Secure Linking of Data from Population-Based Cancer Registries with Healthcare Data to Evaluate Screening Programs

Sichere Verknüpfung von Daten bevölkerungsbezogener Krebsregister und Einrichtungen des Gesundheitswesens zur Evaluation von Screening Programmen

Authors
Sebastian Bartholomäus[1], Yannik Siegert[1], Hans Werner Hense[2], Oliver Heidinger[3]

Affiliations
1 Landeskrebsregister NRW gGmbH, Software Development, Bochum, Germany
2 Department of Epidemiology and Social Medicine, University of Münster, Münster, Germany
3 Landeskrebsregister NRW gGmbH, Managing Board, Bochum, Germany

Key words
record linkage, anonymisation, data protection, screening evaluation, cancer registry

Schlüsselwörter
Record-Linkage, Anonymisierung, Datenschutz, Screening Evaluation, Krebsregister

Bibliography
DOI https://doi.org/10.1055/a-1031-9526
Online-Publikation: 10.12.2019
Gesundheitswesen 2020; 82 (Suppl. 2): S131–S138

ISSN 0949-7013

Correspondence
Mr. Sebastian Bartholomäus
Landeskrebsregister NRW gGmbH, Softwareentwicklung,
Gesundheitscampus 10
44801 Bochum
Germany
bartholomaeus@krebsregister.nrw.de

ABSTRACT

Background The evaluation of population-based screening programs, like the German Mammography Screening Program (MSP), requires collection and linking data from population-based cancer registries and other sources of the healthcare system on a case- specific level. To link such sensitive data, we developed a method that is compliant with German data protection regulations and does not require written individual consent.

Methods Our method combines a probabilistic record linkage on encrypted identifying data with 'blinded anonymisation'. It ensures that all data either are encrypted or have a defined and measurable degree of anonymity. The data sources use a software to transform plain-text identifying data into a set of irreversibly encrypted person cryptograms, while the evaluation attributes are aggregated in multiple stages and are reversibly encrypted. A pseudonymisation service encrypts the person cryptograms into record assignment numbers and a downstream data-collecting centre uses them to perform the probabilistic record linkage. The blinded anonymisation solves the problem of quasi-identifiers within the evaluation data. It allows selecting a specific set of the encrypted aggregations to produce data export with ensured k-anonymity, without any plain-text information. These data are finally transferred to an evaluation centre where they are decrypted and analysed. Our approach allows creating several such generalisations, with different resulting suppression rates allowing dynamic balance information depth with privacy protection and also highlights how this affects data analysability.

Results German data protection authorities approved our concept for the evaluation of the impact of the German MSP on breast cancer mortality. We implemented a prototype and tested it with 1.5 million simulated records, containing realistically distributed identifying data, calculated different generalisations and the respective suppression rates. Here, we also discuss limitations for large data sets in the cancer registry domain, as well as approaches for further improvements like l-diversity and how to reduce the amount of manual post-processing.

Conclusion Our approach enables secure linking of data from population-based cancer registries and other sources of the healthcare system. Despite some limitations, it enables evaluation of the German MSP program and can be generalised to be applicable to other projects.

ZUSAMMENFASSUNG

Hintergrund Die Evaluation bevölkerungsbezogener Früherkennungsprogramme, wie dem deutschen Mammografie-Screening (MSP), erfordert die fallscharfe Verknüpfung von Daten bevölkerungsbezogener Krebsregister und anderen Stellen des Gesundheitswesens. Wir haben eine Methode entwickelt, die ohne individuelle Einwilligung die Verknüpfung solch sensibler Daten im Einklang mit deutschen Datenschutzbestimmungen erlaubt.

Methoden Unser Verfahren kombiniert ein probabilistisches Record-Linkage auf verschlüsselten Identitätsdaten mit einer ‚verblindeten Anonymisierung', sodass sämtliche Daten entweder verschlüsselt sind oder einem definierten Anonymitätsmaß genügen. Die Datenquellen verschlüsseln die identifizierenden Merkmale irreversibel in eine Menge Personenkryptogramme, während die Auswertungsdaten in verschiedenen Stufen aggregiert und reversibel verschlüsselt werden. Ein Pseudonymisierungsdienst verschlüsselt die Personenkryptogramme erneut zu Zuordnungsnummern, die dann von einer nachgelagerten Datensammelstelle zur Verknüpfung der Datensätze mithilfe des Record-Linkage genutzt werden. Die ‚verblindete Anonymisierung' löst das Problem quasi-identifizierender Merkmale in den Auswertungsdaten. Sie ermöglicht, ohne Einsatz von Klartextdaten, aus den verschlüsselten Aggregationsstufen einen k-anonymen Datensatz zu erstellen. Die geprüft anonymen Auswertungsdaten werden an eine evaluierende Stelle übertragen, dort entschlüsselt und ausgewertet. Unser Ansatz erlaubt die Erzeugung verschiedener Generalisierungen, wodurch dynamisch die Informationstiefe gegen die Anforderungen des Datenschutzes abgewogen und der Einfluss auf die Auswertbarkeit hervorhoben werden kann.

