PREPRINT

Preservation of individuals' privacy in shared COVID-19 related data

Stefan Sauermann[1], Chifundo Kanjala[2,§], Matthias Templ[3], Claire C. Austin[4]; and the RDA-COVID19-WG[5]

[1]UAS Technikum Wien, [2]London School of Hygiene and Tropical Medicine, [3]Zurich University of Applied Sciences, [4]Environment and Climate Change Canada, [5]This work was developed as part of the Research Data Alliance RDA-COVID19-WG Recommendations and guidelines on data sharing, and the RDA COVID-19 Epidemiology WG Data sharing in epidemiology, and we acknowledge the support provided by the RDA community and structures.

All views and opinions expressed are those of the co-authors, and do not necessarily reflect the official policy or position of their respective employers, or of any government, agency or organization.

[§] Corresponding author: chifundo.kanjala@lshtm.ac.uk

CITE AS: Sauermann S, Kanjala[§] C, Templ M, Austin CC; and the RDA-COVID19-WG (2020). Preservation of individuals' privacy in shared COVID-19 related data. In *COVID-19 Data sharing in epidemiology, version 0.054*. Research Data Alliance RDA-COVID19-Epidemiology WG. https://doi.org/10.15497/rda00049

ABSTRACT

This paper provides insight into how restricted data can be incorporated in an open-be-default-by-design digital infrastructure for scientific data. We focus, in particular, on the ethical component of FAIRER (Findable, Accessible, Interoperable, Ethical, and Reproducible) data, and the pseudo-anonymization and anonymization of COVID-19 datasets to protect personally identifiable information (PII). First we consider the need for the customisation of the existing privacy preservation techniques in the context of rapid production, integration, sharing and analysis of COVID-19 data. Second, the methods for the pseudo-anonymization of direct identification variables are discussed. We also discuss different pseudo-IDs of the same person for multi-domain and multi-organization. Essentially, pseudo-anonymization and its encrypted domain specific IDs are used to successfully match data later, if required and permitted, as well as to restore the true ID (and authenticity) in individual cases of a patient's clarification. Third, we discuss the application of statistical disclosure control (SDC) techniques to COVID-19 disease data. To assess and limit the risk of re-identification of individual persons in COVID-19 datasets (that are often enriched with other covariates like age, gender, nationality, etc.) to acceptable levels, the risk of successful re-identification by a combination of attribute values must be assessed and controlled. This is done using statistical disclosure control for anonymization of data. Lastly, we discuss the limitations of the proposed techniques and provide general guidelines on using disclosure risks to decide on appropriate modes for data sharing to preserve the privacy of the individuals in the datasets.

KEYWORDS:
Pseudo-anonymization, statistical disclosure control, SDC, data anonymization, data sharing, privacy, personally identifiable information, PII, COVID-19, open science.

BACKGROUND

It is imperative that COVID-19 response efforts be coordinated at a global scale to ensure that no part of the world is left behind and serve as a reservoir for the virus that would spread to regions that had successfully brought it under control. Critical to this globally coordinated approach are the requirements that the data relating to the pandemic are rapidly collected, processed, and shared. Data sharing reduces duplication of effort, increases efficiency, and facilitates transparency thus maximising the return on effort and funding.

Recent years have made advances in addressing the complex and interconnected data demands in a new world of open science, Big Data, and artificial intelligence (NIST 2019; NASEM 2019; Wellcome Trust 2017; European Commission 2020). As a result, unprecedented data sharing led to faster-than-ever outbreak research on SARS-CoV-2 before it resulted in the COVID-19 pandemic (Le Guillou 2020). However, *open science by-default-by-design* in a fully digital framework has not yet been achieved across disciplines and organizations. As a result, from the standpoint of data management and data sharing which underlie detection, response, and implementation of decisions to prevent and subsequently manage the COVID-19 pandemic, our systems were not prepared to meet the challenge (Austin and Widyastuti 2020). Open data, data security, and ethical considerations such as privacy, need to be built into the data system from the start to produce FAIRER (Findable, Accessible, Interoperable, Reusable, Ethical, Reproducible) data. To better respond to the COVID-19 crisis, we need a broad range of FAIRER open data that input seamlessly into certified digital repositories, Big Data analytical workflows and Artificial Intelligence (AI), and that are reusable beyond their original purpose by unimagined systems.

