Methods of Information in Medicine
© F. K. Schattauer Verlagsgesellschaft mbH (1996)

K. Pommerening, M. Miller, I. Schmidtmann, J. Michaelis

Institut für Medizinische Statistik
und Dokumentation
der Johannes-Gutenberg-Universität,
Mainz, Germany

Pseudonyms for Cancer Registries

Abstract: In order to conform to the rigid German legislation on data privacy and security we developed a new concept of data flow and data storage for population-based cancer registries. A special trusted office generates a pseudonym for each case by a cryptographic procedure. This office also handles the notification of cases and communicates with the reporting physicians. It passes pseudonymous records to the registration office for permanent storage. The registration office links the records according to the pseudonyms. Starting from a requirements analysis we show how to construct the pseudonyms; we then show that they meet the requirements. We discuss how the pseudonyms have to be protected by cryptographic and organizational means. A pilot study showed that the proposed procedure gives acceptable synonym and homonym error rates. The methods described are not restricted to cancer registration and may serve as a model for comparable applications in medical informatics.

Keywords: Cancer Registry, Data Protection, Data Encryption, Pseudonyms, Record Linkage.

1. Introduction

Until recently, the rigid German legislation on data privacy and data security has hindered comprehensive cancer registration in major parts of Germany. The new European directive on data protection [1] may pose further difficulties. The basic premise states that permanent storage of an individual's medical data together with his/her identification data is allowed on the basis of informed consent only. However, many cancer patients nowadays are still not completely informed about the nature of their disease and, therefore, cannot be asked for informed consent to report their data to a cancer registry. Hence, it is desirable that physicians should have the right to notify incident cases without obtaining informed consent in order to assure the necessary completeness of cancer registration. Notification without informed consent is regarded as violation of an individual's constitutional right to data privacy, unless it is compensated by anonymity.

A cancer registry, however, needs identification data for record linkage, to identify multiple notifications of the same individual, and to record follow-up information on individuals. On the other hand, scientific analysis of the registry data is generally performed anonymously and does not include any reference to individual identification data.

To minimize the violation of data privacy we developed a new organizational and technical concept for cancer registries which has been approved by data-protection officials and incorporated into the corresponding German federal legislation [2]. In our concept the registry is separated into two offices with complementary functions. The concept makes extensive use of data encryption and provides data privacy by pseudonymous data storage. This mode of data storage allows record linkage by matching of pseudonyms and does not interfere with the scientific requirements of a cancer registry. In certain cases a controlled re-identification of records might be necessary to obtain follow-up information about cases. The concept includes provisions for achieving this.

A pilot study was initiated in 1992 to explore the possibilities for running a population-based cancer registry in Rheinland-Pfalz (Rhineland-Palatinate) on the basis of this concept [3-5]. The results show that the proposed compromise between research interests and privacy issues is practicable and sound. Further overviews have been given in [6-8]. The concept has also been adopted for the pilot phase of the cancer registry of Niedersachsen (Lower Saxonia) [9].

The cryptographic concept of pseudonymity can be adapted to other situations where a fundamental conflict between the goals of privacy and public interest needs to be solved, e.g., to control the effiency of health care [10, 11].

2. Pseudonyms

Pseudonyms are distinct, unlinkable identities that an individual assumes in order to hide his or her true identity. In information technology pseudonyms control the matching of data while preserving privacy. A pseudonym belongs to one person only (henceforth called 'the owner') but does not reveal the identity of that person. If only the owner can uncover the pseudonym, it is called 'untraceable'. This concept was introduced into cryptology by Chaum [12]; it is useful to protect privacy in electronic banking, electronic elections, and other electronic transactions. Possible (but not yet realized) applications in the medical domain are anonymous electronic prescriptions [10] or the settlement of accounts between physicians and insurance companies [11].

Cancer registries need a distinct kind of pseudonyms which must satisfy the following requirements:

1. The registry must be able to recognize multiple notifications of the same case (record linkage).
2. The record linkage procedure should minimize synonym and homonym errors (see section 6) to yield sufficient data quality.
3. Collaborating registries should be able to match their records.
4. In certain controlled circumstances the uncovering of a pseudonym should be possible for obtaining additional information, e.g. within the scope of case-control studies.
5. The owner should not be able to uncover his own pseudonym.

This last point derives from the right to notify a case without informing the patient about his disease. It implies that the owner should not generate his pseudonym; instead, we need a trusted institution that generates the pseudonyms.

To satisfy the first requirement the pseudonym should be generated by an algorithmic procedure that can be reproduced. The prefered method is hashing [13, par. 6.4]. Since the hash values should not reveal any information about the original data, we use a cryptographic hash function [14, chap. 14]. Since no one except the trusted institution should be able to generate a pseudonym for cancer registry, the procedure should depend on a secret key which is kept by the trusted institution. Such a pseudonym can by no means be uncovered; the key-dependent procedure even prevents unauthorized trial encryption, at least from outside.

This kind of pseudonym does not meet requirement 2, the reason is lack of fault tolerance: the encryption process cannot compensate for slight variations in the identification data, e.g., mistakes in spelling the name. This is not a problem when machine-readable identification data on patient cards can be used; but this is not always the case. Certain notifying institutions, such as pathologists, may not have access to the patient card. Old data (from the time before the introduction of patient cards) should also be linked. In any case, requirement 2 conflicts with complete anonymity; the model has to provide a balance between these two conflicting goals. What we need is a concept of error detection and error correction for encrypted data. Finding an optimal solution is an interesting problem for further research. As a first solution we divide the 'one-way' part of the pseudonym into a set of 'linkage data' that satisfy requirements 1, 2 and 5.

In order to meet requirement 4 we add a second part to the pseudonym. This part derives from the identification data of the patient by encryption; the key is known only to the trusted institution. For reasons to be discussed later we use asymmetric encryption with two keys (see section 5.1).

