e-Health – For Continuity of Care
C. Lovis et al. (Eds.)

Matching Study to Registry data: Maintaining Data Privacy in a Study on Family based Colorectal Cancer

Daniel NASSEH [a], Jutta ENGEL [a,b], Ulrich MANSMANN [a], Werner TRETTER [b] and Jürgen STAUSBERG [a]
[a] *IBE, Ludwig-Maximilians-Universität München, Germany*
[b] *Munich Cancer Registry, Germany*

Abstract. Confidentiality of patient data in the field of medical informatics is an important task. Leaked sensitive information within this data can be adverse to and being abused against a patient. Therefore, when working with medical data, appropriate and secure models which serve as guidelines for different applications are needed. Consequently, this work presents a model for performing a privacy preserving record linkage between study and registry data. The model takes into account seven requirements related to data privacy. Furthermore, this model is exemplified with a study on family based colorectal cancer in Germany. The model is very strict and excludes possible violations towards data privacy protection to a reasonable degree. It should be applicable to similar use cases which are in need of a mapping between medical data of a study and a registry database.

Keywords. Confidentiality, record linkage, colonic neoplasms, genealogy

Introduction

Handling personal data is not only a matter of trust. In the worst case, leaked information can be adverse to and abused against an individual and cause severe damage. Therefore, in different fields of application, it is important to supply methods and models that provide a strong security environment. Especially when working with medical data, in particular, patient data, which often contain sensitive information, a high standard concerning obligation of secrecy needs to be maintained. One field of application in this domain is the task of medical record linkage where medical data of different datasets has to be mapped to each other. This is not a trivial task especially if patients' anonymity has to be assured at any stage of the process. In this case, the record linkage is called privacy preserving [1].

In 2013, in Germany, an ongoing study concerning family based colorectal cancer (CRC) has to face this task [2]. There are multiple risk factors contributing to the development of CRC like smoking, lack of exercise, wrong eating habits and predominantly high age [3]. Aside from known genetic dispositions [4] there has been the observation of accumulations of cases of CRC within families. This means, having a family member with CRC is a risk factor in itself and is referred to as family based CRC [5].

Genealogical links are not documented in cancer registries and are not available within the Munich Cancer Registry (MCR) itself (cf. www.tumorregister-muenchen.de). Therefore, the study design aims at gathering family information from newly diagnosed patients who accept to participate in the study. Information needed are identifying data, in specific, first name, last name, date of birth, address and gender of all close relatives of the patient. This data is matched to the MCR by using a probabilistic record linkage approach in order to identify study patients and relatives within the cancer registry who have been diagnosed with CRC or associated kinds of cancer in the past. Their medical data as well as their genealogical information has then to be passed to the analysis center for further research. Information about family members not registered in the MCR contributes to the study's finding as well.

The study design demands a highly complex safety infrastructure including multiple parties in order to fulfill the obligations based on the strict laws of privacy protection in Germany. In this work's section of methods we give a detailed overview about these obligations and requirements and a description of the model itself. In the section of results we present how we fitted the model to the study of family based CRC.

1. Methods

As postulated in guidelines regarding data privacy protection in a medical environment [6] one of the requirements for a model of data protection is the conceptual and institutional division of all participating parties. In case of performing a data matching, this should result in four different parties as illustrated in figure 1. These are typically the two institutions managing the study as well as the registry datasets which are supposed to be mapped together, the location of an independent data trustee whose main task is to perform the matching process as well as the center which will perform the final analysis. The standard procedure used for matching data is referred to as record linkage [7].

In the case of patient data, attributes like first name, last name or date of birth within the identifying data (IDAT) are used to match the patients within the different datasets. However, the IDAT should not be readable for any other participating institution. Consequently, a variant of a standard probabilistic record linkage, so called, anonymous or privacy preserving record linkage is needed that operates on one way encrypted attributes instead of attributes written in plain text. One way encryption can be achieved by applying hash functions to a string of text [8] or alternatively representing the string as a bloom filter filled with different hash values [1]. This task has to be performed by the data trustee. Before sent to the data trustee, the IDAT are marked with an institution specific ID which is a unique identifier as well as a reference to the corresponding medical data (MDAT) labelled with the same ID.

To expand the model, the third stipulation would be that MDAT should in general, aside from its originating institution, only be readable by the center of analysis. As discussed later, at the location of the data trustee a conversion of the study ID within the IDAT and consequently also within the study MDAT is required. Thus, the study's MDAT are sent to the data trustee as well. Due to the fact that it should not be readable by any other institution but the center of analysis the contents of the study MDAT have to be symmetrically encrypted before. Following, the result of the record linkage process is a list of pairs of IDs, referred to as links [9], describing which patients of the

different datasets relate to the same entity. The objective of these links is to request the MDAT of the registry according to the registry's ID within the links.