Individuals have a right to privacy in health care, and the state has the right to override an individual's right to privacy in cases of serious public health risks if revealing private medical information helps to protect public health (Upshure et al. 2005). Before such a situation arises, a balanced approach will have already put in place an *open science by-default--by-design* digital infrastructure for all data, including confidential data that would be subject to tiered access by appropriately credentialed people and machines. In addition to encryption, there are several ways to protect restricted data in the system. These include, for example: (a) Secure model / query servers that integrate differential privacy perturbations to statistics generated/queried from the user; (b) Secure environments with already (pre-tabulated) noise on individuals (cell key method), for example, microdata perturbation before "queries" or pre-calculation of all possible queries (data cubes) and suppression or perturbation of cells, so that cells with low frequencies are protected; and, (c) Microdata sharing that must be anonymised beforehand using SDC methods. The resulting anonymized microdata might be shared for a broader audience. It is essential that security measures be frictionless so that they do not place a burden on users.

Even during a pandemic such as COVID-19, with such a system in place, rapid collection, processing, analysis and sharing of the data need not come at the expense of ethical practices. The privacy and confidentiality of the individuals supplying these much needed data need to be preserved throughout the entire process, whether the data come from a healthcare facility, a community-based study, or some surveillance mechanism. While the technology exists to

address these issues, it is not clear how it could be implemented for the task of preserving data privacy during the pandemic which is requiring rapid data production, sharing and analysis. This task is further complicated by the need to link datasets from disparate sources. The RDA recommendation calls for processes that involve anonymisation and pseudonymisation of data.

The present paper summarises use cases deriving from the RDA COVID-19 recommendation, and illustrates how existing methods and technology can be harnessed and implemented. We also consider how the privacy preservation methods can work within the context of a Common Data Model (CDM) and full spectrum epidemiology proposed by Greenfield et al. (2020). The CDM provides a set of specifications for integrating silos of COVID-19 related data from hospital/clinic surveillance Electronic Health Records (EHR), field-based longitudinal demographic and epidemiological surveillance, population level indicators, air quality monitoring, etc.). In the present paper, we describe how the existing privacy preservation technologies and methods can be customised for implementation in the context of the model proposed by Greenfield et al. (2020) to ensure pseudonymisation and anonymisation.

METHODS

A privacy preservation strategy is presented that combines pseudo-anonymization and statistical disclosure control techniques within the context of bringing together COVID-19 data from diverse sources into a common data model (Greenfield et al. 2020). The model proposed by Greenfield et al (2020) is flexible and caters for the entire gamut of settings from high-income countries (HIC) to low- and middle-income countries (LMIC) settings. In this context, the applied techniques need to be flexible enough to take into account diversity of data sources. In addition, they ought not presume universal health coverage, complete national disease registries or optimal disease surveillance. We therefore present customizations of existing anonymization techniques to accommodate the Covid - 19 data from both siloed and fragmented systems on one hand and also those from the better connected systems on the other.

Further, given that the common data model integrates diverse data sources, we are proposing an approach that iteratively assesses and controls disclosure risks at each stage of data integration. First, when the data are harvested from their sources and also when they are integrated into harmonised databases. Our approach seeks to balance between identity disclosure control and minimising data loss within the context of a common data model for COVID-19 data.

RESULTS

Pseudonymization and anonymization

To ensure privacy, both pseudonymization of direct identifiers (e.g. patient specific ID's) and anonymization of indirect identifiers (e.g. socio-demographic information on individuals) must be applied. Pseudonymization can be done e.g. by first salting the ID's using, for example, the SHA-3 hash function (NIST 2017) and 256 bit salts, and then applying a hash function to the salted ID's for encryption. Different domains-specific salts for the same patient can be used to

calculate domain specific identifiers. Domains-specific salts can be managed by a central agency that stores the salts and is able to link data sets from different domains from the same patients. An appropriate IT-infrastructure is relevant here for ID management and linking data. This article introduces basic concepts that can be used to sustain privacy of patient related data in epidemiologic, research and administration contexts.

Pseudo-anonymization

Health data are recorded and used in different domains, for example medicine, communities, research, administration, and statistics. Person specific electronic identifiers (IDs) link specific information to the individual person records in each domain. Each domain assigns a domain-specific ID (dsID) to each individual. In order to satisfy privacy requirements on direct identifiers, data that carry a dsID in clear text remain within the data holder's secure server and are not shared.

If data are re-used, for example in research or public health purposes, the dsID is removed (anonymization) or replaced by a different ID (pseudonym). Pseudonyms can later be traced back by an authorized person to the original dsID. This must only occur under well-defined conditions, e.g. if a researcher needs to clear ambiguities in incoming information together with the organisation that generated the data.

In multi-domain and multi-organization scenarios such as when data from EHR are brought together with population demographic and epidemiological surveillance in a common data model (Greenfield et al. 2020), consistent management of IDs and pseudonyms on a secure authorized server is needed to enable cross-linking of data from different sources using the domain specific ID's of patients.

The following requirements apply:

- Domain specific IDs (dsIDs) exist for each individual in each domain;
- The dsID of an individual must not leave the domain in clear text, to prevent unintended record linkage between domains;
- When providing data from a source domain to a target domain, the target dsID must become available to users in the target domain; and,
- Appropriate dsID IT infrastructure must be available to enable domains to share and link data, while preventing unintended record linkage.

