scientific data

OPEN

ANALYSIS

Statistical biases due to anonymization evaluated in an open clinical dataset from COVID-19 patients

Carolin E. M. Koll[1] ✉, Sina M. Hopff[1], Thierry Meurers[2], Chin Huang Lee[1], Mirjam Kohls[3], Christoph Stellbrink[4], Charlotte Thibeault[5], Lennart Reinke[6], Sarah Steinbrecher[5], Stefan Schreiber[6], Lazar Mitrov[1], Sandra Frank[7,8], Olga Miljukov[3], Johanna Erber[9], Johannes C. Hellmuth[8,10], Jens-Peter Reese[3], Fridolin Steinbeis[5], Thomas Bahmer[6,11], Marina Hagen[12], Patrick Meybohm[13], Stefan Hansch[14], István Vadász[15,16], Lilian Krist[17], Steffi Jiru-Hillmann[3], Fabian Prasser[2], Jörg Janne Vehreschild[1,12,18] & NAPKON Study Group*

Anonymization has the potential to foster the sharing of medical data. State-of-the-art methods use mathematical models to modify data to reduce privacy risks. However, the degree of protection must be balanced against the impact on statistical properties. We studied an extreme case of this trade-off: the statistical validity of an open medical dataset based on the German National Pandemic Cohort Network (NAPKON), which was prepared for publication using a strong anonymization procedure. Descriptive statistics and results of regression analyses were compared before and after anonymization of multiple variants of the original dataset. Despite significant differences in value distributions, the statistical bias was found to be small in all cases. In the regression analyses, the median absolute deviations of the estimated adjusted odds ratios for different sample sizes ranged from 0.01 [minimum = 0, maximum = 0.58] to 0.52 [minimum = 0.25, maximum = 0.91]. Disproportionate impact on the statistical properties of data is a common argument against the use of anonymization. Our analysis demonstrates that anonymization can actually preserve validity of statistical results in relatively low-dimensional data.

[1]University of Cologne, Faculty of Medicine and University Hospital Cologne, Department I of Internal Medicine, Center for Integrated Oncology Aachen Bonn Cologne Duesseldorf, Cologne, Germany. [2]Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Charitéplatz 1, 10117, Berlin, Germany. [3]University of Wuerzburg, Faculty of Medicine, Institute for Clinical Epidemiology and Biometry, Wuerzburg, Germany. [4]Department of Cardiology and Intensive Care Medicine, Bielefeld Medical Centre, Medical Faculty OWL, University of Bielefeld, Bielefeld, Germany. [5]Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt Universität zu Berlin, Berlin, Germany. [6]Internal Medicine Department I, University Medical Center Schleswig-Holstein Campus Kiel, Kiel, Germany. [7]Department of Anesthesiology, University Hospital of Ludwig-Maximilians-University (LMU), Munich, Germany. [8]Department of Medicine III, University Hospital, LMU Munich, Munich, Germany. [9]Technical University of Munich, School of Medicine, University Hospital rechts der Isar, Department of Internal Medicine II, Munich, Germany. [10]COVID-19 Registry of the LMU Munich (CORKUM), University Hospital, LMU Munich, Munich, Germany. [11]Airway Research Center North (ARCN), German Center for Lung Research (DZL), Großhansdorf, Germany. [12]Department II for Internal Medicine, Hematology/Oncology, University Hospital Frankfurt, Frankfurt am Main, Germany. [13]Department of Anaesthesiology, Intensive Care, Emergency and Pain Medicine, University Hospital Wuerzburg, Wuerzburg, Germany. [14]Department of Infection Prevention and Infectious Diseases, University Hospital Regensburg, Regensburg, Germany. [15]Department of Internal Medicine, Justus Liebig University, Universities of Giessen and Marburg Lung Center (UGMLC), Member of the German Center for Lung Research (DZL), Giessen, Germany. [16]The Cardio-Pulmonary Institute (CPI), Giessen, Germany. [17]Institute of Social Medicine, Epidemiology and Health Economics, Charité-Universitätsmedizin Berlin, Berlin, Germany. [18]German Centre for Infection Research (DZIF), partner site Bonn-Cologne, Cologne, Germany. *A list of authors and their affiliations appears at the end of the paper. ✉e-mail: carolin.koll@uk-koeln.de

Introduction

The Severe Acute Respiratory Syndrome Coronavirus II (SARS-CoV-2) pandemic has now been ongoing for more than two years[1–3]. On 15th of March 2022, worldwide, more than 460 million infected cases have been detected and more than 6 million people have died as a result of the Coronavirus Disease 2019 (COVID-19, https://covid19.who.int/). In order to collect high-quality clinical data, image data and biosamples of COVID-19 patients in Germany, the National Pandemic Cohort Network (NAPKON) was founded in 2020 as part of the Network University Medicine (NUM), a state-funded network to tackle the COVID-19 pandemic in Germany[4]. NAPKON consists of three sub-cohorts, which differ in granularity, severity of disease and sector of recruitment (Cross-Sectoral Platform (SUEP), High-Resolution Platform (HAP), and Population-Based Platform (POP)). In NAPKON, extensive clinical data have been collected during the acute course of COVID-19 and the post-COVID-19 phase, reaching in total over 4000 variables in the SUEP and HAP and over 2000 variables in the POP.

To make important parameters on the clinical course and outcome of COVID-19 openly available without restrictions, a public clinical dataset (Public Use File, PUF) with patient-level information was developed within NAPKON. The dataset is updated on a monthly basis and can be downloaded on the project website (https://napkon.de/statistik/). Most public COVID-19 datasets were generated from governmental sources or contain aggregated data only[5–9]. The NAPKON PUF combines quality-controlled clinical parameters from cohort studies or clinical routine, such as severity of disease, with demographic information, such as age and gender. Access to comprehensive data and biosamples from NAPKON can be requested by internal and external scientists through clearly defined use and access procedure. Although requests are processed promptly, they do not allow immediate access, which has been described as an effective strategy to fight the COVID-19 pandemic[10]. In addition, the reason for access and goal of data usage must be defined in advance. The NAPKON PUF provides an openly accessible overview of the NAPKON cohorts in near real-time. This simplifies the preparations for more complex data analyses and makes the cohort data more accessible for international scientists.

In NAPKON, to ensure that publishing the PUF does not compromise the privacy of individuals, the data is processed through an anonymization pipeline. The pipeline uses mathematical and statistical privacy models preventing re-identification of individuals, singling out as well as the inference of sensitive information. Records of individuals whose publication would not meet the privacy guarantees specified are withhold from the dataset. The anonymization pipeline is based on the approach used in the LEOSS project that has already been successfully used for releasing data about over 10,000 patients to the public[11].

Although data anonymization can significantly contribute to protecting the privacy of individuals, modifying the data can lead to a reduction of its usefulness. While the anonymization process implemented for NAPKON contains optimization procedures to minimize the loss of information, the extent to which the transformations performed impact the usefulness of the dataset can only be investigated in the context of specific usage scenarios. The objective of this work was to evaluate whether and how the anonymization process used in the creation of the NAPKON PUF affects its statistical properties with regard to the three NAPKON sub-cohorts. To address this question, we performed multiple evaluations on the dataset before and after anonymization. The evaluations included descriptive statistics and regression models. Thereafter, we assessed the extent of bias introduced by the anonymization process used in our results and thus its effects on the dataset's usefulness.

Results

Anonymized clinical dataset. The NAPKON PUF used in this study contained clinical data from 3,904 cases captured in 15 variables. Following a qualitative analysis of the attributes contained in the dataset, the risk of linkage or singling out was controlled by reducing the uniqueness of combinations of the variables age, gender, quarter and year of diagnosis, and cohort. The risk of sensitive attribute inference was further controlled for the variables defining the severity of disease, the patient status at the end of acute phase, presence of intensive care treatment or invasive ventilation, and ability and any symptoms at three months follow up. Details can be found in the methods section.

Fraction of cases published. The PUF contained a subset of the cases present in the original dataset, as the anonymization process was configured to withhold cases from release for which the defined privacy guarantees would not hold. Figure 1 provides an overview of the number of cases included in the PUF for increasing sizes of the dataset and for the three sub-cohorts within NAPKON. The size of the original dataset has increased with the time the registry was running.

The fraction of cases from the NAPKON cohort that can be included in the PUF increased to over 85% as soon as at least 2,250 cases were documented (which happened on 2021-05-18). The curve flattened out at 1,600 documented cases. The highest fraction of cases included in the PUF (87%) was reached when 4,350 cases were documented. The absolute number of cases in the PUF was reduced due to anonymization from 1,697 to 1,410 (83%) for the SUEP, from 2,346 to 2,280 (97%) for the POP, and from 519 to 237 (45%) for the HAP.