Ergebnisse Unser Konzept wurde von den deutschen Datenschutzbehörden für die Mortalitätsevaluation des deutschen MSP zugelassen. Wir entwickelten einen Prototyp und erprobten ihn mit 1,5 Mio. simulierten Datensätzen und realistisch verteilten Identitätsdaten. Dabei berechneten wird verschiedene Generalisierungen und die resultierenden Unterdrückungsraten. Wir diskutieren die Limitierungen unseres Ansatzes sowie mögliche Verbesserungen wie die l-Diversität und die Reduktion manueller Nachbearbeitungsschritte.

Schlussfolgerung Unser Ansatz erlaubt die sichere Verknüpfung von Daten aus bevölkerungsbezogenen Krebsregistern und anderen Einrichtungen. Obwohl einige Limitierungen greifen, erlaubt das Konzept die Evaluation des deutschen MSP und kann für den Einsatz in anderen Projekten generalisiert werden.

Introduction

To evaluate the performance and outcomes of cancer screening programs, it is necessary to collect and link data from multiple sources of the healthcare system on a case-specific level. Since the General Data Protection Regulation (GDPR) of the European Union [1] came into effect, this has become a challenging task.

If neither individual written informed consents nor a specific law are available, existing regulations in Germany enforce the usage of anonymised or, if that is not feasible, pseudonymised data wherever possible. Falling short of these directives requires explicit justification by outlining the predominant public interest in the research project and providing evidence that the goals cannot be accomplished by using anonymised or pseudonymised data.

In 2012 the Federal Office for Radiation Protection in Germany (Bundesamt für Strahlenschutz, BfS) commissioned an evaluation study on the impact of the German Mammography-Screening Program (MSP) on breast cancer mortality (Grant numbers: 3610S40002, 3614S40002, 3617S42402).

A major problem was the fact that no single institution holds all necessary data for this kind of evaluation. Instead, these sensitive data need to be collected from cancer registries, institutions of the screening program, health insurance companies and the Associations of Statutory Health Insurance Physicians (Kassenärztliche Vereinigungen).

As part of the research consortium, the Cancer Registry of North Rhine-Westphalia (Landeskrebsregister NRW gGmbH, LKR-NRW) had the task to develop a data flow and processing model. The task was to collect, link and anonymise data from these different sources in a way that does not require individual consent or a specific law and is nevertheless compliant to existing legal regulations in Germany and the European General Data Protection Regulation (GDPR). We already published the basic concept in a preliminary, much shorter paper [2].

Methods

To achieve the goals of the study we had to solve two major problems: How can we link data from different sources reliably and securely, and how do we deal with potentially identifying combinations of attributes in the evaluation data.

As there is no common global identifier like a social security number in Germany, we decided to use the probabilistic record-linkage relying on encrypted identifying data, which the LKR-NRW applied successfully from 2005 to 2016 before its transformation from a purely epidemiological to an integrated clinical and epidemiological cancer registry. The method in the context of cancer registration [3] and its appliance in the LKR-NRW [4] have been published and evaluated [5] before, so we just recap information where necessary to understand the integrated concept of this paper.

As stated before, even data that are essential for evaluation purposes may contain potentially identifying combinations of individual attributes ('quasi-identifiers') [6]. To protect the privacy of individuals contained in such data sets the data needs to be anonymised. The anonymisation of personal data is an active field of research and there are numerous anonymity measures and techniques, some even capable to create an anonymised dataset in a distributed environment [7, 8]. However, to our knowledge, none of these is suitable for the scenario of the MSP in which data are distributed horizontally as well as vertically (i. e. sources contribute different cohorts and also different attributes), the horizontal sub-

sets overlap and there is no global identifier. To solve these difficulties, we exploit a fundamental idea of the record-linkage based on encrypted identifiers, apply it to the anonymisation problem and combine both into an integrated concept.