There are different concepts for pseudo-anonymisation:

- Hashing with key and salt: hash functions (incorporating SHA-3 (Bertoni et al. 2010), for example) whose output depends not only on the input but on a secret key too. Depending on the key, different pseudonyms can be produced for the same ID. By salting, the input to the hash function is being augmented via adding a predefined string (the "salt"). For the same identifier, several different pseudonyms can be produced, according to the choice of the salt.

- Encryption: a key used within a encryption algorithm provides a cipher text that is to be used as a pseudonym (Daemen and Rijmen 2002) . The same secret key is needed for the decryption.

The main difference between encryption and keyed hash functions is that the secret key owner (the source domain) must always store data subjects' initial identifiers, and can identify the pseudonyms through a simple decryption process. The key and the original data (initial values for the ID's) thus really need to be stored safely at the source domain. Randomness might be added to the key, e.g. adding time stamps (see below). If the key is lost, it is not trivial and hardly possible to identify pseudonyms even on the source domain. A major difference to encryption is the use of hash functions. Using Hash functions the data controllers (source domain) does not need to store the initial identifiers for tracking the individuals.

Salting and encryption of identifiers in a multi-domain and multi-organization pseudonymization strategy

To achieve pseudonymisation between different domains, salting and hashing or encryption can be applied. The latter is shown in the following. Encryption assures that no dsID leaves the domain as clear text. Figure 1A describes how data from a health domain can be shared to be linked to records in a research domain and an administration domain in this way. Figure 1B introduces the IT infrastructure needed to map dsIDs between different domains. Mapping can for example be done via lookup tables, or by algorithms that calculate the desired target dsID. Different scenarios can be used when building IT infrastructures: In a strictly centralised scenario the ID mapping server may be provided by local authorities, and conform to legal requirements of disease control. In a more distributed scenario, patients and healthy study participants may use the server to configure the cross-mapping of dsIDs to match their individual privacy needs. Patients may e.g. decide to allow re-use of their data for disease control by administrations, but not for research.

As an organization receives multiple messages over time, it might link the data using even the encrypted dsIDs. This will be possible, if the encryption process generates the same dsIDs for each person and each domain over time. This can be prevented by using the concept of salting. A salt is added to an attribute before encrypting, using a newly generated random value as a salt each time that the attribute is being encrypted. This assures that no two encrypted dsIDs are the same, even if they relate to the same individual in the same domain. Figure 1C shows how salting is applied to the dsIDs, as they are mapped between domains while preserving pseudonymization.

In Austria, for example, eGovernment legislation and IT infrastructures are in place to handle domain specific identifiers e.g. for health care, traffic, taxes, and statistics (Austria 2013, 2018). This is implemented and in operation for example in the Austrian electronic health care record ELGA (ELGA GmbH 2020).