Descriptive statistics before and after anonymization. Table 1 shows the cohort descriptions from the original dataset from 2022-03-15 ($n = 4{,}562$) and the PUF ($n = 3{,}904$) for each of the three NAPKON cohorts. The age distribution (Fig. 2(a)) before and after anonymization differed significantly ($P < 0.001$). It is notable that in particular the age groups under 18 and over 79 years were represented with only a few cases in the original dataset, which is why no minors were included in the PUF and the age group over 79 years was reduced by 73% in size (150/205) through the anonymization process. In contrast, the sizes of the other age groups differed only slightly between the two datasets. The largest impact can be observed for the HAP cohort (Fig. 3). Only 56% (129/230) of its cases in the age group 40–59 and 58% (108/187) of its cases in the age group 60–79, were included

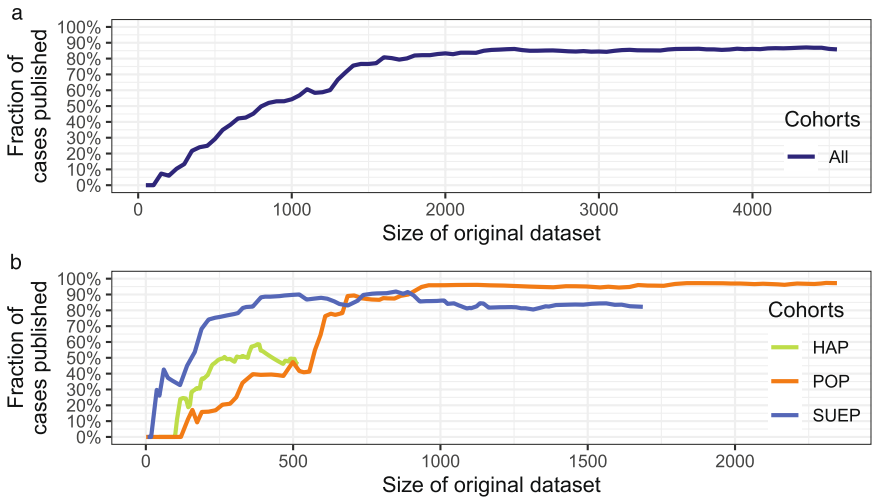

Fig. 1 Fraction of cases published for the complete NAPKON dataset **(a)** and **(b)** the High-Resolution Platform (HAP), the Population-based Platform (POP), and the Cross-Sectoral Platform (SUEP).

in the PUF. The POP shows the smallest case reduction regarding the variable age, however with a significantly different age distribution ($P = 0.011$). The age category over 79 years stands out with a case reduction of 66% (12/35).

The gender distribution is shown in Fig. 2(b). Males were predominant in the original (2,412/4,562; 53%) and anonymized (2,050/3,904, 53%) dataset. The gender distribution did not differ significantly before and after anonymization ($P = 0.74$). Analyzing the three cohorts (Fig. 3), 85% (868/1,020) of male cases could be published for the SUEP, 96% (1,001/1,040) for the POP and 51% (181/352) for the HAP. Consequently, gender distribution differed significantly in the HAP ($P = 0.023$), but not in the SUEP ($P = 0.172$) and in the POP ($P = 0.753$). The cases with a first COVID-19 diagnosis in 2020 were mainly from the POP and in 2021 from the SUEP (Fig. 2(c)). No significantly different distribution of quarter of diagnosis before and after anonymization could be observed for 2020 ($P = 0.477$). For 2021, the distribution differed significantly ($P = 0.044$) with regard to all cohorts, but not for cases in the SUEP ($P = 0.66$). In the first quarter of 2022 the case number was still too low so that no cases were included in the anonymized dataset.

The distribution of the documented disease phases was significantly different between the original and the anonymized dataset, with mild phases overrepresented and moderate phases underrepresented ($P < 0.001$, Fig. 2(d)). The proportion of included cases in the PUF differed for the POP, HAP, and SUEP. 97% (2,228/2,286) of cases from the POP having a documented mild COVID-19 phase (ambulatory treatment) were included in the PUF. A percentage of 84% (1,154/1,382) of cases with a documented moderate phase and 89% (312/349) with a documented severe phase were included in the dataset after anonymization from the SUEP and 44% (171/385) and 47% (93/199), respectively, from the HAP (Fig. 3(d)).

The distribution of patient status at the end of the acute phase significantly differed with more cases in an ambulatory setting and less cases discharged, transferred, or died in the anonymized dataset in comparison to the original dataset ($P < 0.001$, Fig. 4(a)). Especially for the HAP, only 44% (180/407) of the cases with status "discharged" in the original dataset were added to the anonymized dataset (Fig. 5). For the POP, most cases were outpatients and only 2% of cases were removed from the original dataset.

Case fatality rate. We computed the case fatality rate for the SUEP and HAP cohorts (no fatalities in POP at first visit) for different sizes of the NAPKON dataset, showing the impact of the anonymization procedure on increasing documented cases over time (Fig. 4(b)). The case fatality rate was overestimated in the anonymized dataset with a size of up to 1,486 cases. For more than 1,486 cases documented in NAPKON, the case fatality rate was slightly underestimated in the anonymized dataset. In the original dataset containing 2,096 cases (dataset with three cohorts $n = 4,562$, 120 with missing information) the case fatality rate was 8% before and after anonymization (before: 165/2,096, after: 121/1,551). With more cases included in the original dataset, the bias of the case fatality rate before and after anonymization decreased. The differences in bias observed for the different cohorts is presented in Figs. 6 and 7. The median difference of the case fatality rate before and after anonymization for different sizes of the NAPKON datasets was 0.2% for the SUEP (interquartile range [IQR]: 0.1%–0.3%) and 2.9% (IQR: 1.7%–5.1%) for the HAP.

Regression analyses before and after anonymization. We investigated the impact of the used anonymization procedure on associations between parameters by computing four regression models for different sizes of datasets before and after anonymization (Fig. 8). The results of the NAPKON PUFs consistently reflected the trends of the associations found in the respective original datasets.

The odd ratios (ORs) deviated for different sizes of datasets. As an example, in the original dataset, the association between inpatient cases from the SUEP aged older than 59 years and dying during the acute phase was estimated with a minimum of $OR = 1.72$ with 95%-confidence interval (CI) 0.38–7.49 (154 cases) and a maximum

	Original (n = 4,562)			Anonymized (n = 3,904)		
	SUEP (n = 1,697)	POP (n = 2,346)	HAP (n = 519)	SUEP (n = 1,387)	POP (n = 2,280)	HAP (n = 237)
Age in years						
<18	40 (2.4%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
18–39	317 (18.7%)	920 (39.2%)	78 (15.0%)	275 (19.8%)	917 (40.2%)	0 (0%)
40–59	609 (35.9%)	967 (41.2%)	230 (44.3%)	541 (39.0%)	957 (42.0%)	129 (54.4%)
60–79	584 (34.4%)	412 (17.6%)	187 (36.0%)	528 (38.1%)	394 (17.3%)	108 (45.6%)
>79	147 (8.7%)	35 (1.5%)	23 (4.4%)	43 (3.1%)	12 (0.5%)	0 (0%)
Missing	0 (0%)	12 (0.5%)	1 (0.2%)	0 (0%)	0 (0%)	0 (0%)
Gender						
Male	1,020 (60.1%)	1,040 (44.3%)	352 (67.8%)	868 (62.6%)	1001 (43.9%)	181 (76.4%)
Female	677 (39.9%)	1,305 (55.6%)	166 (32.0%)	519 (37.4%)	1,279 (56.1%)	56 (23.6%)
Missing	0 (0%)	1 (0.04%)	1 (0.2%)	0 (0%)	0 (0%)	0 (0%)
Quarter and year of first COVID-19 diagnosis						
Q1 2020	2 (0.1%)	554 (23.6%)	11 (2.1%)	0 (0%)	541 (23.7%)	0 (0%)
Q2 2020	0 (0%)	279 (11.9%)	24 (4.6%)	0 (0%)	271 (11.9%)	0 (0%)
Q3 2020	7 (0.4%)	275 (11.7%)	19 (3.7%)	0 (0%)	265 (11.6%)	0 (0%)
Q4 2020	68 (4.0%)	748 (31.9%)	61 (11.8%)	0 (0%)	740 (32.4%)	0 (0%)
Q1 2021	534 (31.5%)	367 (15.6%)	156 (30.1%)	496 (35.8%)	365 (16.0%)	130 (54.9%)
Q2 2021	397 (23.4%)	86 (3.7%)	108 (20.8%)	390 (28.1%)	76 (3.3%)	85 (35.9%)
Q3 2021	208 (12.3%)	1 (0.04%)	39 (7.5%)	194 (14.0%)	0 (0%)	0 (0%)
Q4 2021	357 (21.0%)	1 (0.04%)	72 (13.9%)	307 (22.1%)	0 (0%)	22 (9.3%)
Q1 2022	66 (3.9%)	0 (0%)	7 (1.3%)	0 (0%)	0 (0%)	0 (0%)
Missing	58 (3.4%)	35 (1.5%)	22 (4.2%)	0 (0%)	22 (1.0%)	0 (0%)
WHO Clinical Progression Scale most severe phase						
Mild	191 (11.3%)	2185 (93.1%)	0 (0%)	149 (10.7%)	2131 (93.5%)	0 (0%)
Moderate	1,135 (66.9%)	124 (5.3%)	320 (61.7%)	919 (66.3%)	114 (5.0%)	144 (60.8%)
Severe	349 (20.6%)	37 (1.6%)	199 (38.3%)	307 (22.1%)	35 (1.5%)	93 (39.2%)
Missing	22 (1.3%)	0 (0%)	0 (0%)	12 (0.9%)	0 (0%)	0 (0%)
Patient status at end acute phase						
Ambulant	192 (11.3%)	2,108 (89.9%)	0 (0%)	149 (10.7%)	2,069 (90.7%)	0 (0%)
Discharged	1,084 (63.9%)	124 (5.3%)	407 (78.4%)	911 (65.7%)	112 (4.9%)	180 (75.9%)
Referral/transfer	221 (13.0%)	0 (0%)	26 (5.0%)	176 (12.7%)	0 (0%)	14 (5.9%)
Dead	103 (6.1%)	0 (0%)	62 (11.9%)	87 (6.3%)	0 (0%)	34 (14.3%)
Missing	97 (5.7%)	114 (4.9%)	23 (4.4%)	64 (4.6%)	99 (4.3%)	9 (3.8%)
Hospitalization						
Yes	1,502 (88.5%)	161 (6.9%)	519 (100%)	1,238 (89.3%)	149 (6.5%)	237 (100%)
No	192 (11.3%)	2,185 (93.1%)	0 (0%)	149 (10.7%)	2,131 (93.5%)	0 (0%)
Missing	3 (0.2%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Intensive Care Treatment						
Yes	403 (23.7%)	33 (1.4%)	200 (38.5%)	356 (25.7%)	31 (1.4%)	99 (41.8%)
No	1,294 (76.3%)	2,313 (98.6%)	319 (61.5%)	1,031 (74.3%)	2,249 (98.6%)	138 (58.2%)
Invasive ventilation						
Yes	153 (9.0%)	18 (0.8%)	111 (21.4%)	128 (9.2%)	18 (0.8%)	60 (25.3%)
No	1,522 (89.7%)	2,328 (99.2%)	327 (63.0%)	1,245 (89.8%)	2,262 (99.2%)	131 (55.3%)
Missing	22 (1.3%)	0 (0%)	81 (15.6%)	14 (1.0%)	0 (0%)	46 (19.4%)
3-month follow-up available						
Yes	968 (57.0%)	2,346 (100%)	158 (30.4%)	859 (63.0%)	2,280 (100%)	103 (43.5%)
No/not yet	729 (43.0%)	0 (0%)	361 (69.6%)	528 (38.1%)	0 (0%)	134 (56.5%)
Any symptoms at 3-month follow-up (if 3-month follow-up available)						
Yes	305 (31.5%)	900 (38.4%)	43 (27.2%)	280 (32.6%)	874 (38.3%)	36 (35.0%)
No	663 (68.5%)	862 (36.7%)	115 (72.8%)	579 (67.4%)	847 (37.1%)	67 (65.0%)
Missing	0 (0%)	584 (24.9%)	0 (0%)	0 (0%)	559 (24.5%)	0 (0%)
Ability to work at 3-month follow-up (if 3-month follow-up available)						
Yes	454 (46.9%)	1,754 (74.8%)	41 (26.0%)	442 (51.5%)	1,729 (75.8%)	30 (29.1%)
No	89 (9.2%)	126 (5.4%)	91 (57.6%)	85 (10.0%)	124 (5.4%)	59 (57.3%)
Missing	425 (43.9%)	466 (19.9%)	26 (16.5%)	332 (38.7%)	427 (18.7%)	14 (13.6%)