In the actual record-linkage algorithm by Fellegi and Sunter [9], all linkage decisions come down to the basic operation of checking whether two given values are equal or not. However, this simple test does not require any plain text information as it can validly be checked on deterministically encrypted data too, whereby the same input is always mapped to the same cryptogram. Fortunately, there are anonymity measures like the well known k-anonymisation [10] that also use equality checks as their fundamental operation, which makes them also viable to be checked on deterministically encrypted data.

By combining the pseudonymised linkage with a 'blinded anonymisation' on deterministically encrypted data, we ensure that all critical data either are encrypted or have a predefined degree of anonymity after leaving their original data sources.

▶ **Figure 1** shows a high-level view of the participating parties in the data processing. Various data sources (DS) hold the necessary data and use a local reporting tool, which performs substantial pre-processing and encryption steps for the pseudonymised record-linkage as well as the blinded anonymisation. The pre-processed data are sent to a pseudonymisation service (PSS) which adds another layer of security and forwards the data to the data-collecting centre (DCC). The data-collecting centre executes the pseudonymised record-linkage as well as the blinded anonymisation algorithms without any plain text information and exports the anonymised but still encrypted data to an evaluation centre (EC). The evaluation centre finally decrypts the anonymised data, performs research on it and provides excerpts of the data for secondary research groups (RG).

Although our concept makes heavy use of encryption technology, we generalize from the actual algorithms, as they are not relevant for understanding the concept and have to be chosen based on current technological standards. In addition to all mentioned cryptographic measures, we protect all point-to-point communication by TLS (Transport Layer Security) encryption. In the Figures and following paragraphs, we use a special notation to highlight different layers of encryption. E.g., once a set of data (DATA) is encrypted in a way that only allows the evaluation center (EC) to decrypt it, we refer to it as $(DATA)_{EC}$. After applying an additional layer of encryption, which can only be decrypted by the data collecting center (DCC), it is noted as $((DATA)_{EC})_{DCC}$.

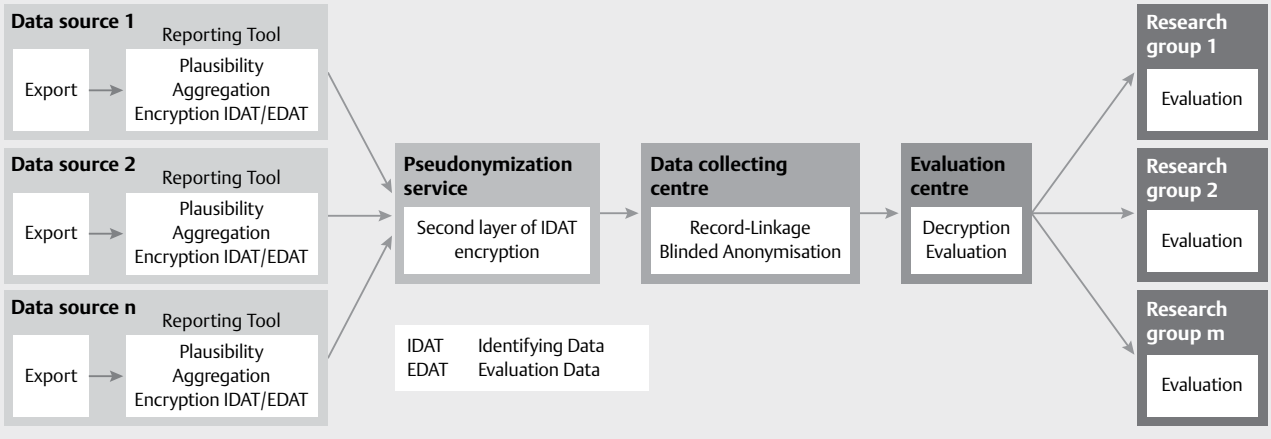
The data processing at each participating party mostly consists of two parts; one for the directly identifying data (IDAT) and another for the evaluation data (EDAT).

The IDAT consist of the forenames, surnames, date of birth, full address and gender of each individual. Each data source needs to provide this information. Additional attributes like titles, birth names and former names can improve the linkage results. The IDAT are processed and encrypted in an irreversible way, and are only used for linking the data from the various sources. They cannot be used for evaluation purposes.