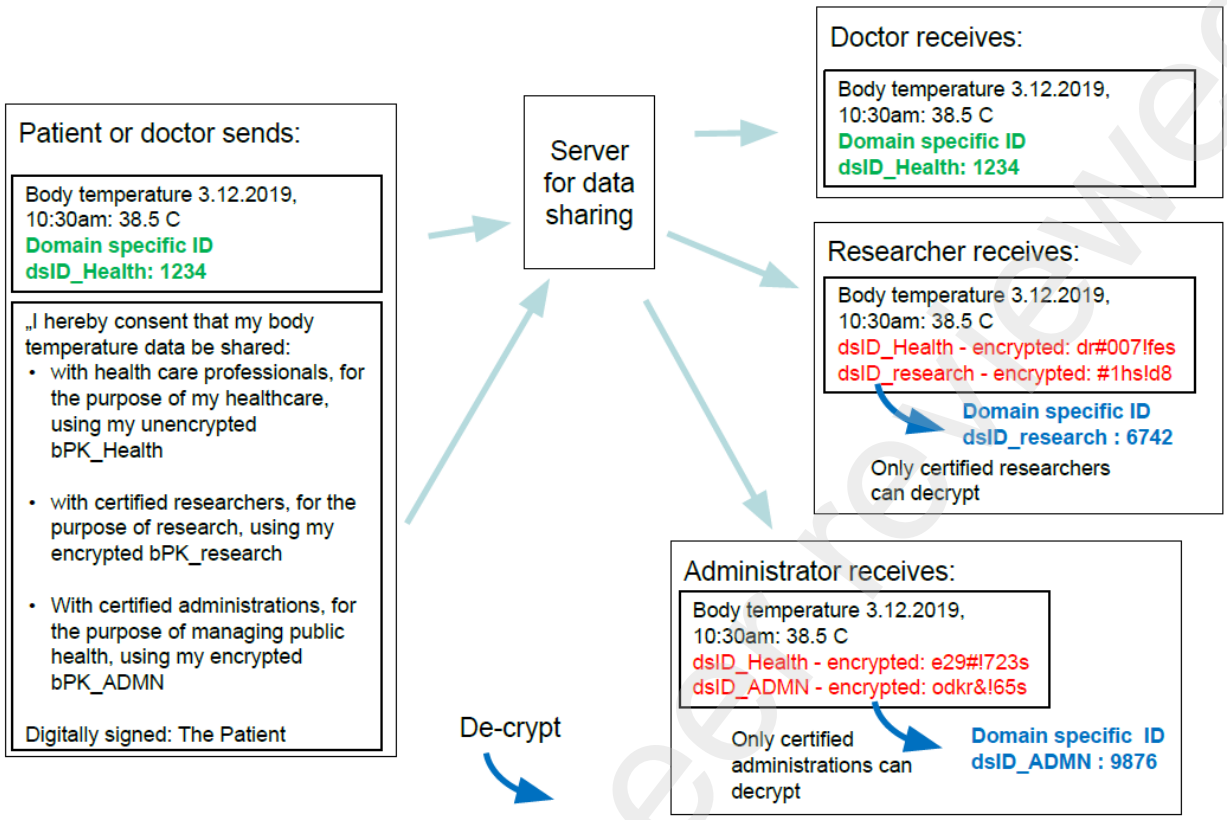

Figure 1A. Sharing or linking a body temperature observation from the healthcare domain with a research and administration domain. In the healthcare domain ID (dsID_Health, green), a patient is identified with a specific ID, (1234, colour green, denoting that it is unencrypted). A doctor will receive this data together with the original ID, as the law allows doctors to share IDs unencrypted. The doctor can attach an encrypted ID (dsID_research - #1hsId8, red, denoting it is encrypted) to the data. The doctor never has access to the unencrypted dsID research ID. A researcher who receives the data decrypts the encrypted ID. This decrypted ID (dsID_research - 6742, blue, denoting that it is a pseudonym) is specific to the research domain. The same method applies as data are provided to administrations, e.g. for public health purposes.