The reason for requirement 3 is that the German Federal States will have separate registries. To enable anonymous data matching between these registries they could use a common cryptographic key, but this is not advisable: A secret loses its value if shared among too many parties. Therefore, for inter-registry linking we propose a re-encryption of the first part of the pseudonym with a temporary (one-time) key (for details, see section 5.3).

Our concept of pseudonymity in cancer registry needs an organizational framework that is described in the next section.

3. Organizational structure of registry

The cancer registry consists of two separate offices at separate locations. The first office (trusted office, "Vertrauensstelle") basically serves for the notification and generates the pseudonyms. The second office (registration office, "Registerstelle") links the records and stores data permanently.

3.1. Identity Data and Epidemiological Data

In the following we distinguish between identity data and epidemiological data. Identity data are:

- surname, former surname(s), given name(s),
- address,
- date of birth, date of death,
- date of diagnosis,
- notifying physician or health-care institution.

Epidemiological data are those data that are needed in every meaningful statistical evaluation of the registry data:

- gender,
- census code of place of residence,
- professional group,
- year of birth, year of death,
- year of diagnosis,
- date of notification,
- tumor classification,
- further medical data.

3.2. The Trusted Office

The trusted office accepts incoming reports from physicians or hospital-based cancer registries. These reports are checked for completeness and plausibility. If necessary, this office obtains additional information from the reporting physicians. It codes the reported diseases according to classification schemes such as ICD-9 and ICD-10. Thereafter, it assigns a pseudonym to the record, and sends the pseudonymous record to the registration office. After a short period of time, when any discrepancies are cleared, the trusted office deletes the records in its database. Death certificates are also sent to the trusted office and handled in the same way as notification forms.

The trusted office is directed by a physician and, therefore, is subject to professional discretion in addition to data-protection laws. It is trusted by all other parties, hence the German name "Vertrauensstelle". Nevertheless, the decryption key – the 'private' key of the asymmetric encryption procedure, henceforth called 're-identification key' – is held in a second trusted institution outside the cancer registry. There are several sensible choices for this institution; in the following we call it the 'supervising office'. The separate handling of the re-identification key emphasizes the 'separation of informational powers' and makes clear that decryption (= re-identification) is an exceptional process. Moreover, it gives additional security in case of a compromised encryption key.

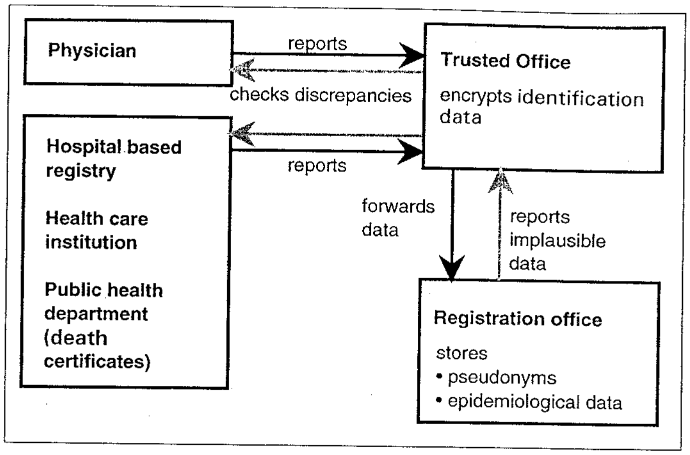

Fig. 1 Organizational structure and information flow.

3.3. The Registration Office

The registration office receives pseudonymous data only. With these data it performs record linkage and detects duplicate notifications; then it stores the pseudonyms and the epidemiological data permanently. If the record linkage reveals any inconsistencies, these are reported back to the trusted office which, in turn, may sort out any discrepancies by contacting the reporting physicians. In the same way the office links a death certificate to an existing patient record. Figure 1 illustrates the data flow. Only the registration office stores records permanently.

3.4. Epidemiological Studies

The pseudonymous records serve for routine analyses of the cancer registry as well as for epidemiological studies. Figure 2 illustrates the procedure for a cohort study: if a well-defined cohort (e.g., occupationally exposed employees of a company) is to be analyzed for the occurrence of cancer, a sequence number is assigned to each individual member of the cohort and possibly also to non-exposed controls. These sequence numbers serve as simple temporary pseudonyms for the study. A research institute (which could also be the registry) obtains a record for each individual containing the sequence number and the exposure data. A record containing the sequence number and personal identification data is sent to the trusted office in parallel. This office generates the pseudonym and sends it to the registration office, together with the sequence number. The registration office performs the record linkage and generates a record which contains the sequence number and the epidemiological data stored in the registry. Thereafter, epidemiological data and exposition data may be linked for further analysis by using the sequence number. This procedure ensures that for the purpose of the study nobody sees which cohort members were diseased.

A corresponding procedure applies to case-control studies if only the epidemiological data which are kept in the registry are needed for such a study.

If it is necessary to obtain additional information from the diseased patients, the identification data may be decrypted using the re-identification key which is kept in the supervising office (see section 3.2). Re-identification has to be approved by an ethics committee and is done in the supervising office; technically this could also be realized with a portable PC operated by an employee of the supervising office. The decrypted identification data are then given to the trusted office. In some cases the necessary data can be retrieved from the notifying institution. If it is necessary to contact the patient for an additional inquiry, the trusted office has to obtain informed consent from the patient via the notifying or treating physician whose identity is stored as part of the (encrypted) identification data of the patient (see section 3.1).