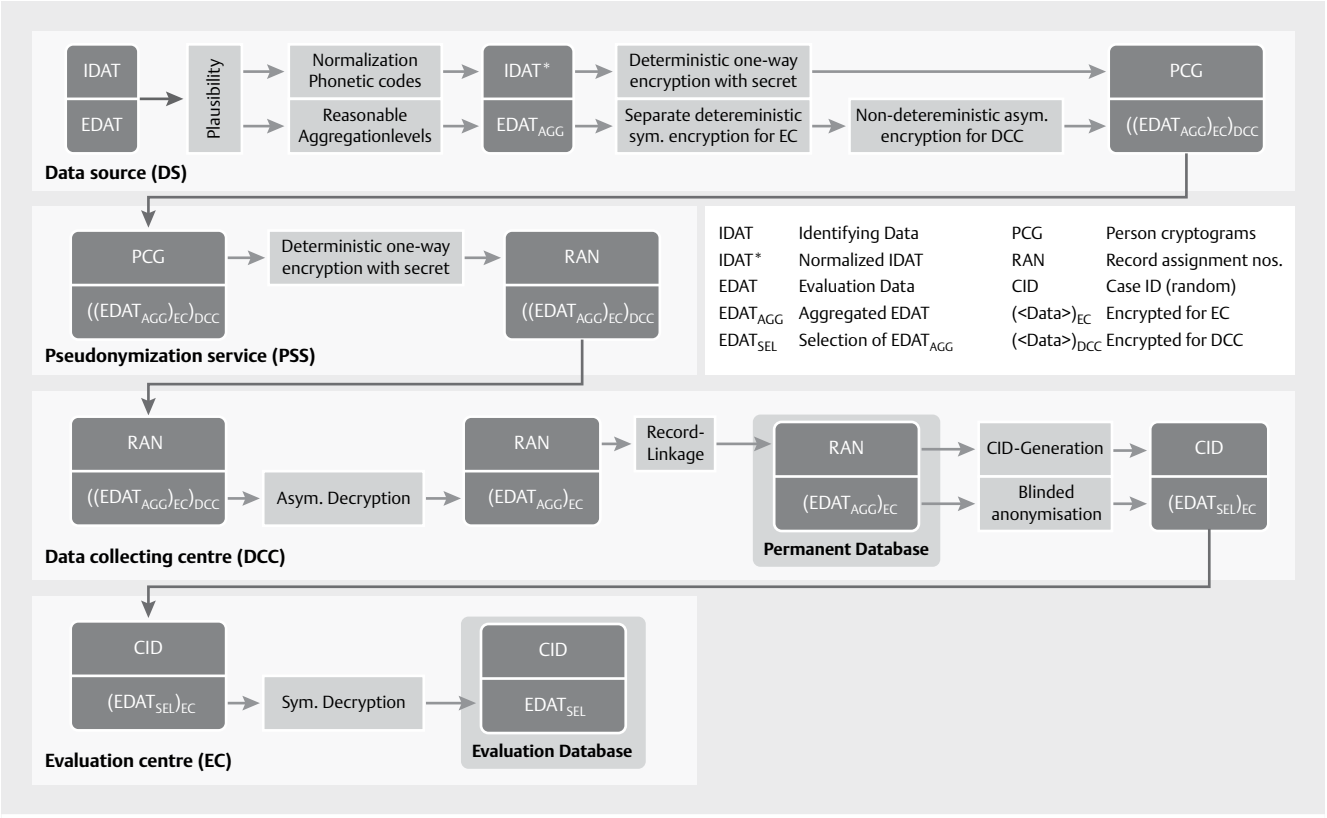
The EDAT instead are processed in way that allows to decrypt specific parts of them in the evaluation centre at the end of the process. Each data source might contribute different EDAT attributes. Beside medical or organisational information, the evaluation data will most likely also contain a selection of plain text identity attributes that are necessary for evaluation purposes and are the strongest candidates for quasi-identifiers. In the context of the German MSP these are e. g. the date of birth and the zip code.

We will now traverse through each step of the process up to the evaluation centre in detail, following the graph in ▶ **Fig. 2**.

Before transmitting any of their collected data, the data sources use a dedicated reporting tool we named SecuNym-RT, which pre-processes the IDAT and EDAT. Quality of the input data is paramount, as the reporting tool will finally encrypt all data, effectively preventing data corrections later on. Therefore, processing starts with domain specific plausibility checks. We created SecuNym-RT in a modular fashion in order to be easily adoptable for different kinds of records and projects.



▶ **Figure 1** High-level view of participating parties and important processes. Source: Fuhs A, Bartholomäus S, Heidinger O et al. Evaluation der Auswirkungen des Mammographie-Screening-Programms auf die Brustkrebsmortalität. Bundesgesundheitsblatt-Gesundheitsforschung-Gesundheitsschutz 2014; 57, 1: 60–67.



