clinical quality measures. AMIA Jt Summits Transl Sci Proc 2019;2019:630–638
4 Schmidt CO, Struckmann S, Enzenbach C, et al. Facilitating harmonized data quality assessments. A data quality framework for observational health research data collections with software implementations in R. BMC Med Res Methodol 2021;21(01):63
5 Schons MJ, Pilgram L, Reese JP, et al. The German National Pandemic Cohort Network (NAPKON): rationale, study design and baseline characteristics. Eur J Epidemiol 2022;37(08): 849–870
6 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(01):341
7 Costa-Santos C, Neves AL, Correia R, et al. COVID-19 surveillance data quality issues: a national consecutive case series. BMJ Open 2021;11(12):e047623
8 Muzoora MR, Schaarschmidt M, Krefting D, Oehm J, Riepenhausen S, Thun S Towards FAIR patient reported outcome: application of

the interoperability principle for mobile pandemic apps. In: Delgado J, Benis A, de Toledo P, et al., eds. Studies in Health Technology and Informatics. Amsterdam: IOS Press; 2021:85–86
9 Yusuf K, Rainers M, Hanß S, Krefting D Medizinische Informatik - Öffentliche Projekte / mi-num-public / NAPKON-to-Gecco-Convert. GitLab. Accessed April 12, 2022 at: https://gitlab.gwdg.de/medinfpub/mi-num-public/napkon-to-gecco
10 Kahn MG, Callahan TJ, Barnard J, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. EGEMS (Wash DC) 2016;4(01):1244
11 Embury SM, Brandt SM, Robinson JS, et al. Adapting integrity enforcement techniques for data reconciliation. Inf Syst 2001;26(08):657–689
12 Yusuf K, Tahar K, Sax U, Hoffmann W, Krefting D Assessment of the consistency of categorical features within the DZHK biobanking basic set. In: Röhrig R, Grabe N, Hoffmann VS, et al., eds. Studies in Health Technology and Informatics. Amsterdam: IOS Press; 2022: 98–106
13 Herzinger S, Gu W, Satagopam V, et al; eTRIKS Consortium. SmartR: an open-source platform for interactive visual analytics for translational research data. Bioinformatics 2017;33(14): 2229–2231
14 Covid-19 Research-Dataset - Datasets. Accessed May 22, 2022 at: https://art-decor.org/art-decor/decor-datasets–covid19f-?id=2.16.840.1.113883.3.1937.777.53.1.1&effectiveDate=2020-04-08T13%3A04%3A13&language=de-DE
15 Nakamura K. Central circuitries for body temperature regulation and fever. Am J Physiol Regul Integr Comp Physiol 2011;301(05): R1207–R1228
16 Mackowiak PA. Concepts of fever. Arch Intern Med 1998;158(17): 1870–1881
17 Geneva II, Cuzzo B, Fazili T, Javaid W. Normal body temperature: a systematic review. Open Forum Infect Dis 2019;6(04): ofz032