4. A Registry Model

Since a strict formalization of the procedures of the previous section in

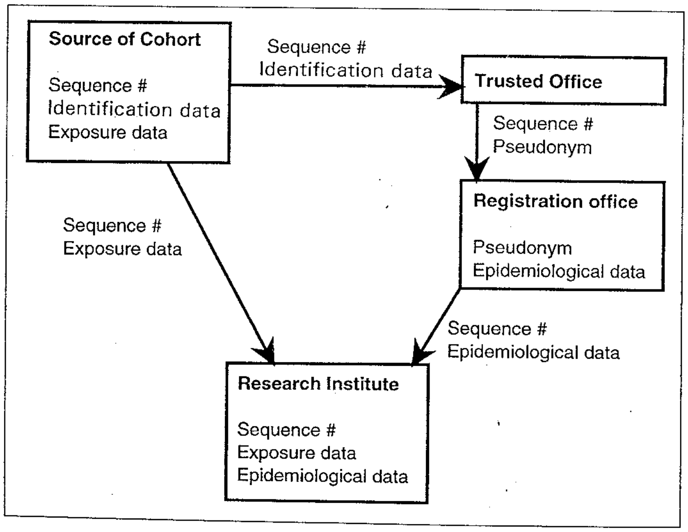

Fig. 2 Record linkage for cohort studies.

the sense of [15] would be too technical for this paper, we only give a systematic verbal (semi-formal) description and the access matrix of the registry model; some of the less relevant details are given in a slightly simplified form.

Every assumption of the model should be critically examined as to whether it is sound. For instance, can a party do things it is not supposed to do? What can two or more parties achieve through collaboration? The model will not give absolute security but will show where additional (organizational) means should be provided. The organizational framework has to guarantee the model assumptions and fill the security gaps that the cryptographic procedures leave open.

In discussing the security of the model we assume that the cryptographic algorithms are secure and that they are implemented in a secure way. The first assumption is justified by using state-of-the-art cryptographic techniques. The second assumption is more problematic and needs careful organizational measures.

4.1. Data and Parties

In the semi-formal description of the model we speak of *the* patient, *the* cooperating registry, *the* sequence number etc., although in reality there are several instances of each of these classes.

The knowledge (or data) in our model consists of the following parts:

- The identity data (see 3.1).
- The pseudonym
  - the encrypted identity (see 5.1),
  - the linkage data (see 5.3); they occur in 'pure hash' format, in 'linkage' format, in 'storage' format, and in 'exchange' format (see Fig. 5).
- The epidemiological data (see 3.1).
- The sequence number, a temporary pseudonym for a research project as in 3.4.
- The encryption key for asymmetric encryption of identification data.
- The re-identification key for re-identification of identity data.
- The linkage data key for generating the linkage data (see 5.3).
- The storage key for permanent storage of the linkage data (see 5.3).
- The exchange key for inter-registry record linkage (see 5.4).

Moreover, we have the identification data of the notifying institution for clearing discrepancies, for obtaining follow-up information, for reporting follow-up information in the case where the notifying institution is a clinical cancer registry, and for compensating the reporting physician for his notification. The trusted office also stores other administrative data.

The relevant parties for our model are the following; for each of these parties we have to define what knowledge it has or transfers and which other parties it trusts:

- The patient has access to his own data, but only via his treating physician.
- The notifying institution knows the data of its own patients:
  - The treating physician notifies the registry of his patients and can be asked by the trusted office about them.
  - Other health-care institutions which also send notifications are clinical cancer registries, after-care institutions, and Public Health offices.
- The trusted office sees all the data except the re-identification key and the storage key. It permanently stores only the encryption key and the linkage data key.
- The supervising office keeps the re-identification key and sees the identity data of re-identified cases.
- The registration office sees the pseudonym, the epidemiological data, the sequence number, the storage key, and also stores these data permanently (except the sequence number).
- The cooperating registry:
  - The trusted office sees the exchange key and the pseudonyms, even in pure hash format.
  - The registration office sees the linkage data in its own linkage format. In case of a match it gets the full registry data, which is the aim of the linking procedure.
- The research institute gets the sequence number and the epidemiological data as well as the exposure data which are outside the scope of the registry model (see 3.4).
- The outsider: any person or institution other than those listed above – has access only to communication paths and perhaps to storage media, if these leave the registration office, say, in case of a hardware defect.

The bank where the notifying physician has his account is ignored. Only a very small amount of information can be gained by observing the financial transfers, e.g., that a certain physician has a cancer patient at a certain time.

In the following we discuss only the parts of the model that are relevant for the pseudonymity aspect. For example, data on storage and communication media should be useless for the outsider; this is achieved by encryption of all communication paths and all storage media. In particular, the notifying institutions should communicate with the trusted office in a secure manner, i.e., using encrypted data transfer. Henceforth, we assume that the outsider can gain data access only through collaboration with some other institution, and leave the security of communication and storage outside the scope of this paper.

4.2. The Access Matrix

Figure 3 gives the access matrix of the registry model. We have to show that no party can get additional information by inferencing, in other words, that the access matrix as shown in Fig. 3 is complete. Since the model involves cryptographic keys, i.e., data that imply access to other data, the question is what subsets of the set of data in the access matrix are 'closed' with respect to inferencing. This gives only a 'naive' proof of security; there are indirect ways for getting additional informations (see section 4.3).

We have a single inference that needs no key:

$id \rightarrow ld_h$,

where the symbols are taken from Fig. 3 and the arrow denotes the inference. In other words: whoever has the identification data can derive the linkage data in pure hash format, because the hash algorithm is publicly known and needs no key. The complete list of key-dependent inferences is as follows:

k_e: $id \rightarrow ps$,
k_{re}: $ps \rightarrow id$,
k_{ld}: $ld_h \leftrightarrow ld_l$,
k_{st}: $ld_l \leftrightarrow ld_s$,
k_x: $ld_h \leftrightarrow ld_x$.