▶ **Figure 2** Detailed graph of all relevant processes for identifying and evaluation data. Source: Hense H-W, Barlag H, Bartholomäus S et al. Evaluation der Brustkrebsmortalität im Deutschen Mammographie-Screening-Programm – Vorhaben 3610S40002 und 3614S40002 2017; Edition: Ressortforschungsberichte zur kerntechnischen Sicherheit und zum Strahlenschutz, Publisher: Bundesamt für Strahlenschutz (BfS), ISBN: http://nbn-resolving.de/urn:nbn:de:0221-2017050314273.

Following the plausibility checks, the reporting tool normalizes all IDAT. This includes splitting street names into up to five parts, separating multiple fore-, sur-, and birth names as well as generating phonetic codes, to take phonetic similarities in names into account that often lead to misspellings. For each record the values of the resulting set of up to 31 normalized attributes (IDAT *) are individually encrypted, transforming them into a set of person cryptograms (PCG). Unlike the pseudonymised linkage used by the LKR-NRW, which still kept some plaintext information for the record-linkage (month and year of birth, zip code, residence and gender), in our concept all normalized IDAT are used only in encrypted form. The encryption is a deterministic one-way function with a secret key (e. g. an Hash-based Message Authentication Code HMAC) which is shared between all data sources. The secret key prevents the pseudonymisation service from trying to resolve the person cryptograms by mass encrypting plaintext names and thus creating a reference table.

For each attribute of the evaluation data, SecuNym-RT creates multiple reasonable levels of aggregation ($EDAT_{AGG}$). Following the principle of data minimization, some aggregation levels are omitted:

- low aggregation levels with overly specific information (e. g. a full date of birth or diagnosis), and
- high aggregation levels that are insufficiently precise for the evaluation.

▶ **Figure 3** depicts an example of reasonable aggregation levels for the attributes date of birth, zip code and date of diagnosis.For all attributes, the value of each remaining aggregation level is then encrypted separately for the evaluation center ($(EDAT_{AGG})_{EC}$). We use a deterministic symmetric algorithm so that the encrypted values can be used by the blinded anonymisation algorithm and finally be decrypted by the evaluation centre at the end of the process. The secret key needs to be shared between all data sources and the evaluation centre; a limitation that we will discuss later.

To prevent the pseudonymisation service from learning anything out of the aggregated and separately encrypted evaluation data, we add an end-to-end encryption using an asymmetric non-deterministic algorithm. By applying the public key of the data-collecting centre we create the $((EDAT_{AGG})_{EC})_{DCC}$, which are a single block of encrypted data for each record.

The data sources transmit the person-cryptograms alongside the encrypted evaluation data to the pseudonymisation service. The pseudonymisation service in turn transforms the set of person-cryptograms of each record into a set of record assignment numbers (RAN) by using a deterministic one-way function with a secret key. The record assignment numbers are more secure then the person-cryptograms because the secret to create them only resides at the pseudonymisation service. In addition, this prevents the data-collecting centre to communicate with the data sources about individual records (six eyes principle). The PSS then transmits the

Aggregated evaluation data				
$EDAT_{AGG}$	date of birth	zip	date of diagnosis	Lower bound: data avoidance
level 0	~~06.10.1944~~	~~66879~~	~~15.09.2010~~	
level 1	10. 1944	6687#	09. 2010	Reasonable aggregations
level 2	Q4 1944	668##	Q3 2010	
level 3	H2 1944	66###	~~H2 2010~~	Upper bound: analysability
level 4	~~1944~~	~~6####~~	~~2010~~	

Aggregated evaluation data, encrypted for the evaluation centre				
$(EDAT_{AGG})_{EC}$	date of birth	zip	date of diagnosis	deterministic encryption
level 1	gtz54D230oi34	f3409gkn439	i2nf23ng43u89	
level 2	54g4w5h5676u	43t89u43gn4	h44u7jsfglkeds	
level 3	32j3fk65ds	r302q94ufk5i		

▶ **Figure 3** The reporting tool generates reasonable aggregation levels and encrypts each value for the EC. Source: Bartholomäus S, Hense H W, Heidinger O. Blinded anonymization: A method for evaluating cancer prevention programs under restrictive data protection regulations. In: Studies in health technology and informatics. 2015: 210; 424–428.

record-assignment-numbers (RAN) and the encrypted evaluation data to the DCC.