Table 1. Description of the three NAPKON sub-cohorts in the original and anonymized (PUF) dataset from 2022-03-15. 3-month follow-up obtained 10 to 14 weeks after day of first COVID-19 diagnosis (for POP retrospective documentation). SUEP = Cross-Sectoral Platform; POP = Population-Based Platform; HAP = High-Resolution Platform.

of OR = 4.29 with 95%-CI = 2.47–7.91 (908 cases). Comparing the models derived from the NAPKON PUF with those derived from the original dataset, the ORs and CIs are getting closer to the original results when more cases were included. The median absolute deviations of estimated ORs before and after anonymization for datasets of different sizes were for inpatient cases from the SUEP aged older than 59 years and dying during the acute phase (a) 0.2 with minimum (min) 0 and maximum (max) 1.48 difference (Fig. 8(a)), for 49 to 59 years old inpatient cases of the SUEP and HAP that were in a severe phase having any symptom at three month follow up median = 0.03, min = 0.01 and max = 0.13 (Fig. 8(b)), for female cases from the POP having any symptom at three month follow-up median = 0.01, min = 0, and max = 0.58 (Fig. 8(c)), and for cases from the HAP aged older 59 years having intensive care treatment median = 0.52, min = 0.25, max = 0.91 (Fig. 8(d)).

Reidentification risk. Figure 9 illustrates how anonymizing the dataset affects the re-identification risk for the patients included. For both, the original and the anonymized dataset, the lowest, highest, and the average re-identification risk is provided. As the original dataset contained unique records, the highest re-identification risk was 100%. In the NAPKON PUF, the highest risk was reduced to not more than 9.09%, as one of the used privacy models required each record to be indistinguishable from at least 10 other records ($1/11 = 0.\overline{09}$). As expected, the average risk was much lower and further decreased with an increasing number of documented cases. Furthermore, there was no difference in the lowest re-identification risk between the datasets. The original dataset contained records with a low re-identification risk requiring no additional protection.

Discussion

In this study, we found that statistical bias introduced by anonymization for the NAPKON PUF is small. Descriptive statistics as well as regression analyses showed acceptable differences with only little biases in statistical results. Cases with less frequent characteristics were excluded from the anonymized dataset. However, regression models showed comparable results for the anonymized and the original dataset if parameters with relatively rare values were used as independent variable. The bias decreased as the size of the original dataset increased. Overall, the NAPKON PUF contains only a few variables of the original dataset with a reduced case number, but it preserves important information and a high utility.

We note that statistical results obtained from the original dataset vary with an increasing case number. Small deviations may therefore generally be acceptable. However, since anonymized datasets are likely to contain additional biases and may vary across cohorts with different underlying characteristics, it is important to make the anonymization process transparent when they are shared. In addition to enabling an analysis team to adequately interpret the results for themselves from the anonymized data provided, transparency of the data preprocessing steps is important for research integrity and for making the limitations of studies visible[12].

Comparing the fraction of cases published in the PUF over time (different sizes of original dataset) with the recruitment rate in NAPKON (https://napkon.de), a relation can be seen between the high recruitment rates in NAPKON in Q4/2020, Q1/2021 as well as Q2/2021 and a larger number of cases that could be included in the PUF in these time periods. This is caused by the fact that the privacy models utilized are based on the principle of 'hiding in the crowd'[13]. The more cases are included in the original dataset; the more cases can be included in the anonymized dataset as well. Due to a high-granularity data collection in the HAP, the case number was lower than in the SUEP and the POP. This explains why a lower fraction of cases from the HAP can be included in the PUF compared to the other cohorts and why the course of cases included in the PUF flattens out after Q2 2021. Furthermore, the results of the descriptive analyses showed no significant differences in the distribution of gender and quarters of diagnosis in 2020 between the original and the anonymized dataset. For age, disease severity and quarters of diagnosis in 2021 the differences between the two datasets were significant with regard to all of the three cohorts.

The results of this study show that the bias in descriptive characteristics due to anonymization processes can be small. We therefore hypothesize that anonymized data with a comparable complexity may be suitable for various applications. In the following, four possible applications of anonymized data sets are presented, as evidenced by examples from NAPKON. (i) Anonymized data could support researchers in cohort discovery. For the NAPKON PUF, the patient-level information on general characteristics and clinical course severity can help to describe the recruited patient collective, providing insights into the different recruitment strategies in the NAPKON cohorts. (ii) Anonymized datasets further could contribute to facilitate the feasibility process when applying for comprehensive datasets, as the public dataset is an extract from the comprehensive dataset. In NAPKON, researchers can explore the data and feasibility for their research question on their own. A majority of requests for data usage were based mainly on the parameters included in the NAPKON PUF. (iii) Furthermore, anonymized dataset with same parameters originating from different cohorts could contribute to the assessment of the generalizability of results. In NAPKON, the transferability of results from one cohort to another could be assessed by comparing characteristics between the cohorts. For example, we know that the age structure in the SUEP and the HAP differs. Therefore, it can be assumed that the computed OR for age older than 65 years and death in the acute phase of COVID-19 slightly differs in the populations. (iv) In addition, the possibility to explore general patient characteristics in anonymized datasets could enhance the assessment of a present selection bias. For example, patient characteristics of NAPKON could be compared to databases containing data from any reported SARS-CoV-2 infected hospitalized patient to assess the generalizability of results. Although open aggregated descriptive statistics of clinical datasets would already enable the estimation of a selection bias, an open clinical dataset containing patient-level information increases flexibility and empowerment of researchers.

We further showed that an anonymized dataset might reflect trends of associations between parameters. These findings open a new field of application for an anonymized dataset in addition to the use for i - iv. If the research question, as well as the covariates and confounders of the intended analysis are known and well defined,

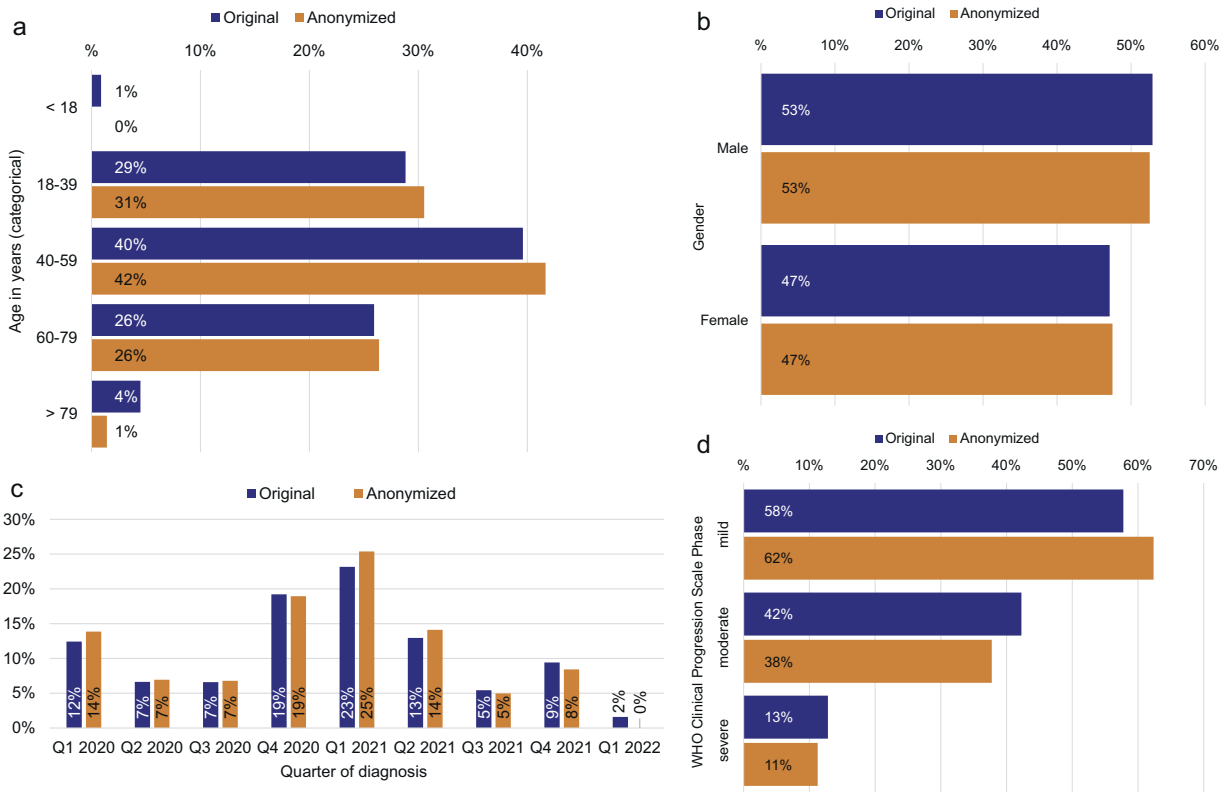

Fig. 2 Comparison of demographic parameters of patients for the original dataset (n = 4,562) and the anonymized dataset (PUF; n = 3,904). The proportions are given in percentage. Note: The percentages in the PUF may be larger if the number of censored cases is unbalanced. (**a**) Age distribution in years, (**b**) gender distribution, (**c**) distribution of quarter and year of first positive SARS-CoV-2 test, and (**d**) distribution of the disease severity in the course of disease. WHO = World Health Organization.

anonymization procedures could be used to create an open dataset specific to pre-defined research questions. Routinely collected clinical data or cohort data could be used to generate an open dataset, overcoming strict privacy regulations and logistical as well as personal costs. In particular, for regression analyses, where the number of variables included is often limited, a small anonymized dataset may be sufficient. We therefore performed simple regression analysis to show the impact of even a small public dataset.