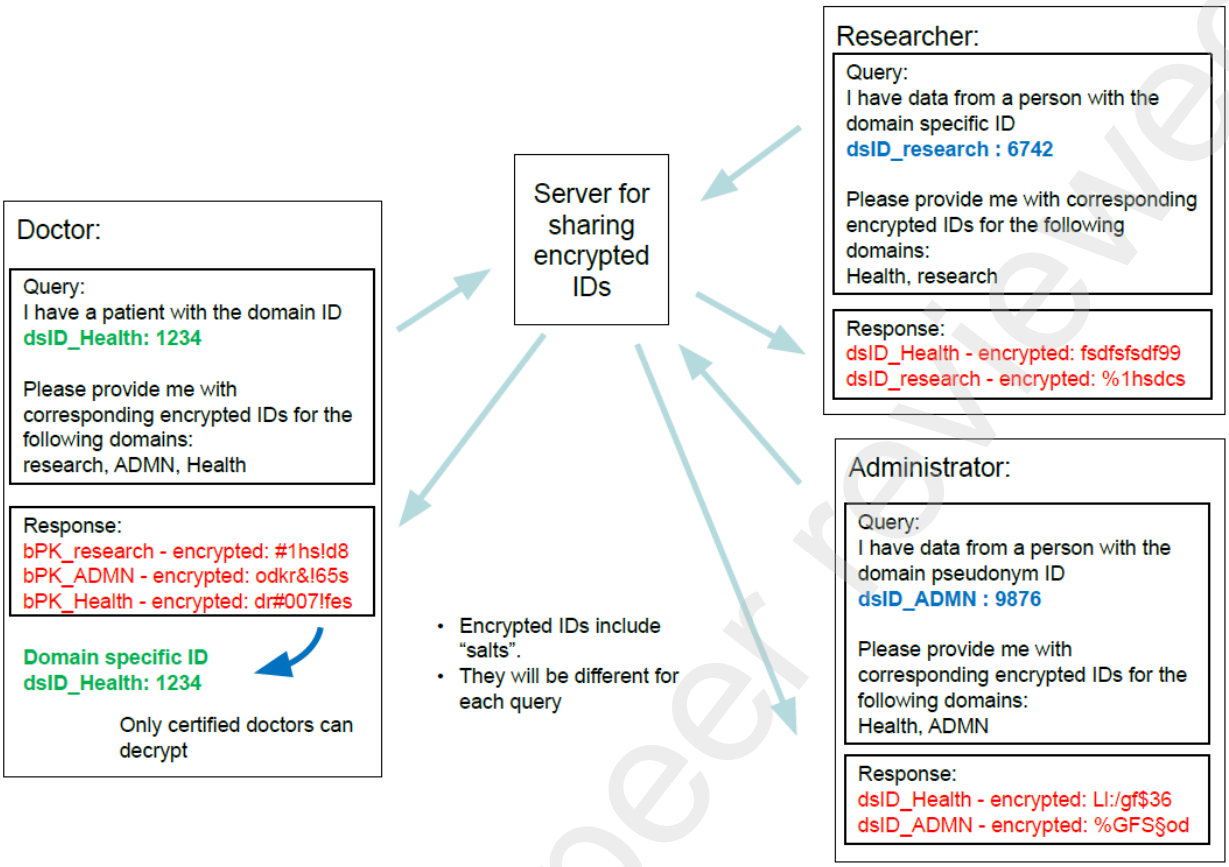

Figure 1B. IT infrastructure for cross-domain IDs management. A user in the medical domain queries the IT service using the dsID of the health domain, asking for encrypted dsIDs of other domains. The service responds with the encrypted dsIDs. Users in other domains can use the same mechanism. This enables users in different domains to co-operate: For example the researcher can attach the encrypted dsID_Health ID to a message to the doctor, asking for details to support clearing ambiguities in the data the doctor provided earlier. The doctor can then decrypt the patient ID, access the patient related information in the health IT system, and finalise the clearing with the researcher.

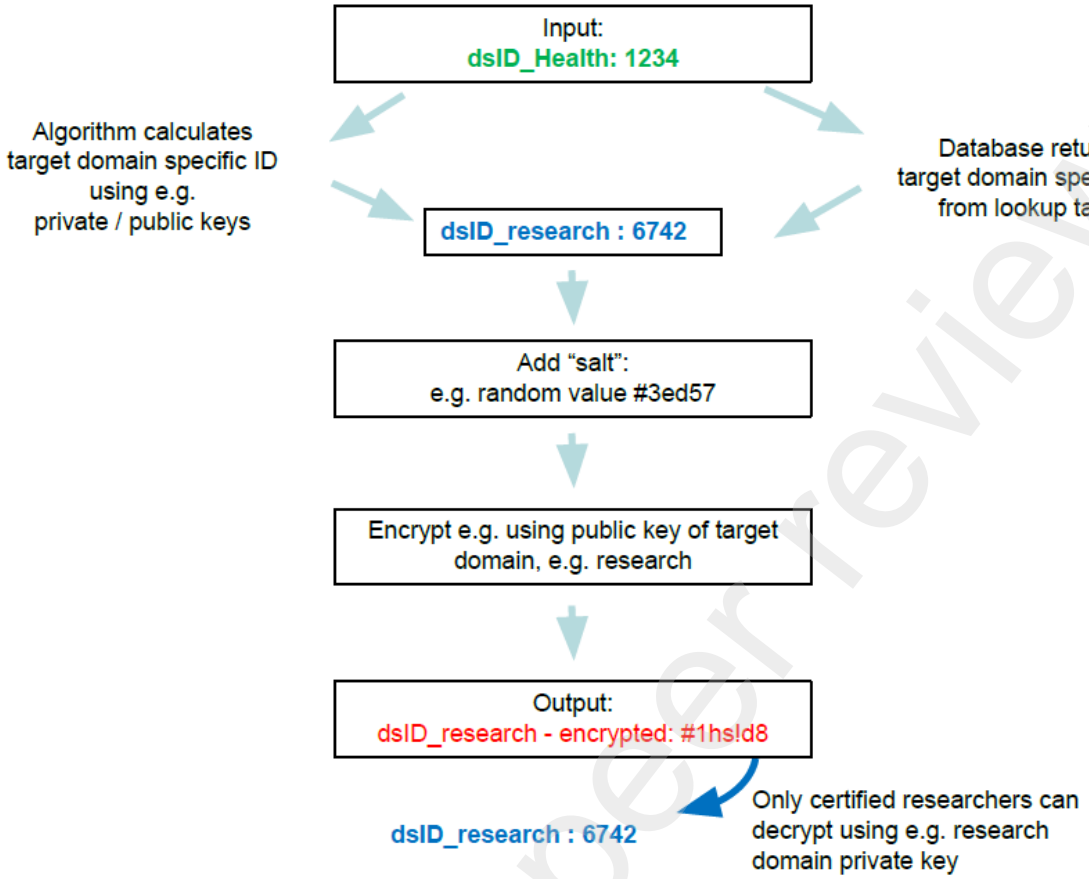

Figure 1C. Data flow for deriving an encrypted ID (#1hs!d8) for a target domain (research). The source ID (1234) can be mapped to the target identifier for example in two ways. A mathematical algorithm is used (left branch) to calculate the target domain ID (6742), or a database query returns the ID. A time stamp (as kind of salting) is then attached to this ID. ID and timestamp together are then encrypted, e.g. using the public key of the target domain. This assures that no two encrypted IDs for the same patient and the same domain are identical, in this way preventing the unintended linking of records

Anonymization: Controlling statistical disclosure risk

A pseudonymised data set does not imply that individuals cannot be re-identified. The dataset still needs to be anonymized using different methods other than those for pseudo-anonymization. Intruders can reveal patients' identities using a combination of indirect identifiers such as gender, age, education, location, race/ethnicity, etc.