The DCC reverts the non-deterministic encryption and uses a subset of the record assignment numbers (all name-parts and phonetic codes, day, month and year of birth, gender, zip code and residence) to perform the probabilistic record-linkage. The process differs slightly from the linkage performed by the LKR-NRW. The main difference is that there is no plain-text information at all, which makes the manual post-processing of uncertain matches more complicated. In the LKR-NRW revisers used a rich set of information to make the manual decisions. These included:

- all available record assignment numbers (including those not used in the automatic linkage like street-name parts, house-number parts, title parts, etc.),
- the available plain-text information (year of birth, gender and residence) and
- additional medical information from the records.

This allowed to make decisions based on common sense, e. g. that the equality of all name parts, age and gender in a big city less likely indicates a match than in a small town or that the ICD diagnostic codes in the records belong to the same kind of cancer. In our setup there is no such plain-text information, but many aspects important for the manual post-processing can be emulated: E.g.

- by providing the revisers the relative frequency of a given cryptogram for a zip code or
- comparing the different aggregation values for an ICD code and see if they share a common value on a higher level of aggregation.

The linked records are stored at the data-collecting centre. Periodically, the data-collecting centre starts the blinded anonymisation process to create an anonymised data export for the evaluation centre. In the course of this process, the data-collecting centre selects a generalisation (i. e. a set of aggregation levels for each potential quasi-identifier attribute in the EDAT) such that there are always at least k records that are identical with regard to their quasi-identifiers. The selected aggregation levels of the quasi-identifiers ($(EDAT_{SEL})_{EC}$) fulfil the requirement of k-anonymity and are ultimately transmitted to the evaluation centre (▶ **Fig. 4**). Using this approach the data-collecting centre can ensure that the evaluation data meet a predefined degree of anonymity expressed by the k-value and the set of quasi-identifiers, without actually knowing the sensitive content of the records.

In order to achieve a given k-value the data-collecting centre might have to suppress a certain amount of records because they contain very rare quasi-identifiers and would result in generalisations with unavailable or undesirably high aggregation levels. The chosen aggregation levels and the necessary suppression rate are an important indicator for the analysability of the resulting dataset and should be minimised.

Before transferring the composed data to the evaluation centre, the data-collecting centre replaces the record assignment numbers of the records with a single random case id (CID). This allows to solve technical problems in the process, but does not allow the evaluation centre to link different exports in order to gain additional knowledge.

The evaluation centre finally decrypts the selected $(EDAT_{SEL})_{EC}$ and gains the final anonymised dataset with plain-text epidemiological and medical evaluation data.

Results

Our concept has been approved by German data protection authorities for use in the German MSP evaluation study. Since then we have implemented a “proof-of-concept” prototype containing

all key processes and tested it against a simulated data set with more than 1.5 million records. The simulated records contained realistically distributed IDAT for women in the age of 50–69 years living in North Rhine-Westphalia (NRW) and some potentially 'quasi-identifying' EDAT attributes, like the date of diagnosis. We generated different generalisations and analysed the suppression rates resulting from different k-values. ▶ **Fig. 5** depicts the suppression rates for four different generalisations. Setting k = 5, an aggregation level of a four-digit postal code and a date of birth as MMYYYY (▶ **Fig. 5** data series A) proved barely practical for an evaluation dataset, because the suppression rate was above 2.5 %. A one step higher aggregation level for one of the attributes resulted in a significantly lower suppression rate: e. g., using the date of birth by quarters of a year QYYYY (▶ **Fig. 5** data series C), the suppression rate was below 0.1 % and values up to k = 11 would still allow a suppression rate of less than 1.0 %.

In the actual study, we have to add more attributes into the set of quasi-identifiers, which will naturally lead to higher suppression rates.

However, the intriguing advantage of our approach is that all aggregation levels will be available in the data-collecting centre, which allows it to create all kinds of generalisations which are required (or desired) for the respective research questions and calculate the suppression rates. This way one can also suggest sets of quasi-identifi-

Aggregated evaluation data, encrypted for the evaluation centre

$(EDAT_{AGG})_{EC}$	date of birth	zip	date of diagnosis
level 1	gtz54D230oi34	f3409gkn439	i2nf23ng43u89
level 2	54g4w5h5676u	43t89u43gn4	h44u7jsfglkeds
level 3	32j3fk65ds	r302q94ufk5i	

Select minimal levels of aggregation that satisfy predefined k value

Selection of aggregated evaluation data, encrypted for the evaluation centre

$(EDAT_{SEL})_{EC}$	date of birth	zip	date of diagnosis
level 1		f3409gkn439	
level 2	54g4w5h5676u		h44u7jsfglkeds
level 3			