Therefore, the access matrix is complete. The only way to infer the identification data *id* is by knowledge of *ps* and k_{re}, the encrypted identification data and the re-identification key. Hence this can only be done by the supervising office.

4.3. Indirect Ways for Re-identification

The goal of the registry model is to make unauthorized re-identification as difficult as possible. However, what is possible, if the access matrix is guaranteed by the implementation of the model? The multitude and nature of indirect ways for making inferences about the data cannot be completely delineated. This is the main difficulty in proving the validity of any security model formally. Some relevant methods that should be considered are:

- trial encryption (guessed plain-text attack),
- data matching with outside sources [16],
- statistical attacks [16],
- covert channels [17],
- social engineering (voluntary or forced collaboration).

The outsider sees none of the data. He could gain access only by collaboration with another party.

The research institute sees the epidemiological data and could try an unauthorized matching with an external data source. This danger is inherent in the granularity of the epidemiological data and cannot be made smaller by any model whatsoever. Therefore, the release ofsubsets of epidemiological data is restricted according to a specific project.

The cooperating registration office only sees the linkage data in its own linkage format. It could try a statistical attack to find out some frequent names or use distribution anomalies of birth data. But this will hardly suffice to identify even a single case other than those that this registry has among its own records.

s = sees (and temporarily stores)
k = keeps (= permanently stores)
d = can derive

	Identification data [id]	Pseudonym (encrypted identity) [ps]	Linkage data (pure hash f.) [ld_h]	Linkage data (linkage format) [ld_l]	Linkage data (storage format) [ld_s]	Linkage data (exchange f.) [ld_x]	Epidemiological data [ep]	Sequence number [sq]	Encryption key [k_e]	Reidentification key [k_{re}]	Linkage data key [k_{ld}]	Storage key [k_{st}]	Exchange key [k_x]
Patient	s		d										
Notifying institution	k[1]		d				k[1]						
Trusted office	s	s	s	s		s	s	s	k		k		s
Supervising office	s[2]	s[2]	d							k			
Registration office		k		s	k		k	s				k	
Cooperating trusted office			s			s							s
Cooperating reg. office				s[3]									
Research institute							s	s					
Outsider													

Fig. 3 Access matrix of the registry model. [1] only own patients; [2] only re-identified cases; [3] in its own linkage format.

The cooperating trusted office sees the linkage data even in pure hash format and could perform a trial encryption. However, it is trusted by definition.

The registration office could try illegal data matching with the epidemiological data and a statistical attack at the linkage data in linkage format.

The supervising office sees the identity data of re-identified cases. However, it is also trusted, and it gets only few data.

The trusted office sees the identification data and the epidemiological data, but it is trusted by definition.

The notifying institution and the patient get no knowledge of data they should not know. They know their own data only.

The question what a party can do that has unauthorized knowledge of an additional piece of data, say, by collaborating with another party, can be answered by the analysis in section 4.2. Covert channels could be exploited, for instance, by faking notifications; we come back to this in section 7.1. Unauthorized matching with epidemiological data is only possible for an employee of the registration office or of the research institute; the trusted office that also sees the epidemiological data sees the identity anyway.

5. Encryption Procedures

Encryption of identifying data is performed by using different techniques which are suited for different purposes. A detailed technical description of the basic algorithms is given in [14]. As a basis to assess the performance of the procedures one has to take an expected number of 50,000 notifications each year for Rheinland-Pfalz. The efficiency of the procedures also suffices for larger registries.

5.1. Asymmetric Encryption of Identification Data

Asymmetric encryption techniques use two different keys for encryption and decryption, often called 'public key' and 'private key'. This notation, however does not fit in the present context. Therefore we speak of 'encryption key' and 're-identification key'. Knowledge of one of the keys does not help in any way to derive the other.

The identity data of each incoming record are encrypted in the trusted office using the encryption key, see Fig. 4. If, under special circumstances (as in 3.4), the decryption of some identification data becomes necessary, the registration office sends the encrypted identity data back to the trusted office that initiates the re-identification, see section 3.4.

The most suitable asymmetric encryption method, according to the state-of-the-art, is the RSA algorithm [14, 18, 19]. It uses the mathematical operation of modular exponentiation, $x \rightarrow x^e \bmod n$; character strings are treated as numbers according to their bit patterns and decomposed into blocks such that each block represents a number smaller than n. The modulus n is a very large number. The exponent e is the encryption key. The re-identification key d has a size similar to n and the property that $x^{ed} \equiv x \pmod{n}$. Thus, modular exponentiation with d is the inverse operation of modular exponentiation with e. Deriving e from n and d requires decomposition of n into its prime factors, a task that is mathematically infeasible, if n is large enough. Experts recommend a key length of >700 bits [20]. Since in a cancer registry data are stored for a long time, one should rather choose a key length of >1,000 bits to be prepared for possible technological progress. For performance reasons, instead of RSA one could use a hybrid encryption method [19, section V.1.7] such as RSA + DES or PGP (RSA + IDEA) [14, section 17.9]. This makes sense as soon as the data to be encrypted are longer than a single RSA block. DES and IDEA are symmetric encryption procedures, meaning that encryption and decryption use the same key. The exact description is too complicated to be given here; we refer to [14, 17]. They are several orders of magnitude faster than all known asymmetric procedures but do not fit directly to our model which relies on asymmetric encryption. Therefore, a hybrid combination with RSA has to be used.

If an employee of the registration office gains knowledge of the encryption key, or if an outsider gains knowledge of the encryption key and access to the registered data, he could perform a trial encryption ('chosen plain-text attack') with the corresponding identity data. In order to prevent this possible misuse, each record is complemented by a random number before encryption. As shown in Fig. 4, this random number is kept in the encrypted part of the record.