Our results confirm findings from other studies analyzing the impact of anonymizing real-world data for use in real-world contexts, of which, however, there are very few to our knowledge. In the LEOSS project, it was shown that the association between age and death could be replicated in the anonymized dataset with regard to significance and trend of ORs[11]. However, clinical data collected in LEOSS are also comparable to the NAPKON data (less granularity in LEOSS)[14]. This was one reason why we decided to use the principles of the LEOSS anonymization process in this study. In a dataset from the social sciences, a study showed that statistical bias introduced by k-anonymity with k = 5 and six key variables (four patient characteristics, two activity variables) was small[15].

Our analyses performed on the NAPKON PUF demonstrate that for specific usage scenarios the bias due to anonymization of a reduced dataset may be acceptably small. However, there are limitations regarding the complexity of the dataset and the generalizability of the results to other use cases. (i) The extent to which the chosen anonymization method affects a dataset and subsequent analyses must be examined on a case-by-case basis, which is why our findings cannot necessarily be generalized to other datasets and analyses. In particular, it is difficult to make a statement about the transferability of our results a priori. (ii) In addition, the configuration of the anonymization process must be assessed on case-by-case basis. Publishing more sensitive information, such as information on additional infections with the human immunodeficiency virus, may require more stringent anonymization techniques. (iii) Furthermore, the assessment of the bias introduced by anonymization also needs to be performed from the perspective of individual usage scenarios. In a public clinical dataset used for research on discrimination or underprivileged sub-populations anonymization may mask the severity of disparity[16]. (iv) Moreover, for generating the NAPKON PUF, it was necessary to reduce the datasets' complexity from the beginning. This resulted in a subset of 15 variables. Some variables, such as age or clinical states defined by the WHO Clinical Progression Scale[17], underwent a categorization to reduce their granularity. By reducing the datasets' complexity, the number of cases withhold by the anonymization process can be reduced. However, this limits the possibilities of complex statistical analyses as some variables needed are missing or too much generalized for accurate evaluation.

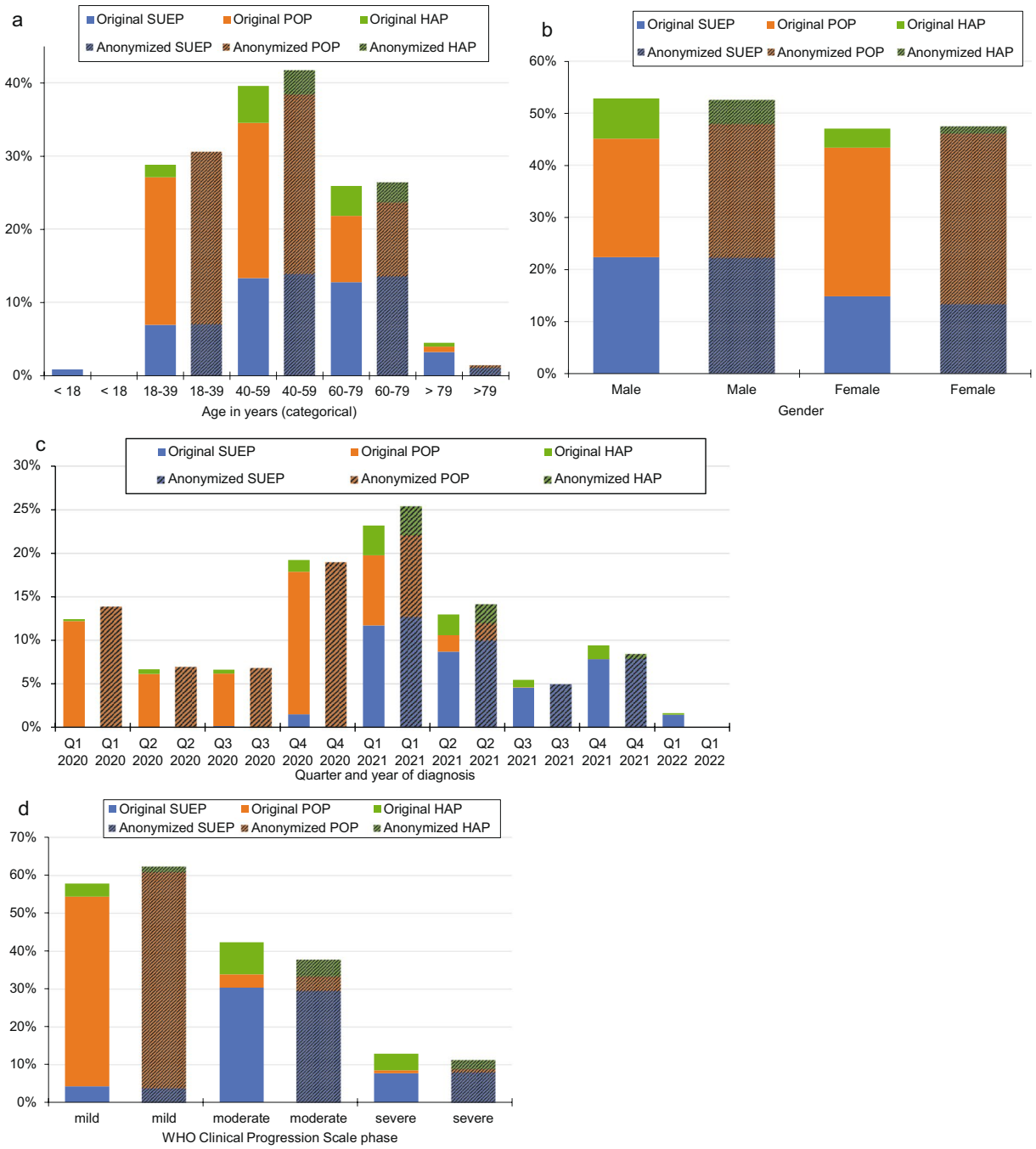

Fig. 3 Comparison of demographic parameters of patients for the original data set (n = 4,562) and the anonymized data set (PUF, n = 3,904). (**a**) Age distribution in years, (**b**) gender distribution, (**c**) distribution of quarter and year of first positive SARS-CoV-2 test, and (**d**) distribution of the disease severity in the course of disease. HAP = High-Resolution Platform; POP = Population-based Platform; SUEP = Cross-Sectoral Platform; WHO = World Health Organization.

The NAPKON PUF is another practical example of an anonymized dataset with high utility. Nevertheless, data protection legislation still complicates the publication of individual-level anonymous data in Germany and other countries. Therefore, further progress is needed to establish a more standardized way of data anonymization to provide a solid bridge between legal requirements and technical implementation options. As a next step, a standardized framework could be supported by legal opinions, helping to remove uncertainty of whether datasets can be considered legally anonymous.

In addition to statistical considerations regarding the utility of an anonymized dataset, the economic and social benefits must be emphasized. Anonymized clinical data can be easily shared resulting in maximal benefit. In comparison to that, pseudonymized clinical data are access restricted and protected so that time-consuming and potentially costly use and access processes are necessary. Complex data sets like in NAPKON are usually

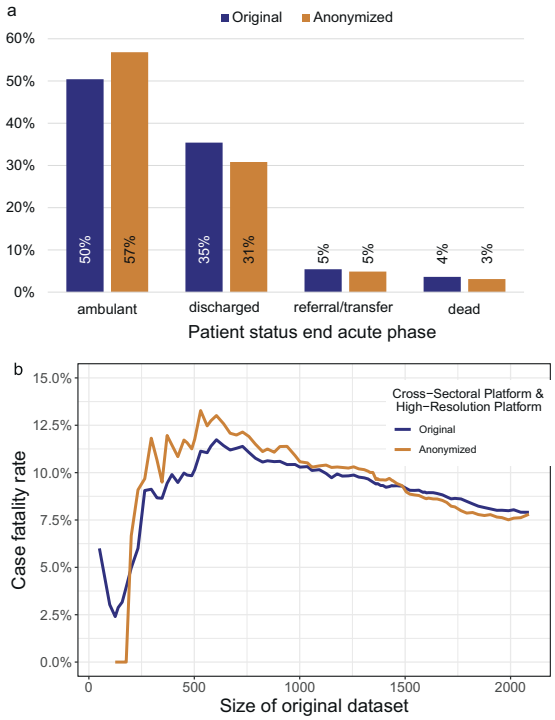

Fig. 4 Comparison of patient status at end of acute phase before and after anonymization (anonymized dataset = PUF). The proportions are given in percentage. Note: The percentages in the PUF may be larger if the number of censored cases is unbalanced. (**a**) Distribution for the original dataset containing n = 4,562 and resulting anonymized dataset (n = 3,904). (**b**) Case fatality rates (patient status dead) are computed for the Cross-Sectoral Platform (SUEP) and High- Resolution Platform (HAP) cohorts over different sizes of original dataset. In the plot, the size of the original dataset is adjusted by the number of HAP and SUEP patients. To note, the Population-based Platform (POP) has recruited patients that survived SARS-CoV-2 infection only.