Since the first data protection scandals in the 1990s (Barth-Jones 2012), we have learned that removing or pseudo-anonymization of directly identifying attributes (ID's) such as names, addresses and social insurance numbers is generally not sufficient to prevent data protection violations. Beyond masking these direct identifiers, we need to control statistical disclosure risk. This is the risk of intruders using a combination of indirect identifiers such as race/ethnicity,

education level, sex, age, etc., to identify individuals and their their health status. The levels of such risks need to be quantified, accessed, and controlled in the shared datasets to ensure privacy and trust of the individuals contributing their data and to meet scientific integrity requirements.

The anonymization methods differ in terms of anonymization of microdata (each row corresponds to an individual), anonymization of tabular data (aggregated information) and anonymization of queried information from a database which stores original, non-anonymized data.

COVID-19 microdata. To measure the risk of re-identification of individuals/persons, frequencies of persons with certain attributes are generally counted and evaluated. The lower the frequency, the higher the risk. For example, consider in a COVID-19 cases dataset exists an 83-year-old man with COVID-19 in a certain community. If he is the only 83-year-old man in the community, then this man can easily re-identified and people living in the community know that he has (had) COVID-19 (neighbor matching scenario) or intruders can match COVID-19 data with, e.g., a voters database including the names and addresses of people, gender and age. In this case, the match is certainly successful and unique and the re-identification is successful. To make this person anonymous, various methods can be used. For example, post-randomization of the place of residence (Gouweleeuw et al. 1998), or adding a noise to the actual age, or regrouping age in age groups. After the anonymization the data utility should be checked. While the risk for re-identification should be very low, the data utility should be as high as possible, i.e. the data quality of the anonymized data should be very high.

In the context of integration of data from various sources, the steps outlined in the previous paragraph will need to be carried out on the individual sources prior to bringing the data into the common data model and then again after integration. Data integration may bring together indirect identifiers that were originally in separate data sources. Combined together, these potential indirect identifiers may reveal some information that was not possible to reveal from each source taken individually. For instance, consider a dataset containing individuals' education and occupation information. When this dataset is integrated with another dataset from a healthcare facility containing variables on demographics, location and COVID-19 disease status. This combination may reveal the identity of an individual with a rare occupation in their community and thus their health status.

COVID-19 tabular data. If queries are published that report aggregated information in the form of tabular data (aggregates) instead of microdata, methods for tabular data protection/anonymization are used. Again, the frequency of cells in a table can be checked (how many persons contribute to the cell value) and if the number of cases is too small, blocking/suppression may be applied (primary suppression). This is, for example, currently (= this is work-in-progress) done for the corresponding COVID-19 data cubes from the tracing app data of Switzerland with a threshold of 5 persons. All cells with less than 5 contributors are suppressed in the data cubes. However, this information can be easily compromised by differentiating and secondary cell suppression or swapping techniques would be necessary (Gouweleeuw et al. 1998; Fischetti and González 2000). For the secondary cell suppression

approach, these primary locked cells are protected by a secondary suppression to such an extent that the primary suppressed information can no longer be estimated accurately enough. Other alternatives to cell suppression in tabular data are controlled tabular adjustment (Pérez and Giessing 2005), or the cell-key method implemented by Thompson et al. (2013) in https://github.com/sdcTools/cellKey. See, also, an introduction to statistical disclosure control (Duncan et al. 2011; Templ 2017).

COVID-19 model servers: Another concept, differential privacy, aims to maximize the accuracy of responses to queries to databases while minimizing the probability of identifying the records used to respond. The higher the probability of re-identification and the more often the (same) query has been made, the more noise is added to the queried information. Note that there are a lot of implications and restrictions of differential privacy leading to inconsistencies of anonymized datasets and potential low utility of queried data (Bambauer et al. 2014).

In the LMIC, the statistical disclosure control software in use is the sdcMicro software (Templ 2020). The World Bank Group and the International Household Survey Network have supported the development and implementation of this software in the region (LSMS 2020).