5.2. Key Management

The keys have to be generated in a secure manner under special organizational precautions, e.g., in the supervising office. The encryption key is kept in the trusted office. It has not necessarily to be kept secret because the encryption is randomized (see section 5.1). Therefore, there is no need for a cryptographic token, like a smart card, to hold this key. But a smart card is desirable as access-control token. It could then also hold the key. On the other hand, the 'need to know' principle says that it is better keeping the key secret.

There are two cases where a change of the encryption and re-identification keys becomes necessary:

- The actual keys are compromised; at least there is suspicion that an unauthorized person has got the keys.
- The progress of cryptanalysis or the performance of hardware have advanced to a great extent such that the chosen key length can no longer be assumed to be sufficient.

In these cases a new, more secure pair of encryption and re-identification keys has to be generated and used. This could be done by decrypting and then re-encrypting all the stored records in the trusted office. However, the German BSI ('Bundesamt für Sicherheit in der Informationstechnik', Federal Office for Security in Information Technology) proposed a more efficient method: define the new encryption method to be the composition of the old one and the "over-encryption" with the new key, thereby avoiding even a temporal exposition of the plain-text data; the future decryption key is the composition of the old and the new keys. Over-encryption of the old records can be done in the registration office under special security precautions. An analogous procedure also applies in case the chosen encryption method is invalidated by new research results.

An alternative method to handle key changes without temporarily generating plain text was proposed by Miller [21]. It eliminates the need of superimposing the old and new encryption procedures and keeping the old key. On the other hand, it works only with a slightly restricted version of the RSA algorithm.

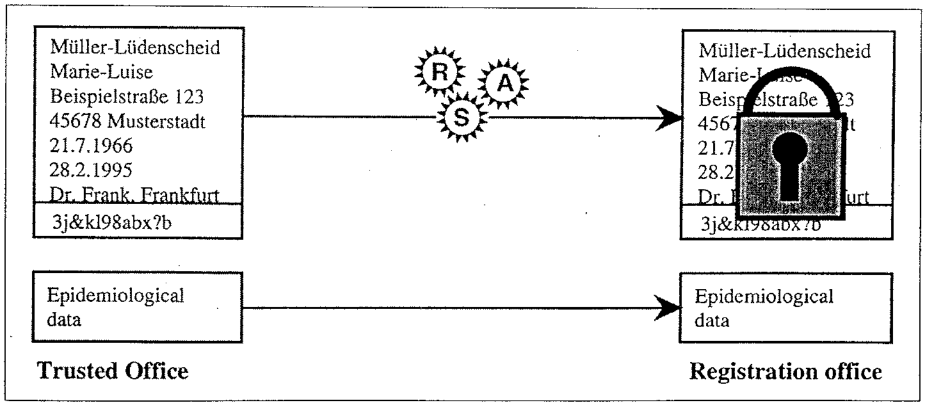

Fig. 4 Asymmetric encryption of identification data.

5.3. Linkage Data and Anonymous Data Matching

To generate the linkage data we extract the following components from the identity data: Name(s), surname(s), phonetic codes, the name code of the former GDR, day and month of birth. Then these components are separately encrypted, in a first step by using a one-way hash function [14], in a subsequent step by using a symmetric encryption algorithm [14] with the 'linkage-data key'; then they are in 'linkage format'. For permanent storage in the registration office the encrypted components are combined, complemented by a random number and once more encrypted using a symmetric algorithm with an independent key, the 'storage key'. The first key is kept in the trusted office, the second key is kept in the registration office. Both keys are secrets of their owners. The use of the one-way algorithm prevents any direct decryption of the identification data. The additional encryption steps and addition of a random number prevent any trial encryption. This would only be possible by illegal cooperation of the trusted office and the registration office. The second encryption in the registration office has the effect that the linkage data in the linkage format appear only in the main memory of the registration office's computer and never in permanent storage. Only a person who has the storage key can use the linkage data for record linkage. The storage encryption of the linkage data could be omitted if we assume that data are stored via a cryptographic device driver [19, section II.5.2].

As one-way hash function we use the MD5 algorithm [14], for symmetric encryption the DES algorithm [14, 19]. MD5 gives a cryptographically strong check sum (hash value) of 128 bits; it is computationally infeasible to find another plain-text string that gives the same check sum – the probability that another string has the same hash value is $1/2^{128}$. This procedure alone does not protect against a trial encryption, hence the key-dependent encryption in the second step. To prevent a trial encryption attack by outsiders, the linkage-data key has to be kept secret, e.g., on the smart cards of the employees.

Because record linkage with other cancer registries in Germany is planned, all these registries have to use the same linkage-data algorithm, but should use independent keys. If requirement 3 of section 2 were not important, one could employ another key-dependent one-way hash function with better performance, see [14, section 14.14], instead of MD5 + DES (e.g., MD5 with a secret 128-bit appendix to the plain text). This would prevent the registries from the procedure in section 5.4.

Figure 5 illustrates the various formats that the linkage data assume. From left to right the security increases: The clear-text format shows the full information; the pure hash format allows trial encryption and record linkage; the linkage format allows record linkage only; and the storage format gives complete anonymity.

For record linkage the registration office compares the linkage data and other unencrypted identifying data of a new case with all the stored records. In case of small differences, if there is a reasonable evidence of match, the case is reported back to the trusted office that tries to clarify the case. In very few exceptional cases this procedure could necessitate a re-identification as in section 3.4.

5.4. Inter-registry Matching

From time to time, e.g., once per year, the collaborating registries are allowed to link their records in order to detect common notifications, e.g., caused by change of residence, or notifications by a treating physician and a hospital in the hinterland of the other registry.