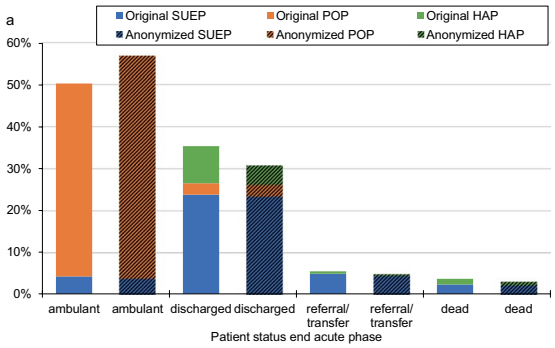

Fig. 5 Comparison of patient status at end of acute phase before and after anonymization. Distribution for the original dataset containing n = 4,562 and resulting anonymized dataset (PUF, n = 3,904). HAP = High-Resolution Platform; POP = Population-based Platform; SUEP = Cross-Sectoral Platform.

particularly costly to generate and hence often require public funding. Access to anonymized data can help to justify the cost of data collections[18] and data access is expedited, which is particularly important in times of pandemics. Furthermore, anonymization could improve the cost-effectiveness in case of scarce scientific resources, as anonymized datasets without access restrictions do not need continuous financial expenditures for data distribution[18].

Finally, we applied a mathematical anonymization procedure to a large clinical dataset containing demographics and clinical information on COVID-19 disease courses. In our study, the statistical bias introduced by anonymization was small. Therefore, we advocate for the use of anonymized clinical datasets in research, supporting use cases such as feasibility analyses or pre-defined non-complex statistical analyses. However, it is difficult to estimate the impact of anonymization in novel datasets a priori. Therefore, statistical interpretation of an anonymized dataset should be carried out with caution.

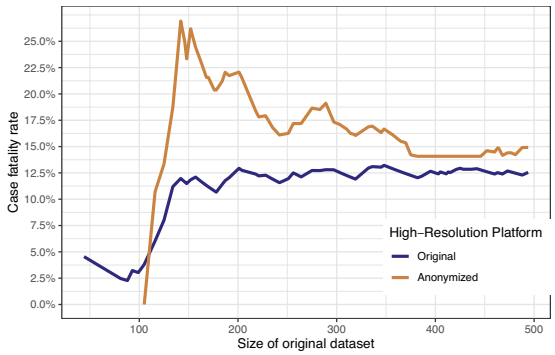

Fig. 6 Case fatality rate for the High-Resolution Platform (HAP). Anonymized dataset = PUF.

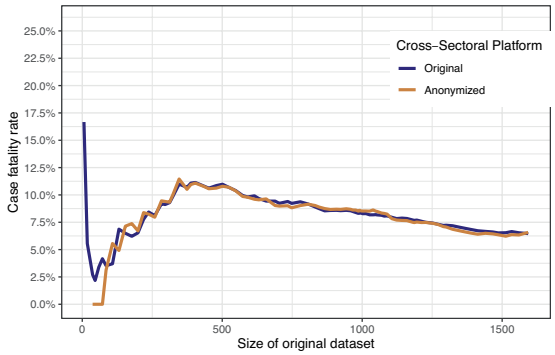

Fig. 7 Case fatality rate for the Cross-Sectoral Platform (SUEP). Anonymized dataset = PUF.

Methods

We investigated the statistical bias due to anonymization within the PUF from the NAPKON cohort. The open dataset contains clinical data from SARS-CoV-2 infected patients treated in hospitals, by general practitioners or infected individuals that were identified and contacted via the local public health authorities, collected in three sub-cohorts.

Ethical statement. NAPKON was approved by local ethics committees of participating sites (primary approval for the SUEP: Ethics Committee of the Department of Medicine at Goethe University Frankfurt (local ethics ID approval 20–924), for the HAP: Ethics Committee of the Charité – Universitätsmedizin Berlin (local ethics ID approval EA2/226/21 and EA2/066/20), for the POP: Ethics Committee of the Department of Medicine at Christian-Albrechts-University Kiel (local ethics ID approval D 537/20)). Patients consent was obtained for data collection, storage, and processing. The data from the anonymized PUF were considered anonymous. The PUF did not contain directly personal information and the re-identification risk was lowered applying the anonymization pipeline.

NAPKON cohort platforms. The Cross-Sectoral Platform (SUEP) cohort recruits SARS-CoV-2 infected patients from university and non-university medical centers as well as outpatient settings across 40 study sites, which are followed up over a 12-month period. Both pediatric and adult patients are included in the SUEP. Additional cases from the CORKUM cohort were subsequently added to the SUEP dataset[19]. Part of the cases were already recruited at the beginning of 2020. The High-Resolution Platform (HAP) cohort follows a deep phenotyping protocol at eleven university sites and has established follow-up investigations up to 36 months after initial COVID-19 diagnosis. Cases from the preceding Pa-COVID study[20] were also added to the HAP dataset, including data from patients recruited since the beginning of 2020. The Population-Based Platform (POP) cohort differs from the other two cohorts in that it contains retrospectively collected data from the acute course and prospectively collected data, imaging data, and biosamples from six to 12 months after the initial COVID-19 diagnosis. It focuses on health consequences of SARS-CoV-2 infection in the general adult population[21].

Dataset. The dataset used for the evaluation of the statistical effects of anonymization (NAPKON PUF) was extracted from the comprehensive clinical dataset of NAPKON (original dataset), including all cases documented or integrated from 2020-11-01 to 2022-03-15. 15 variables were included in the NAPKON PUF, containing demographic variables, cohort information, and clinical course and outcome parameters (Table 2). Table 3 compares the different features of the NAPKON PUF and the original dataset.

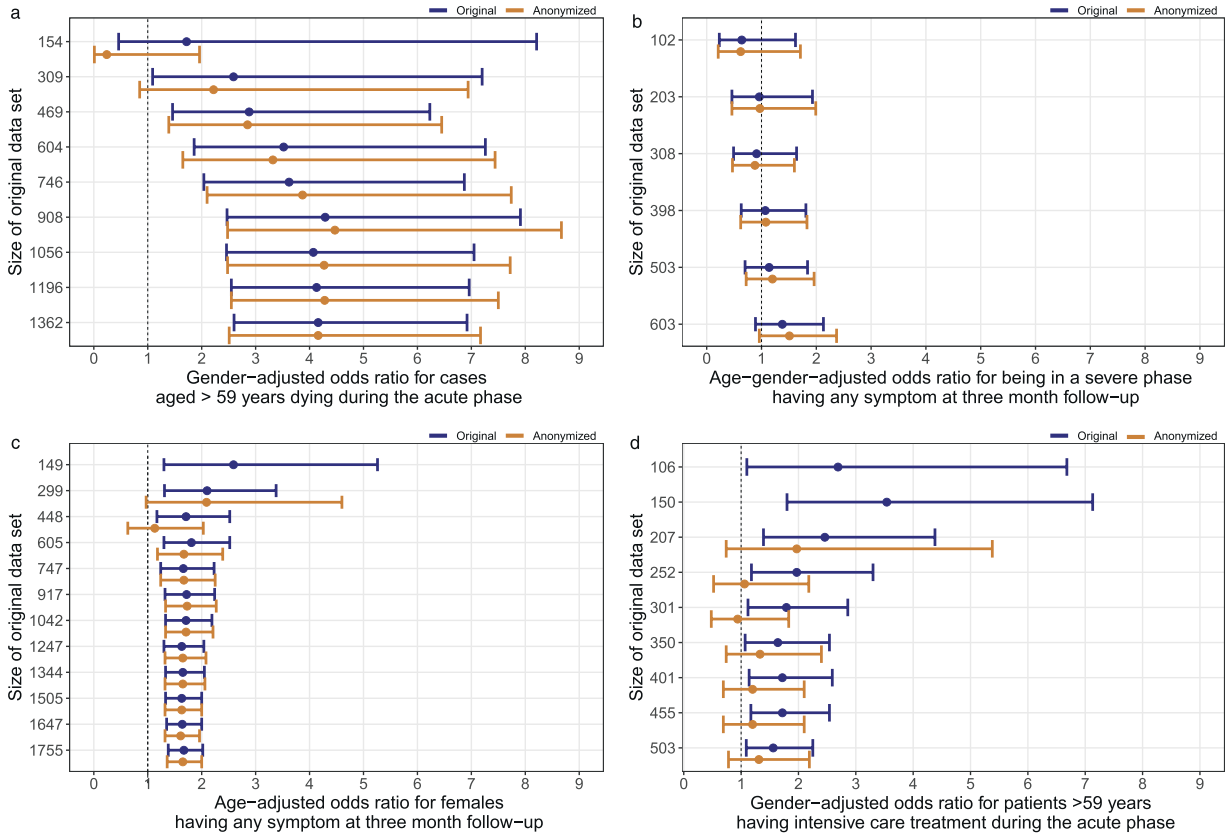

Fig. 8 Odds ratios (OR) and 95%-confidence intervals (CI) of patient characteristics and outcomes in the dataset before and after anonymization for different sizes of the original dataset (anonymized dataset = PUF). In the graphs, the number of records in the original dataset was adjusted according to the number of cases included in the regression analysis, excluding missing data. To note, datasets with no ORs and CIs do not contain the relevant information for the respective regression model. **(a)** Inpatient cases from the Cross-Sectoral Platform (SUEP). **(b)** Cases from the High-Resolution Platform (HAP) and inpatient SUEP aged between 49 and 59 years that survived the acute phase of COVID-19 (ambulant or discharged). **(c)** Cases from the Population-Based Platform (POP). **(d)** Cases from the High-Resolution Platform (HAP).