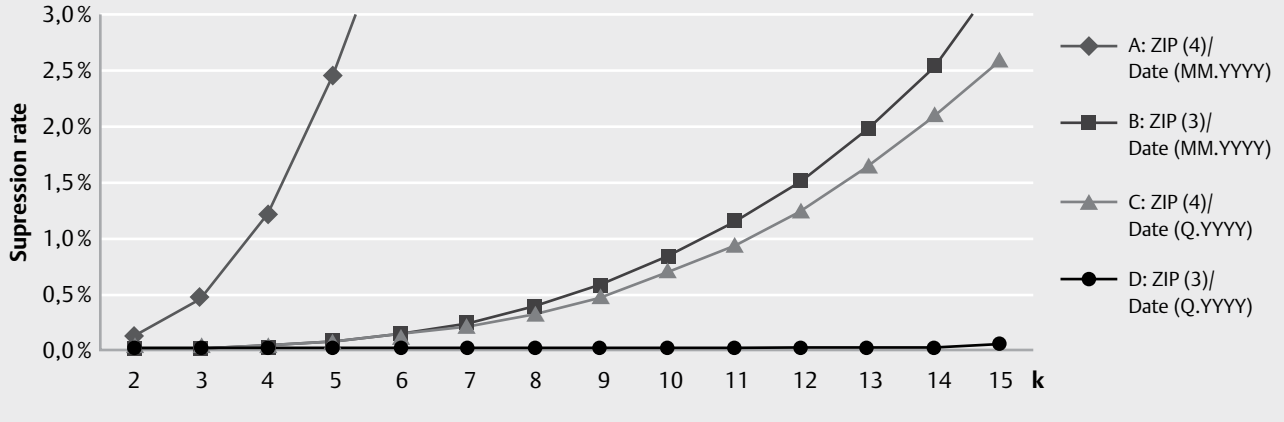
Selected data that in sent to the EC

Selection of aggregated evaluation data, plain text

$EDAT_{SEL}$	date of birth	zip	date of diagnosis
plaintext	Q4. 1944	6687#	Q3. 2010

Anonymised data

▶ **Figure 4** Choosing the aggregation levels that satisfy predefined k values. Source: Bartholomäus S, Hense H W, Heidinger O. Blinded anonymization: a method for evaluating cancer prevention programs under restrictive data protection regulations. In: Studies in health technology and informatics. 2015: 210; 424–428.



▶ **Figure 5** Suppression rates for combinations of 3/4 digit zip codes and monthly/quarterly date of birth. Source: Bartholomäus S, Hense H W, Heidinger O. Blinded anonymization: a method for evaluating cancer prevention programs under restrictive data protection regulations. In: Studies in health technology and informatics. 2015: 210; 424–428.

ers and required k-values to data protection authorities such that the final generalisations yield suppression rates compatible with the evaluation purpose. Ultimately, our approach is able to balance information depth with data protection very specifically and based on empirical data, which is completely encrypted.

Additionally, the data-collecting centre can produce generalisations that are specifically tailored to a particular research question. For example, the data-collecting centre could sacrifice precision in the spatial resolution, i. e. higher aggregation level for the zip code, to gain a higher precision for the birth date or vice versa (▶ **Fig. 5**, data series B and C).

We are currently developing the actual software suite for the project and are testing it in the model region of NRW. SecuNym-RT has been deployed at one of the data sources of the MSP project and we received over 80.000 simulated records that were derived from real data.

Discussion

An important precondition for the usability of our approach is the quality and structured nature of the primary data. As the main data pre-processing occurs on the side of the data sources, all imported data have to be standardized and of high validity, thus the data sources need to be selected carefully. In the setting of the MSP the primary data source are cancer registries , which contribute to the project. The other data sources mainly contribute the identity data of chosen cohorts for a case specific linkage and limited additional information about the outcome and participation in screening or grey screening, which are also highly structured.

The main technical limitation of our approach is the number of data sources that share the secret keys for the deterministic symmetric encryption of the aggregated evaluation data. Although approaches for secure deterministic asymmetric encryption are also discussed [11] they rely on a high entropy on the plain-text information which does not hold true for the limited domain of zip codes, birth dates or encoded cancer classifications. An adversary could simply use the public key to encrypt all possible inputs, effectively decrypting the values. So to prevent the keys from being compromised we need to limit the overall number of data sources. In the actual MSP evaluation cancer registries also provide complete mortality information even for persons not affected by cancer. This reduces the number of data sources by a great amount, as otherwise hundreds of registration offices would have to become data sources too.