For this purpose two registries *A* and *B* agree upon a temporary one-time 'exchange' key. Registration office *A* transfers a file with the linkage data to its trusted office which removes the encryption, getting the 'pure' hash values, and encrypts these with the exchange key. Then it sends them to the trusted office of registry *B*, which removes the exchange encryption and does the usual linkage-data encryption for its associated registration office. The same procedure applies in the other direction.

The weak point in this procedure is the removal of the linkage-data encryption at the two trusted offices; here, the linkage data are exposed to a trial encryption attack by everyone who gets them. This is no problem with the employees who are subject to professional discretion, but requires additional security precautions in any case. An alternative procedure proposed by Miller [6] avoids this weakness. Unfortunately, it uses an algorithm that is too slow to be of practical value. Another organizational approach is to match the data of all cooperating registries in a common exchange format at a central office.

5.5. The Security of the Linkage Data

The linkage data, together with the epidemiological data, could be used for an unauthorized matching attack. The better the registry can minimize the errors in record linkage, the easier an unauthorized data matching works. We have conflicting goals: the re-identification of cases by the linkage data should be as difficult as possible, but they should provide sufficient data quality for epidemiological research. The re-identification need not be more difficult than by the epidemiological data. One of the usual proposals for the anonymization of statistical databases is the intentional addition of errors [16, section 1.1.6]. The linkage data give rise to the contrary effect: they facilitate the data matching despite errors in the data. This weakening of data protection has to be compensated by organizational measures according to the access matrix.

A change of the linkage-data key makes it necessary to transfer all linkage data in linkage format back to the

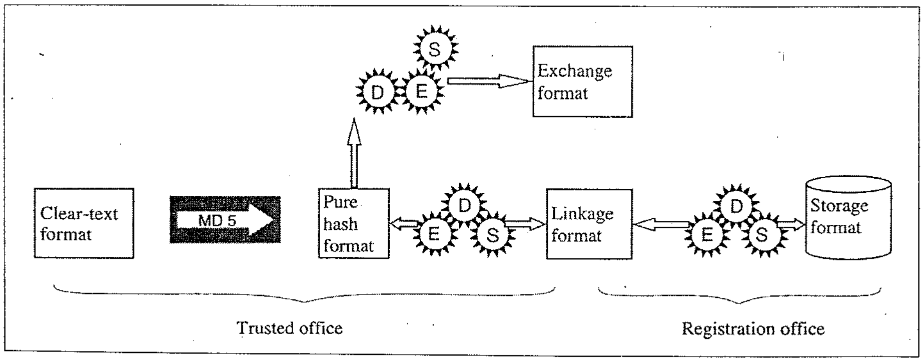

Fig. 5 Formats of linkage data.

trusted office that decrypts them to pure hash format and re-encrypts them with the new key. This procedure is analogous with the procedure for exchange with a cooperating registry and, therefore, not a major problem. A change of the storage key is even easier; it takes place in the registration office and involves no other party.

5.6. Random Generation

Several of the procedures use random numbers:

- key generation for encryption and re-identifying keys,
- key generation for linkage data and storage keys,
- key generation for exchange (one-time) keys,
- randomization of the asymmetric encryption,
- randomization of the linkage-data encryption (in the registration office).

These random numbers should be reasonably secure. Therefore, a source of 'true' random bits is needed. For appropriate procedures we refer to [22].

6. Record Linkage Study

In this section we show what order of magnitude of errors has to be expected with the proposed pseudonymous record linkage procedures.

6.1. Approach

Record linkage is performed by a stochastic approach using an algorithm proposed by Fellegi and Sunter [23] and modified by Jaro [24]. The basic idea of this approach is that two records referring to the same person show better agreement than two randomly selected records. The agreement can be described by a vector γ. Each of its components refers to one of the attributes considered for record linkage. The probability $m(\gamma)$ to observe γ when two records refer to the same person is obtained from knowledge of the probability for miscoding and change of attributes. The probability $u(\gamma)$ to observe γ for two randomly selected records can be estimated from the frequencies of values in the data sets that are to be matched. When observing the agreement vector γ, $m(\gamma)/u(\gamma)$ is computed. If this ratio exceeds a specified threshold S_1, the record pair is considered a match; if the ratio is below another threshold S_2, the record pair is considered a non-match. If $S_1 > S_2$, record pairs with ratios in the "grey area" between S_1 and S_2 should be clerically resolved.

For our present analyses we used the computer program AUTOMATCH provided by Matchware Technologies Inc., Silver Spring, MD [25]. Technically, there is no difference between using original data or pseudonyms for record linkage as both lead to identical results, if similarities in the clear-text data are neglected that have no counterpart in the pseudonyms. The effectiveness of different identifiers was evaluated by using the following data sets:

- 75,694 hospital records from the Mainz University Hospital which contained 27,829 multiple records from the same patients, and
- 62,881 records from the national cancer registry of the former German Democratic Republic (GDR) containing 907 multiple notifications.

The identity data in the multiple records were partly correct (= identical) and partly contained spelling errors, or different names or birth dates. In these data sets the record linkage was performed for the purpose of identifying the multiple records. A third data set was used in order to study the record linkage of cancer-registry data with death certificates. This was performed with 115,175 records from the cancer registry of Saarland and 8,731 death certificates which were received in Saarland in 1992.

Table 1 Results of AUTOMATCH application to different data sets.

Data Set	Homonym error (%)	Synonym error (%) (1)	(2)	(3)
75,694 hospital records	0.62	0.14	0.37	1,69
62,881 records from former GDR cancer registry	0.74	0.02	1.2	14
123,906 records and death certificates from the Saarland cancer registry	0.16	<0.01	2.2	2.2

Two types of errors can be made in record linkage: false-negative matches lead to synonym errors, e.g., a single individual is counted twice in the calculation of incidence rates on the basis of two separate reports which cannot be matched. False-positive matches lead to homonym errors, e.g., a surviving cancer patient is falsely counted as dead if his record was linked to the death certificate of another individual. Whereas it appears straightforward for record-linkage studies to calculate the homonym error rate by dividing the number of false positive matches by the number of synonym-free records in a given data set, it is not so obvious which denominators one should use for the calculation of synonym error rates. In order to describe the results of our studies we use the following denominators (see Table 1):

1. The number of all records involved in the study,
2. the number of all multiple records, and
3. the number of multiple records which do not have identical identifying data.