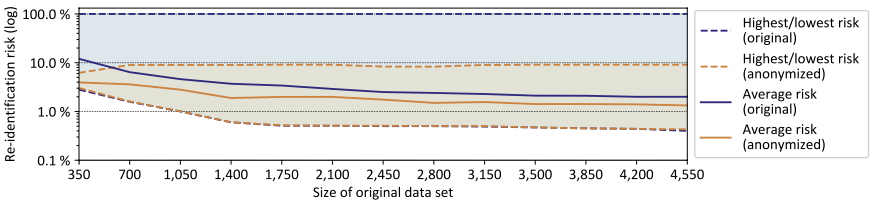

Fig. 9 Re-identification risks based on the uniqueness of k-variables before and after anonymization for different sizes of the original dataset (anonymized = PUF).

Variable selection. The variables of the NAPKON PUF are based on the parameters defined in the German COVID-19 core dataset 'the German Corona Consensus Dataset (GECCO)'[22]. The demographic variables provide a basic overview of the cohorts. By specifying the cohort of included patients, the case numbers can be queried on a cohort-specific basis. This is particularly important, because the cohorts cover different health sectors and data of varying depth. Therefore, not all cohorts are suitable for answering all research questions. For example, the SUEP captures the acute course in the outpatient setting, whereas the HAP only includes patients from the inpatient setting. Therefore, for questions intended to cover outpatient cases, it is possible to filter specifically by "SUEP" and "no hospitalization". The WHO Clinical Progression Scale focuses on ventilation parameters. Categorization into mild, moderate, and severe does not allow a precise conclusion to be made about an intensive care stay, since the use of different ventilation modalities is possible in normal or intensive care units, depending on the hospital. In addition, other reasons such as the need for dialysis also lead to an intensive care stay for a COVID patient who is not ventilated. Therefore, intensive care treatment is requested individually in the PUF. In addition, the severe phase of the WHO Clinical Progression Scale includes patients with non-invasive ventilation

VARIABLE	Categories	R	A	D	Is Key
Age at diagnosis	< = 17 years, 18–39 years, 40–59 years, 60–79 years, > = 80 years	3	3	3	Yes (9)
Gender	Male, female, diverse, unknown/missing	3	3	2	Yes (8)
Quarter first diagnosis	Q1, Q2, Q3, Q4, unknown/missing	1	3	2	Yes (6)
Year first diagnosis	2020, 2021, 2022, 2023, unknown/missing	1	3	2	Yes (6)
Cohort	HAP, POP, SUEP	3	3	2	Yes (8)
Mild disease phase	Yes, no, unknown/missing	1	2	1	No (4)
Moderate disease phase	Yes, no, unknown/missing	1	2	2	No (5)
Severe disease phase	Yes, no, unknown/missing	1	2	2	No (5)
Patient status at end of acute phase	Discharged, ambulant, referral/transfer, dead, unknown/missing	1	2	2	No (5)
Hospitalization during acute phase	Yes, no, unknown/missing	1	2	2	No (5)
Intensive care treatment	Yes, no, unknown/missing	1	2	2	No (5)
Invasive ventilation	Yes, no, unknown/missing	1	1	2	No (4)
Availability of 3- month follow-up	Yes, no/not yet	1	2	1	No (5)
Ability to work at 3- month follow-up	N/a, yes, no, unknown/missing	1	2	2	No (5)
Any symptom at 3- month follow-up	N/a, yes, no, unknown/missing	1	1	2	No (4)

Table 2. Assessment of the re-identification risk associated with individual variables. Key variables are defined by R (replicability) + A (availability) + D (distinguishability) >5. 1 = low risk, 2 = medium risk, 3 = high risk. HAP = High-Resolution Platform (hospitalized patients only), POP = Population-Based Platform (retrospective documentation after SARS-CoV-2 infection), SUEP = Cross-Sectoral Platform (in- and outpatients).

	NAPKON PUF	Original NAPKON dataset
Number of variables	15	>2.000
Cases included	3,904	4,562
Highest re-identification risk	9,09%	100%
Accessibility	public	for selected scientists after application
Linkage with additional data	No further linkage possible	Linkage to further data, image data and biosamples possible
Usage	Feasibility checks, basic statistics, cohort descriptions	Comprehensive in-depth analyses

Table 3. Comparison of different features of the NAPKON PUF and its original dataset.

(NIV), high flow and mechanical ventilation. To delineate the number of patients receiving invasive ventilation, this variable was additionally added to the PUF.

In addition to the acute phase of the COVID-19 disease, the long-term outcome plays a main role in scientific analyses, as the long-term consequences of the disease are not yet clear[23,24]. Furthermore, clinical courses differ with regard to SARS-CoV-2 variants[25,26]. The quarter of first COVID-19 diagnosis can help with the assignment to a certain variant. Therefore, asking for the ability to work and the persistence of symptoms three months after original infection greatly enhances the informative strength of the anonymous dataset. The availability of the 3-months follow-up is captured in the dataset (retrospective documentation for POP) and it is particularly important for feasibility queries.

Anonymization. The anonymization pipeline used for the NAPKON PUF is based on the qualitative and quantitative principles developed for the public dataset of LEOSS[11]. LEOSS is one of the largest COVID-19 registries in Germany with comprehensive clinical data of mainly hospitalized SARS-CoV-2 infected patients[14,27].

Analogously to the approach implemented for LEOSS, we used the method by Malin *et al.*[28] and rated all variables along the axes replicability, availability and distinguishability (1 = low risk, 2 = medium risk, 3 = high risk). In the following the rating for two variables is explained exemplarily. The age categorization hardly changes over the observational period (replicability = high risk). The age is often known and can be determined by appearance (availability = high risk), and there is a great variation within the society (distinguishability = high risk). The need of intensive care treatment, from a medical point of view, is only slightly likely to occur repetitively, as the forms of treatment change in the course of time (replicability = low risk). However, for long hospital stays with known absence and possible subsequent rehabilitation, the need for intensive treatment could be suspected (availability = medium risk). Intensive care treatment as a binary variable has low distinguishability, but median discriminability is assessed as only about one-fifth of cases are documented with intensive care (distinguishability = medium risk).

Based on the rating we assessed which variables should be considered "key variables" and need to be modified to protect records from singling out and linkability. Table 2 shows the result of the assessment and the key variables (variables with a score >5) identified. All remaining variables (i.e. non-key variables) are considered sensitive information and will be transformed to protect against inference.

To protect records from singling out, linkability, and inference, the anonymization pipeline used for the NAPKON PUF requires records to satisfy the following criteria to be released: (i) k-anonymity with $k = 11$ for all key variables (i.e. "age at diagnosis", "gender", "quarter first diagnosis", "year first diagnosis", "cohort") and (ii) t-closeness with $t = 0.5$ for sensitive variables (i.e. "mild disease phase", "moderate disease phase", "severe disease phase", "patient status at end of acute phase", "intensive care treatment", "invasive ventilation", "ability to work at 3-month follow-up" and "any symptom at 3-month follow-up"). The variables "availability of 3-month follow-up" and "hospitalization during acute phase" are not explicitly protected by t-closeness, as they are perfectly correlated with other sensitive variables and thus implicitly protected as well. Further, the anonymization pipeline must guarantee that the criteria mentioned above also hold for continuous releases of new data.

Evaluation of statistical properties. We evaluated the impact of the implemented anonymization procedures by comparing descriptive statistics and associations in the dataset before and after anonymization. We computed a regression model for each cohort and, in addition, a regression model that combined data from the SUEP and the HAP to analyze the effects of anonymization for the NAPKON cohort in general. The medical validity of analyses using all cohort data would have been limited due to different cohort populations and recruitment strategies. Additionally, we assessed the impact of anonymization for datasets of different sizes, ranking cases by the date of their diagnosis. In doing so, we intended to simulate the scenario of a continuous data release starting with few cases. The month of recruitment, when cases were included in the NAPKON cohort, can be found on the NAPKON homepage (https://napkon.de).

Descriptive statistics were computed using relative frequencies, and distribution of variables before and after anonymization were compared using Chi-Squared test, defining $P < 0.05$ as statistically significant. The associations were computed using logistic regression models, adjusting for age and gender. The odds ratios and confidence intervals were presented. Cases with unknown/missing data were excluded. To verify the effect of the implemented anonymization methods of the re-identification risk, we computed highest, average, and lowest re-identification risk for the dataset before and after anonymization, again following the approach developed for LEOSS[11]. The risk was calculated based on the sizes of groups of records which are indistinguishable from one another in regard to the attributes which we assume could be used to identify an individual, i.e. key variables.

Data availability

A current version of the NAPKON public dataset is released as a CSV file on the NAPKON website (https://napkon.de/statistik/). In addition to the dataset, the homepage offers to explore the dataset in more detail by presenting regularly updated figures of descriptive statistics. The NAPKON dataset used for these analyses is published on Zenodo[29]. The original NAPKON dataset is available from the NAPKON Use and Access Commitee but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available.

Code availability

The code developed for the NAPKON public dataset is available as open-source software[30].