Although k-anonymity is a comparatively weak measure of anonymity [12], we nevertheless employ it due to its property to be usable with deterministically encrypted data, too. This also applies for the slightly stronger l-diversity [13], which additionally requires a specific amount of diversity of the sensible information in each k-group. We plan to add l-diversity to our approach in the future.

So far, aggregation levels have to be configured manually and acceptable suppression rates are determined via trial and error. We currently examine algorithms, e. g. Incognito [14], that allow a highly automated search for optimal generalisations.

Another limitation is the need for a manual post-processing of the record-linkage. For around 5 % of all records, the probabilistic record-linkage algorithm cannot clearly classify whether a new report belongs to an individual already contained in the database or not. These uncertain matches need to be manually resolved. To approach this issue, we presently examine a set of machine learning techniques [15]. We trained the classifiers with routine decision data of the LKR-NRW and the results appear to indicate a reduction of manual work by 80 % without sacrificing quality.

Conclusion

Our approach to combine a probabilistic record-linkage based on encrypted identifiers with an anonymisation performed on pre-aggregated and encrypted evaluation data seems very promising. German data protection authorities have approved the application of this concept in the evaluation of the German mammography-screening program.

Although there are some limitations on where our concept can be applied, the prototype implementation highlights the advantages of our approach: avoiding any plain-text data, the method enables the creation of all kinds of required (or desired) generalisations. In fact, the set of quasi-identifiers and required k-values may even be negotiated with data protection authorities based on the empirical data. Likewise, precision can be reduced in one attribute to gain precision in another for particular research questions. Thus, our toolset permits balancing information depth with data protection.

The actual implementation of our software SecuNym aims for a higher degree of automation by applying machine-learning approach, in particular to the post-processing of the record-linkage.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

[1] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). 2016; OJ L119/1

[2] Bartholomäus S, Hense HW, Heidinger O. Blinded anonymization: A method for evaluating cancer prevention programs under restrictive data protection regulations. In Studies in health technology and informatics; 2015: 210;424–428

[3] Meyer M. Kontrollnummern und Record-Linkage. In: Hentschel S, Katalinic A, Das Manual der epidemiologischen Krebsregistrierung. Zuckschwerdt; 2008: 57–68

[4] Krieg V, Hense H-W, Lehnert M et al. Record Linkage mit kryptografierten Identitätsdaten in einem bevölkerungsbezogenen Krebsregister. Das Gesundheitswesen 2001; 63: 6: 376–382

[5] Schmidtmann I, Sariyar M, Borg A et al. Quality of record linkage in a highly automated cancer registry that relies on encrypted identity data. GMS Medizinische Informatik, Biometrie und Epidemiologie 2016; 12: 1

[6] Dalenius T. Finding a needle in a haystack or identifying anonymous census records. Journal of Official Statistics 1986; 2, 3: 329

[7] Jiang W, Clifton C. Privacy-preserving distributed k-anonymity. In Data and Applications Security XIX. 3654. Springer Berlin / Heidelberg; 2005: 924–924

[8] Kohlmayer F, Prasser F, Eckert C et al. A flexible approach to distributed data anonymization. Journal of Biomedical Informatics 2014; 50: 62–76

[9] Fellegi IP, Sunter AB. A theory for record linkage. Journal of the American Statistical Association 1969; 64: 328: 1183–1210

[10] Sweeney L. k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 2002; 10: 557–570

[11] Brakerski Z, Segev G. Better security for deterministic public-key encryption: The auxiliary-input setting. Journal of cryptology 2014; 27: 210–247

[12] Domingo-Ferrer J, Torra V. A critique of k-anonymity and some of its enhancements. In Third International Conference on Availability, Reliability and Security; Barcelona, Spain: 2008: 990–993

[13] Machanavajjhala A, Kifer D, Gehrke J et al. l-diversity: Privacy beyond k-anonymity. ACM Trans. Knowl. Discov. Data 2007; 1: 1

[14] LeFevre K, DeWitt DJ, Ramakrishnan R. Incognito: Efficient full-domain k-anonymity. InProceedings of the 2005 ACM SIGMOD international conference on management of data; New York, USA: 2005: 49–60

[15] Siegert Y, Jiang X, Krieg V et al. Classification-based record linkage with pseudonymized data for epidemiological cancer registries. IEEE Transactions on Multimedia 2016; 18: 1929–1941