6.2. Results

Table 1 shows results from linkage analyses where we used name(s), surname(s), the three components of the birth dates and either zipcode and place of residence (Mainz data), or census code of place of residence (GDR and Saarland data) (year of birth and place of residence are taken from the epidemiological data). One can see that with the automated procedure the homonym error could be kept below 1% in all data sets as well as the synonym error of definition 1. The other definitions of synonym errors lead to higher rates. This partly reflects problems which also occur in the process of manual matching. For example, all multiple records with deviant identifying data from the former GDR registry which represented the denominator for the synonym error 3 had been detected by the automated record-linkage procedure and had been previously unknown to the staff of the registry in spite of manual matching.

The automated record-linkage procedure can marginally be improved by using a phonetic transcription of the names, which is comparable to the SOUNDEX procedure and by taking into account possible permutations of the components within composed names. Additional improvement of the record linkage may also be achieved by a clerical review of individual records taking into account additional information (e.g., diagnosis, reporting physician, etc.) within certain thresholds of the linkage algorithm.

7. Discussion

7.1. Other Security Considerations

The distinction between 'sees' and 'keeps' in the access matrix is weak. It assumes that the involved party has no memory. But whoever sees some data could clandestinely store them. In particular, an employee of the trusted office could build his own private registry. This can be prevented by organizational counter-measures only, in particular by auditing data processing and data export.

Whoever has access to the epidemiological data could use a covert channel [17] by spoofed notification: *A* fakes a notification for *B*, say, with some unusual data, and by observing the effect on the registry reidentifies *B*'s record [26]. This can hardly be prevented. However, it is quite unlikely because it assumes the collaboration with a notifying institution or an employee of the trusted office.

Another goal that could motivate faked notifications is the forgery of statistical results from the registry. This could be to disguise potential causes for cancer. It would, however, require spoofed mass notifications. Another way is by collaboration with an employee of the registration office. This can be prevented by counters for notifications, cryptographic check sums of the database, and organizational measures. Faking of notifications to get some money does not pay because the compensation for a notification is quite low.

A problem that is also hard to avoid is that an employee of the trusted office or the supervising office can see his own record. This violates requirement 5 of section 1 and is unwanted if he is not told about his disease by his physician. Also an employee of the registration office could recognize his own record, as could an employee of the research institute. For these cases the access matrix of section 4.2 is not accurate because a person assumes more than one role at the same time.

Data matching with the linkage-data requires the linkage-data key and the file of linkage-data, i.e., the collaboration with an employee of the trusted office and the registration office. Because of the fine granularity of the linkage data, however, also a statistical attack on them is relatively easy. This can be done only by the registration office and the cooperating registration office, and also by the trusted offices in the short time period where they have access to the file of linkage data for interregistry matching.

The epidemiological data alone could help in identifying an individual in certain cases. Note, however, the coarse classification of the attributes; an attacker would need quite detailed additional information to gain an advantage in matching the data to outside data sources [16]. On the other hand, there is no point for the pseudonym in supplying more anonymity than the epidemiological data do.

7.2. The Matching Errors

The results from the record-linkage study show that the matching errors are within tolerable limits. Synonym errors lead to a small increase of the estimated incidence and survival times which appear to be negligible with respect to other possible errors which occur even in careful cancer registration. If falsely not matched records are used for case-control studies, this may lead to the detection and subsequent correction of the corresponding errors. The observed homonym errors would lead to a small decrease of incidence estimates. They are not likely to be detected if one performs case control studies. Therefore it appears reasonable to keep the homonym error rate especially low. If false-positive or false-negative mismatches occur with the record linkage which is performed for conducting cohort studies as described in section 3.4, this may lead to over- or underestimation of standardized incidence or mortality rates. If control groups are also linked with the registry, matching errors will lead to decreased estimates of risk ratios due to nondifferential misclassification. However, the observed error rates show that all these possible consequences will be relatively small with respect to other possible sources of errors in epidemiological studies.

8. Conclusion

The use of adequate data encryption techniques prevents the access to identification data by unauthorized persons and thus leads to enhanced security of the long-term data storage in cancer registries. It must be realized, however, that encrypting identification data leads only to quasi-anonymized records. Reidentification of individual data will be possible in some instances by combining identifying data from other sources with epidemiological data.

Using very strong cryptographic algorithms seems to be an overkill in view of certain weaknesses in the overall scheme. On the other hand, there is no point in using weaker algorithms – they have a similar performance behavior and similar requirements on the organizational and technological infrastructure.

There could be different kinds of pseudonyms that are suited for error detection and correction in data matching that better minimize the goal conflict between data quality and anonymity. Since the problem is a classification task, maybe an approach using neuronal nets could lead to a better solution. With our concept of pseudonyms, the matching errors are within tolerable limits for scientific analyses of the registry data. All the requirements of section 2 are satisfied. Paper [7] gives an overview of the current application of the model for cancer registry in Germany.

The use of pseudonyms for long-term storage of person-related data may also be adequate for other situations. The application of asymmetric encryption techniques should be considered for routine application within medical communication systems. Our finding that automated record linkage with pseudonyms may be performed with very small error rates opens a perspective for future epidemiological research which takes into account the increasing demand for data protection. Thus our model of a cancer registry may serve as a model for other research with person-related data.