Received: 27 May 2022; Accepted: 30 August 2022;
Published online: 21 December 2022

References

1. Ahn, D. G. *et al*. Current Status of Epidemiology, Diagnosis, Therapeutics, and Vaccines for Novel Coronavirus Disease 2019 (COVID-19). *J Microbiol Biotechnol* **30**, 313–324 (2020).
2. Bchetnia, M., Girard, C., Duchaine, C. & Laprise, C. The outbreak of the novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2): A review of the current global status. *J Infect Public Health* **13**, 1601–1610 (2020).
3. Sarangi, M. K. *et al*. Diagnosis, prevention, and treatment of coronavirus disease: a review. *Expert Rev Anti Infect Ther* **20**, 243–266 (2022).
4. Schons, M. *et al*. The German National Pandemic Cohort Network (NAPKON): rationale, study design and baseline characteristics. *Eur J Epidemiol* (2022).
5. Naqvi, A. COVID-19 European regional tracker. *Sci Data* **8**, 181 (2021).
6. Berry, I. *et al*. A sub-national real-time epidemiological and vaccination database for the COVID-19 pandemic in Canada. *Sci Data* **8**, 173 (2021).
7. Xu, B. *et al*. Epidemiological data from the COVID-19 outbreak, real-time case information. *Sci Data* **7**, 106 (2020).
8. Publications Office of the European Union. The official portal for European data, https://data.europa.eu/en (2022).
9. Belgian-government. *COVID-19 data sets*, https://data.gov.be/en/dataset/1030d556bc6489a9d1e85994e25d6bd01d53ce6b (2022).
10. Vuong, Q.-H. *et al*. Covid-19 vaccines production and societal immunization under the serendipity-mindsponge-3D knowledge management theory and conceptual framework. *Humanit and Soc Sci Commun* **9**, 22 (2022).
11. Jakob, C. E. M., Kohlmayer, F., Meurers, T., Vehreschild, J. J. & Prasser, F. Design and evaluation of a data anonymization pipeline to promote Open Science on COVID-19. *Sci Data* **7**, 435 (2020).
12. Vuong, Q. H. Reform retractions to make them more transparent. *Nature* **582**, 149 (2020).
13. Heatherly, R., Denny, J. C., Haines, J. L., Roden, D. M. & Malin, B. A. Size matters: how population size influences genotype-phenotype association studies in anonymized data. *J Biomed Inform* **52**, 243–250 (2014).
14. Jakob, C. E. M. *et al*. First results of the "Lean European Open Survey on SARS-CoV-2-Infected Patients (LEOSS)". *Infection* **49**, 63–73 (2021).
15. Daries, J. P. *et al*. Privacy, Anonymity, and Big Data in the Social Sciences. *Commun ACM* **57**, 56–63 (2014).
16. Xu, H. & Zhang, N. Implications of Data Anonymization on the Statistical Evidence of Disparity. *Manag Sci* **0** (2021).
17. WHO Working Group on the Clinical Characterisation and Management of COVID-19 infection. A minimal common outcome measure set for COVID-19 clinical research. *Lancet Infect Dis* **20**, e192–e197 (2020).

18. Vuong, Q. H. The (ir)rational consideration of the cost of science in transition economies. *Nat Hum Behav* **2**, 5 (2018).
19. COVID-19 registry of the LMU Munich. *CORKUM - DRKS00021225*, https://www.drks.de/drks_web/navigate.do?navigationId=trial.HTML&TRIAL_ID=DRKS00021225 (2020)
20. Kurth, F. *et al*. Studying the pathophysiology of coronavirus disease 2019: a protocol for the Berlin prospective COVID-19 patient cohort (Pa-COVID-19). *Infection* **48**, 619–626 (2020).
21. Horn, A. *et al*. Long-term health sequelae and quality of life at least 6 months after infection with SARS-CoV-2: design and rationale of the COVIDOM-study as part of the NAPKON population-based cohort platform (POP). *Infection* **49**, 1277–1287 (2021).
22. Sass, J. *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* **20**, 341 (2020).
23. Thye, A. Y. *et al*. Psychological Symptoms in COVID-19 Patients: Insights into Pathophysiology and Risk Factors of Long COVID-19. *Biology (Basel)* **11** (2022).
24. Yelin, D. *et al*. Long-term consequences of COVID-19: research needs. *Lancet Infect Dis* **20**, 1115–1117 (2020).
25. Huang, C. *et al*. 6-month consequences of COVID-19 in patients discharged from hospital: a cohort study. *Lancet* **397**, 220–232 (2021).
26. Zhan, Y. *et al*. SARS-CoV-2 immunity and functional recovery of COVID-19 patients 1-year after infection. *Signal Transduct Target Ther* **6**, 368 (2021).
27. Pilgram, L. *et al*. The COVID-19 Pandemic as an Opportunity and Challenge for Registries in Health Services Research: Lessons Learned from the Lean European Open Survey on SARS-CoV-2 Infected Patients (LEOSS). *Gesundheitswesen* **83**, S45–S53 (2021).
28. Malin, B., Loukides, G., Benitez, K. & Clayton, E. W. Identifiability in biobanks: models, measures, and mitigation strategies. *Hum Genet* **130**, 383–392 (2011).
29. NAPKON Public Use File. *Zenodo* https://doi.org/10.5281/zenodo.6576177 (2022).
30. NAPKON Public Use File Version 1.0.0. *Zenodo* https://doi.org/10.5281/zenodo.6576533 (2022).

Acknowledgements

The study was carried out using the clinical-scientific infrastructure of NAPKON (Nationales Pandemie Kohorten Netz, German National Pandemic Cohort Network) of the Network University Medicine (NUM), funded by the Federal Ministry of Education and Research (BMBF). We gratefully thank the NAPKON Study Group that is composed of the representatives of the NAPKON sites that contributed 5 per mille to the analysis (NAPKON Study Site Group, alphabetical order), of the NAPKON Infrastructure Group, of the NAPKON Steering Committee, and of the NAPKON Use and Access Committee. The project National Pandemic Cohort Network (NAPKON) is part of the Network University Medicine (NUM) and was funded by the German Federal Ministry of Education and Research (BMBF) (FKZ: 01KX2021). Parts of the infrastructure of the Würzburg study site were supported by the Bavarian Ministry of Research and Art to support Corona research projects. Parts of the NAPKON project suite and study protocols of the Cross-Sectoral cohort platform are based on projects funded by the German Center for Infection Research (DZIF). Parts of the infrastructure for the Population-Based Platform received funding of the State of Schleswig Holstein (COVIDOM) and DFG Exzellenzcluster. The Open Access publication was supported by the DEAL project.

Author contributions

Carolin E.M. Koll and Sina M. Hopff contributed equally to this analysis (shared-first). Due to the translational content of this analysis, Fabian Prasser and Jörg Janne Vehreschild share the last authorship (shared-last). C.S., J.E., S.H., C.T., J.H., I.V., L.R., J.R., L.K., were the main contributors for the analyzed NAPKON cohort. J.V., M.K., S.J., O.M., L.M., M.H., S.S., F.S., S.S., T.B., S.F., P.M., S.H., C.K., C.L. were representatives of the NAPKON infrastructure. S.H., C.K., C.L., O.M., M.K., S.J. developed and revised the data computation. C.L. and C.K. extracted the data. F.B., C.K., S.H., C.L., T.M. developed the anonymization pipeline. C.L. and F.P. programmed the anonymization pipeline. S.H. and C.K. performed the evaluation analyses of this manuscript. C.K., S.H., F.P., J.V., T.M., C.L. interpreted the results. C.K., S.H. and T.M. drafted the manuscript. J.V., F.P., J.R., O.M., M.K., S.J., L.K. revised the manuscript critically for important intellectual content. All authors revised and approved the final version of the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to C.E.M.K.

NAPKON Study Group

NAPKON Infrastructure Group I. Bernemann[19], T. Illig[19], M. Kersting[19], N. Klopp[19], V. Kopfnagel[19], S. Muecke[19], G. Anton[20], M. Kraus[20], A. Kuehn-Steven[20], S. Kunze[20], M. K. Tauchert[20], J. Vehreschild[21], M. Brechtel[22], S. Fuhrmann[22], S. M. Hopff[22], C. E. M. Koll[22], C. Lee[22], L. Mitrov[22], S. M. Nunes de Miranda[22], M. Nunnendorf[22], G. Sauer[22], K. Seibel[22], M. Stecher[22], K. Appel[23], R. Geisler[23], M. Hagen[23], M. Scherer[23], J. Schneider[23], C. Weismantel[23], B. Balzuweit[24], S. Berger[24], M. Hummel[24], S. Schmidt[24], M. Witzenrath[24], T. Zoller[24], A. Krannich[24], F. Kurth[24], J. Lienau[24], R. Lorbeer[24], C. Pley[24], J. Schaller[24], C. Thibeault[24], C. Bauer[25], C. Fiessler[25], M. Goester[25], A. Grau[25], P. Heuschmann[25], A. L. Hofmann[25], S. Jiru-Hillmann[25], K. Kammerer[25], M. Kohls[25], O. Miljukov[25], J. P. Reese[25], K. Ungethuem[25], M. Krawczak[26], J. C. Hellmuth[27], T. Bahls[28], W. Hoffmann[28], M. Nauck[28], C. Schäfer[28], M. Schattschneider[28], D. Stahl[28], H. Valtentin[28], I. Chaplinskaya[29], S. Hanß[29], D. Krefting[29], C. Pape[29] & J. Hoffmann[30]