DISCUSSION

For COVID-19 data, it must be possible for authorized persons to know the true ID of a patient, i.e. to reverse the encryption of an ID, either in the case where several data sets are merged with this ID or because clarifications are necessary for the patient. To prevent the merging of patient data, the true ID's are salted and encrypted. In a multi-domain and multi-organization framework, a different salt for the same person can be used for each domain and organization. The management of these salts is an IT-technical challenge, but it is well established in practice. As an example, a complex pseudo anonymisation based on the Austrian health data was shown.

Pseudo anonymization only prevents the merging of persons across different data sets. However, it will still be possible to clearly identify persons by a combination of attribute values, and further anonymization is necessary as soon as the data is used by persons who want to analyse the data, but are not authorized to know the true ID of a patient. This is of great importance, if the data is made available to an extended circle of users.

The choice of indirect identifiers is very important. This means asking "*What is the knowledge of an attacker*"? In other words, what possible databases with intersectional populations may be available to someone who receives the data?

For indirect identifying variables, prior to the application of SDC techniques, the re-identification risks of the data should be modeled and estimated. For categorical key variables apply global recoding and local suppression to establish k anonymity - that at least k observations/persons in the data set shares the same values for the indirect identifiers (e.g. age 65, female, lives in region XY, ...), or alternatively, swapping techniques could be used to change, for example, the

place of residence with another person. Each time an SDC technique is used, the risk of re-identification and the increase of information loss should be measured.The anonymization process is typically iterative. Different anonymization methods are tried until the risk of re-identification has been decisively reduced and at the same time the loss of information is considerably low.

The acceptable level of risk depends on many factors. If COVID-19 data are shared as public-use files, they should have much less disclosure risk than scientific-use files for selected researchers or research institutions. The latter are restricted to certain users under certain conditions. COVID-19 data enriched with information on patients other comorbidities containing sensitive information such as HIV/AIDS information may require greater intervention for anonymisation, compared to data with less sensitive information. Risk is never zero, but this is not required. De-facto anonymity is reached whenever the burden and efforts of re-identification is higher than the value for the data intruder.

Data protection must also be guaranteed when delivering data internally within the data holders organization. But it is clear that, in general, the risk for internal data delivery is smaller than externally, since (and only if) the planned processing tasks are described internally, the purposes of the processing should be better known, the risks to data storage can be minimized (the data may be less likely) can be stolen), and it can also be logged who works when with the data, i.e. to be able to control this better than with external data delivery.

FUTURE RESEARCH

- Quantification of the disclosure risk of COVID-19 data cubes (tabular data)
- Quantification of the disclosure risk of COVID-19 microdata
- Applying different anonymization strategies to COVID-19 microdata including the assessment of the disclosure risk and data utility

Author roles

All authors accept responsibility for the content of the article.
Conceptualization: SS, CK, MT. Methodology: SS, CK, MT. Investigation: SS, CK, MT. Formal analysis: SS, CK, MT. Validation: SS, CK, MT, RDA-COVID19 WG. Writing: SS, CK, MT, CCA. Bibliographic review and analysis: SS, CK, MT. Visualization: SS. Supervision and project administration: CCA

Funding: None.
Conflict of interest: None to declare.

REFERENCES

Austin, C. C., Widyastuti, A.; and the RDA-COVID19-WG (2020). COVID-19 Population level data sources: Review and analysis. In *COVID-19 Data sharing in epidemiology, version 0.054*. Research Data Alliance RDA-COVID19-Epidemiology WG. https://doi.org/10.15497/rda00049

Austria. (2013). *Austrian RIS - E-Government-Bereichsabgrenzungsverordnung—Bundesrecht konsolidiert, Fassung vom 22.04.2020*. Austrian Parliament. https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=20003476

Austria. (2018). *Austrian RIS - E-Government-Gesetz—Bundesrecht konsolidiert, Fassung vom 22.04.2020*. https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=20003230

Bambauer, J. R., Muralidhar, K., & Sarathy, R. (2014). Fool's Gold: An Illustrated Critique of Differential Privacy. *Vanderbilt Journal of Entertainment & Technology Law*, *16*(4), 701–755. https://papers.ssrn.com/abstract=2326746

Barth-Jones, D. (2012). The re-identification of Governor William Weld's medical information: A critical re-examination of health data identification risks and privacy protections, then and now. *SSRN Preprints*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2076397

Bertoni, G., Daemen, J., Peeters, M., & Van Assche, G. (2010). Sponge-Based Pseudo-Random Number Generators. In S. Mangard & F.-X. Standaert (Eds.), *Cryptographic Hardware and Embedded Systems, CHES 2010* (pp. 33–47). Springer. https://doi.org/10.1007/978-3-642-15031-9_3

Daemen, J., & Rijmen, V. (2002). *The Design of Rijndael: AES - The Advanced Encryption Standard*. Springer-Verlag. https://doi.org/10.1007/978-3-662-04722-4