Acknowledgments

This work was supported by the Ministerium für Arbeit, Familie, Soziales und Gesundheit Rheinland-Pfalz and the Bundesministerium für Gesundheit. We thank P. Mustroph, Berlin, and Ch. Stegmaier, Saarbrücken, for their help in the record-linkage testing. We also thank the 'Bundesamt für Sicherheit in der Informationstechnik' (BSI) for helpful comments on several aspects of the proposed procedures.

REFERENCES

1. EU Directive on Data Privacy, http://www.rewi.hu-berlin.de / Datenschutz / EURichtlinie/directive.html.
2. Gesetz über Krebsregister (Krebsregistergesetz), Bundesgesetzblatt Nr. 79 vom 11.11.1994.
3. Krtschil A, Michaelis J. Aufbau des bevölkerungsbezogenen Krebsregisters für Rheinland-Pfalz. Ärzteblatt Rheinland-Pfalz 1992; 45: 434-8.
4. Schmidtmann I, Michaelis J, Pommerening K. Pilotstudie zum Aufbau eines bevölkerungsbezogenen Krebsregisters in Rheinland-Pfalz. In: Pöppl SJ, Lipinski HG, Mansky T, eds. *Medizinische Informatik – ein integrierender Teil arztunterstützender Technologien, 38. Jahrestagung der GMDS.* München: MMV Medizin Verlag, 1993: 399-403.
5. Michaelis J, Krtschil A, Schmidtmann I, Schüz J. Bundeskrebsregistergesetz und Stand der Pilotstudie "Krebsregister Rheinland-Pfalz". Ärzteblatt Rheinland-Pfalz 1995; 48: 71-4.
6. Miller M, Michaelis J, Pommerening K. Cryptographic protection for cancer registries. In: Pernul G, ed. *IT-Sicherheit '94.* Wien: Oldenbourg Verlag, 1995: 239-46.
7. Michaelis J. Towards nation-wide cancer registration in the Federal Republic of Germany. Ann Oncol 1995; 6: 344-6.
8. Michaelis J, Miller M, Pommerening K, Schmidtmann I. A new concept to ensure data privacy and data security in cancer registries. In: Greenes RA, Peterson HE, Protti DJ, eds. *MEDINFO 95.* Edmonton: Healthcare Computing & Communications Canada, 1995: 661-5.
9. Thoben W, Appelrath HJ, Rettig J, Sauer S. Berücksichtigung von Datenschutzaspekten in einem bevölkerungsbezogenen Krebsregister. In: Kunath H, Lochmann U, Straube R, Jöckel KH, Köhler CO, eds. *Medizin und Information. 39. Jahrestagung der GMDS.* München: MMV Medizin Verlag, 1994: 88-90.
10. Struif B. Datenschutz bei elektronischen Rezepten und elektronischem Notfallausweis. In: *Vertrauenswürdige Informationstechnik für Medizin und Gesundheitsverwaltung.* Erfurt: TeleTrusT Deutschland e.V., 1994: 15/1-6.
11. Pommerening K. Datenschutz in Krankenhausinformationssystemen. In: Brüggemann HH, Gerhardt-Häckl W, eds. *Verläßliche IT-Systeme VIS '95.* Braunschweig: Vieweg, 1995: 5-22.
12. Chaum D. Security without identification: Transaction systems to make Big Brother obsolete. Comm ACM 1985; 28: 1030-45.
13. Knuth DE. *The Art of Computer Programming, Vol. 3, Sorting and Searching.* Reading MA: Addison-Wesley, 1973.
14. Schneier B. *Applied Cryptography,* New York: John Wiley, 1994.
15. Biskup J, Bleumer G. Reflections on security of database and datatransfer systems in health care. In: Brunnstein K, Raubold E, eds. *Applications and Impacts – Information Processing '94 (IFIP).* Amsterdam: North Holland, 1994: 549-56.
16. Paaß G, Wauschkuhn U. *Datenzugang, Datenschutz und Anonymisierung.* München: Oldenbourg Verlag, 1985.
17. Denning DE. *Cryptography and Data Security.* Reading MA: Addison-Wesley, 1982.
18. Beutelspacher A. *Kryptologie,* 3rd ed. Braunschweig: Vieweg, 1993.
19. Pommerening K. *Datenschutz und Datensicherheit.* Mannheim: BI-Wissenschaftsverlag, 1991.
20. *RSA Lab's Frequently Asked Questions about today's cryptography.* http://www.rsa.com/rsalabs/faq/, 1996.
21. Miller M. *RSA-Schlüsselwechsel in einer unsicheren Umgebung,* Bericht 13, Musikinformatik und Medientechnik. Mainz: Musikwissenschaftliches Institut der Johannes Gutenberg-Universität, März 1994.
22. Eastlake D, Crocker S, Schiller J. *Randomness Recommendations for Security,* Network Working Group, Request for Comments 1750, ftp://ds.internic.net/rfc/rfc1750, 1994.
23. Fellegi IP, Sunter AB. A theory for record linkage. J Am Stat Ass 1969; 64: 1183-210.
24. Jaro M. Advances in record linkage methodology as applied to matching the 1985 census of Tampa, Florida. J Am Stat Ass 1989; 84: 414-20.
25. Jaro M. Probabilistic linkage of large public health data files. Stat Med 1995; 14: 491-8.
26. Miller M. *Verfahren zur Anonymisierung und Registrierung in Krebsregistern,* Bericht 11, Musikinformatik und Medientechnik. Mainz: Musikwissenschaftliches Institut der Johannes Gutenberg-Universität, Dezember 1993.

Address of the authors:
Prof. Dr. Klaus Pommerening
Institut für Medizinische Statistik und Dokumentation
der Johannes-Gutenberg-Universität
D-55101 Mainz
Germany
Email: Pommerening@imsd.uni-mainz.de