NAPKON Study Site Group J. Fricke[24], T. Helbig[24], M. Hummel[24], T. Keil[24], L. Kretzler[24], L. Krist[24], L. Lippert[24], M. Mittermaier[24], M. Mueller-Plathe[24], M. Roennefarth[24], L. E. Sander[24], S. Schmidt[24], F. Steinbeis[24], S. Steinbrecher[24], D. Treue[24], P. Triller[24], M. Witzenrath[24], T. Zoller[24], S. Zvorc[24], F. Hammer[31], L. Horvarth[31], A. Kipet[31], M. Schroth[31], M. T. Unterweger[31], I. Bernemann[32], N. Drick[32], M. Hoeper[32], T. Illig[32], M. Kersting[32], N. Klopp[32], V. Kopfnagel[32], I. Pink[32], M. Ratowski[32], F. Zetzsche[32], C. M. Bremer[33], H. H. Halfar[33], S. Herold[33], L. H. Nguyen[33], C. Ruppert[33], M. Scheunemann[33], W. Seeger[33], A. Uribe Munoz[33], I. Vadasz[33], M. Wessendorf[33], H. Azzaui[34], M. Gräske[34], M. Hower[34], J. Kremling[34], E. Landsiedel-Mechenbier[34], A. Riepe[34], B. Schaaf[34], S. Frank[27], J. C. Hellmuth[27], M. Huber[27], S. Kaeaeb[27], O. T. Keppler[27], E. Khatamzas[27], C. Mandel[27], S. Mueller[27], M. Muenchhoff[27], L. Reeh[27], C. Scherer[27], H. Stubbe[27], M. von Bergwelt[27], L. Weiß[27], B. Zwißler[27], M. Milovanovic[35], R. Pauli[36], M. Ebert[37], W. K. Hofmann[37], M. Neumaier[37], F. Siegel[37], A. Teulfel[37], C. Wyen[38], C. Allerlei[39], A. Keller[40], J. Walter[40], R. Bals[41], C. Herr[41], M. Krawczyk[41], C. Lensch[41], P. M. Lepper[41], M. Riemenschneider[41], S. Smola[41], M. Zemlin[41], C. Raichle[42], G. Slesak[42], S. Bader[43], J. Classen[43], C. Dhillon[43], M. Freitag[43], V. Gruenherz[43], B. Maerkl[43], H. Messmann[43], C. Roemmele[43], M. Steinbrecher[43], M. Ullrich[43], H. Altmann[44], R. Berner[44], S. Dreßen[44], T. Koch[44], D. Lindemann[44], K. Seele[44], P. Spieth[44], K. Tausche[44], N. Toepfner[44], S. von Bonin[44], D. Kraska[45], A. E. Kremer[45], M. Leppkes[45], J. Mang[45], M. F. Neurath[45], H. U. Prokosch[45], J. Schmid[45], M. Vetter[45], C. Willam[45], K. Wolf[45], M. Addo[46], A. L. F. Engels[46], D. Jarczak[46], M. Kerinn[46], S. Kluge[46], R. Kobbe[46], K. Roedl[46], C. Schlesner[46], P. Shamsrizi[46], T. Zeller[46], C. Arendt[23], C. Bellinghausen[23], S. Cremer[23], A. Groh[23], A. Gruenewaldt[23], Y. Khodamoradi[23], S. Klinsing[23], G. Rohde[23], M. Vehreschild[23], T. Vogl[23], K. Becker[47], M. Doerr[47], K. Lehnert[47], M. Nauck[47], N. Piasta[47], C. Schaefer[47], E. Schaefer[47], M. Schattschneider[47], C. Scheer[47], D. Stahl[47], R. Baber[48], S. Bercker[48], N. Krug[48], S. D. Mueller[48], H. Wirtz[48], G. Boeckel[49], J. A. Meier[49], T. Nowacki[49], P. R. Tepasse[49], R. Vollenberg[49], C. Wilms[49], A. Arlt[50], F. Griesinger[50], U. Guenther[50], A. Hamprecht[50], K. Juergens[50], A. Kluge[50], C. Meinhardt[50], K. Meinhardt[50], A. Petersmann[50], R. Prenzel[50], A. Brauer-Hof[51], C. Brochhausen-Delius[51], R. Burkhardt[51], M. Feustel[51], F. Hanses[51], M. Malfertheiner[51], T. Niedermair[51], B. Schmidt[51], P. Schuster[51], S. Wallner[51], D. Mueller-Wieland[52], N. Marx[52], M. Dreher[52], E. Dahl[52], J. Wipperfuerth[52], T. Bahmer[53], J. Enderle[53], A. Friedrichs[53], A. Hermes[53], N. Kaeding[53], M. Koerner[53], M. Krawczak[53], C. Kujat[53], I. Lehmann[53], M. Lessing[53], W. Lieb[53], C. Maetzler[53], M. Oberländer[53], D. Pape[53], M. Plagge[53], L. Reinke[53], J. Rupp[53], S. Schreiber[53], D. Schunk[53], L. Tittman[53], W. Barkey[54], J. Erber[54], L. Fricke[54], J. Lieb[54], T. Michler[54], L. Mueller[54], J. Schneider[54], C. Spinner[54], F. Voit[54], C. Winter[54], M. Bitzer[55], S. Bunk[55], S. Göpel[55], H. Häberle[55], K. Kienzle[55], H. Mahrhofer[55], N. Malek[55], P. Rosenberger[55], C. Struemper[55], F. Trauner[55], S. Frantz[56], A. Frey[56], K. Haas[56], C. Haertel[56], K. G. Haeusler[56], G. Hein[56], J. Herrmann[56], A. Horn[56], N. Isberner[56], R. Jahns[56], M. Kohls[56], J. Liese[56], P. Meybohm[56], C. Morbach[56], J. Schmidt[56], P. Schulze[56], S. Stoerk[56], B. Weissbrich[56], F. Brinkmann[57], Y. Brueggemann[57], T. Gambichler[57], K. Hellwig[57], T. Luecke[57], A. Reinacher-Schick[57], W. E. Schmidt[57], C. Schuette[57], E. Steinmann[57], C. Torres Reyes[57], K. Alsaad[58], B. Berger[58], E. Hamelmann[58], H. Heidenreich[58], C. Hornberg[58], N. S. A. Kulamadayil-Heidenreich[58], P. Maasjosthusmann[58], A. Muna[58], C. Olariu[58], B. Ruprecht[58], J. Schmidt[58], C. Stellbrink[58], J. Tebbe[58], D. August[59], M. Barrera[59], V. Goetz[59],

A. Imhof[59], S. Koch[59], A. Nieters[59], G. Peyerl-Hoffmann[59], S. R. Rieg[59], A. Amanzada[29], S. Blaschke[29], A. Hafke[29], G. Hermanns[29], M. Kettwig[29], O. Moerer[29], S. Nussbeck[29], J. Papenbrock[29], M. Santibanez-Santana[29], S. Zeh[29], S. Dolff[60], C. Elsner[60], A. Krawczyk[60], R. J. Madel[60], M. Otte[60], L. Brochhagen[60] & O. Witzke[60]

NAPKON Steering Committee S. Herold[61], P. Heuschmann[25], R. Heyder[24], W. Hoffmann[28], H. Neuhauser[62], S. Schreiber[63], J. Vehreschild[21], M. von Lilienfeld-Toal[64], T. Illig[32] & M. Witzenrath[24]

NAPKON Use & Access Committee S. Blaschke[29], C. Ellert[65], A. Friedrichs[63], P. Heuschmann[25], P. Meybohm[56], K. Milger[27], A. Petersmann[50], G. Schmidt[54], S. Schreiber[63], J. Vehreschild[21], M. von Lilienfeld-Toal[64], O. Witzke[66], S. Frank[27], T. Illig[32] & M. Witzenrath[24]

[19]Hannover Unified Biobank, Hannover Medical School, Hannover, Germany. [20]Institute of Epidemiology, Helmholtz Center Munich, Munich, Germany. [21]University Hospital Cologne and University Hospital Frankfurt, Cologne and Frankfurt, Germany. [22]University Hospital Cologne, Cologne, Germany. [23]University Hospital Frankfurt, Frankfurt am Main, Germany. [24]Charité - Universitaetsmedizin Berlin, Berlin, Germany. [25]University of Wuerzburg, Wuerzburg, Germany. [26]University of Kiel, Kiel, Germany. [27]University Hospital LMU Munich, Munich, Germany. [28]University Medicine Greifswald, Greifswald, Germany. [29]University Medical Center Goettingen, Goettingen, Germany. [30]German Center for Cardiovascular Diseases (DZHK), Berlin, Germany. [31]Cnopfsche Children's Hospital, Nuernberg, Germany. [32]Hannover Medical School, Hannover, Germany. [33]Justus Liebig University Giessen & Marburg, Gießen and Marburg, Germany. [34]Klinikum Dortmund, Dortmund, Germany. [35]Malteser Hospital St. Franziskus Hospital, Flensburg, Germany. [36]MVZ am Isartor, Muenchen, Germany. [37]Medical Faculty Mannheim, Mannheim, Germany. [38]Practice for general medicine Am Ebertplatz, Cologne, Germany. [39]Practice for general medicine Dr. Allerlei, Frankfurt am Main, Germany. [40]Saarland University, Homburg, Germany. [41]Saarland University Hospital, Homburg, Germany. [42]Tropical Hospital Paul-Lechler-Krankenhaus, Tuebingen, Germany. [43]University Hospital Augsburg, Augsburg, Germany. [44]University Hospital Carl Gustav Carus, Dresden, Germany. [45]University Hospital Erlangen, Erlangen, Germany. [46]University Hospital Hamburg-Eppendorf, Hamburg, Germany. [47]University Hospital Greifswald, Greifswald, Germany. [48]University Hospital Leipzig, Leipzig, Germany. [49]University Hospital Muenster, Muenster, Germany. [50]University Medicine Oldenburg, Oldenburg, Germany. [51]University Hospital Regensburg, Regensburg, Germany. [52]University Hospital RWTH Aachen, Aachen, Germany. [53]University Hospital Schleswig-Holstein, Kiel and Luebeck, Germany. [54]University Hospital Technical University Munich, Munich, Germany. [55]University Hospital Tuebingen, Tuebingen, Germany. [56]University Hospital Wuerzburg, Wuerzburg, Germany. [57]University Hospitals of the Ruhr University Bochum, Bochum, Germany. [58]Bielefeld University, Medical School and University Medical Center East Westphalia-Lippe, Bielefeld, Germany. [59]University Medical Center Freiburg, Freiburg, Germany. [60]University Medicine Essen, Essen, Germany. [61]University Hospital Giessen and Marburg, Giessen, Germany. [62]Robert Koch Institute, Department of Epidemiology and Health Monitoring, Berlin, Germany. [63]University Hospital Schleswig-Holstein, Kiel, Germany. [64]Jena University Hospital, Jena, Germany. [65]Lahn-Dill-Clinics, Wetzlar, Germany. [66]University Hospital Essen, Essen, Germany.