Duncan, G. T., Elliot, M., & Salazar, G. J. J. (2011). *Statistical Confidentiality: Principles and Practice*. Springer-Verlag. https://doi.org/10.1007/978-1-4419-7802-8

ELGA GmbH. (2020). *ELGA: Technischer Aufbau im Überblick*. http://www.elga.gv.at/technischer-hintergrund/technischer-aufbau-im-ueberblick/

European Commission. (2020). *Final Report of the Open Science Policy Platform*. https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-policy-platform

Fischetti, M., & González, J. J. S. (2000). Models and Algorithms for Optimizing Cell Suppression in Tabular Data with Linear Constraints. *Journal of the American Statistical Association*, *95*(451), 916–928. https://doi.org/10.1080/01621459.2000.10474282

Fischetti, M., & Salazar, J. J. (2000). *Complementary Cell Suppression for Statistical Disclosure Control in Tabular Data with Linear Constraints*. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.24.1017

Gouweleeuw, J. M., Kooiman, P., Willenborg, L., & Dewolf, P. (1998). Post randomisation for statistical disclosure control: Theory and implementation. *Journal of Official Statistics*, *14*(4), 463–478. https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwjsxoWK4rjqAhUvTt8KHRf2AEIQFjACegQIBBAB&url=https%3A%2F%2Fpdfs.semanticscholar.org%2Fcd28%2F1be11657b944b74169b8fe35dddb91d558e8.pdf&usg=AOvVaw1pesVXBk0j_WlBMbEG7mkh

Greenfield J, Sears M, Nagrani R, Mazzaferro G, Widyastuti A, Austin C C; and the RDA-COVID19-WG. (2020). Common Data Models and Full Spectrum Epidemiology: Epi-STACK architecture for COVID-19 epidemiology datasets. In *COVID-19 Data sharing in epidemiology, version 0.053*. Research Data Alliance RDA-COVID19-Epidemiology WG. https://doi.org/10.15497/rda00049

Le Guillou, I. (2020). *Covid-19: How unprecedented data sharing has led to faster-than-ever outbreak research*. Horizon: The EU Research & Innovation Magazine. https://horizon-magazine.eu/article/covid-19-how-unprecedented-data-sharing-has-led-faster-ever-outbreak-research.html

LSMS (2020). *Tools for privacy protection*. World Bank Development Data Group - Living Standards Measurement Study. Retrieved 6 July 2020, from http://surveys.worldbank.org/sdcmicro

NASEM (2019). *Read 'Reproducibility and Replicability in Science' at NAP.edu*. National Academies of Sciences, Engineering, and Medicine. https://doi.org/10.17226/25303

NIST (2017). *SHA-3 Project - Hash Functions*. National Institute of Standards and Technology - Computer Security Resource Center. https://content.csrc.e1a.nist.gov/projects/hash-functions/sha-3-project

NIST (2019). *Big Data Interoperability Framework, Volumes 1-9*. National Institute of Standards and Technology, USA. https://bigdatawg.nist.gov/V3_output_docs.php

Pérez, J. C., & Giessing, S. (2005, November 9). Testing variants of minimum distance controlled tabular adjustment. *Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*. United Nations Statistical Commission, and Economic Commission for Europe Conference of European Statisticians, Geneva, Switzerland. https://www.semanticscholar.org/paper/Testing-variants-of-minimum-distance-controlled-P%C3%A9rez-Giessing/ec76784d361216db481493cfeecc1e98be42b19d

Templ, M. (2017). *Statistical disclosure control for microdata: Methods and applications in R* (p. 287). Springer International Publishing; Scopus. https://doi.org/10.1007/978-3-319-50272-4

Templ, M., Meindl, B., & Kowarik, A. (2020). *sdcMicro: Statistical Disclosure Control Methods for Anonymization of Data and Risk Estimation*. https://cran.r-project.org/web/packages/sdcMicro/index.html

Thompson, G., Broadfoot, S., & Elazar, D. (2013, October 28). *Methodology for the Automatic Confidentialisation of Statistical Outputs from Remote Servers at ABS (Australian Bureau of Statistics)*. UNECE Work Session on Statistical Data Confidentiality, Ottawa, Canada. https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2013/Topic_1_ABS.pdf

Upshur, R. E. G., Faith, K., Gibson, J. L., Thompson, A. K., Tracy, C. S., Wilson, K., & Singer, P. A. (2005, November). *Stand on Guard for Thee: Ethical Considerations in Preparedness Planning for Pandemic Influenza*. Eweb:289060; University of Toronto Joint Centre for Bioethics. Pandemic Influenza Working Group. https://repository.library.georgetown.edu/handle/10822/978222

Wellcome Trust (2020, April 30). *Outputs Management Plan—Grant Funding | Wellcome*. https://wellcome.ac.uk/grant-funding/guidance/how-complete-outputs-management-plan