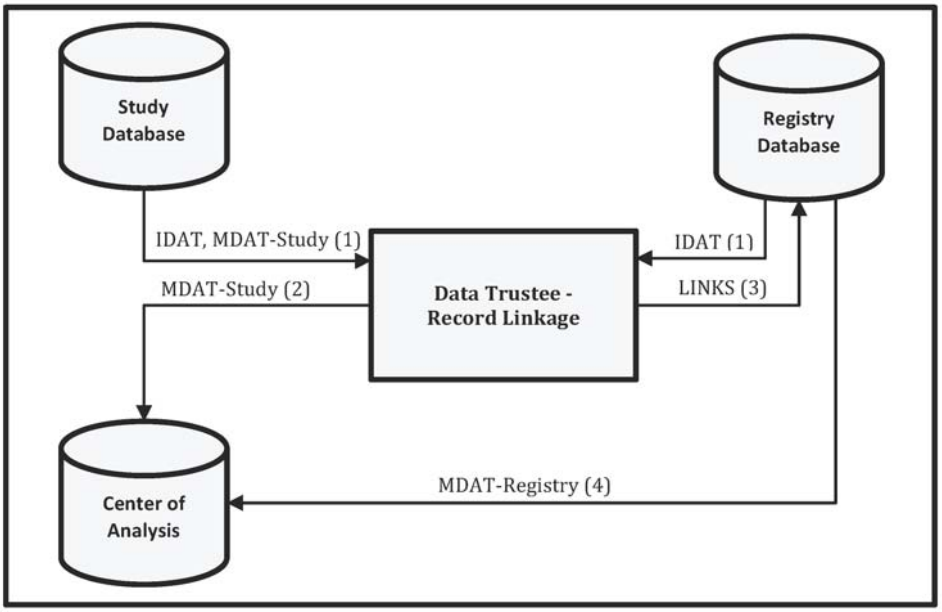

Figure 1. (1) IDAT and MDAT of the study database as well as IDAT of the registry database are sent to the data trustee. (2) After replacing the ID within IDAT and MDAT of the study database with a new random ID, the MDAT of the study database is redirected to the center of analysis. (3) After the record linkage process, the links (pairs of IDs) are sent to the registry database. (4) The requested MDAT of the registry are sent to the center of analysis. Based on the new random ID, the MDAT of the study as well as the registry database can be mapped to each other.

In order to prevent a possible identification based on the information within the MDAT, another requirement should be the realization of *k*-anonymity on the quasi identifiers within this data [10].

As a fifth request, the model demands that study and registry IDs may only be known by the original institutions as well as the data trustee. Thus, the IDs have to be replaced by a new random ID through the data trustee. This new random ID is the only ID transferred to the center of analysis.

Premise six requires that the new random ID may only be shared by the data trustee and the center responsible for analysis. Therefore this ID has to be symmetrically encrypted within the links before shared with the registry. According to the registry's ID within the links, the registry can now extract the needed MDAT, map it to the encrypted new random ID, remove the registry's ID and send the MDAT along with the encrypted new random ID to the center of analysis. In parallel, the center of analysis receives the encrypted MDAT of the study labeled with the same new random ID from the data trustee. At this point, both the MDAT of the registry as well as the study can be mapped together based on the new random ID.

To offer an additional layer of security, all transported data handled between institutions should be asymmetrically encrypted throughout the whole process. All used methods should be considerably save and up to date according to national laws, in this

case according to the recommendations of the BSI which is the bureau for security in IT technology in Germany [11].

2. Results

Within the study of family based CRC an institutional as well as conceptual division is given for both the MCR as well as the study database. The role of the data trustee is performed by an independent person which is the data protection commissioner of the University Hospitals of the Ludwig-Maximilians-Universität (LMU) while the center of analysis is represented by a statistical workgroup of the Institute of Medical Information Processing, Biometry and Epidemiology of the LMU.

The used record linkage system is a probabilistic record linkage system based on the widely used algorithm of Felligi and Sunther [12] as well as on technical guidelines formulated by Martin Meyer [13]. The system has been implemented in java 1.7 and uses some custom technologies to better adapt to the scenario. Some of these customizations have been described in a previous publication [14]. Precedent to the record linkage, the IDAT of both the MCR as well as the study database are first standardized according to UNICON guidelines [15] and one way encrypted by using a hash function of the SHA-2 family [16].

The study's MDAT consist of simple questions as well as the genealogical relations (family structure) of the patients and their relatives. To render this information unreadable when passed to the data trustee it is symmetrically encrypted by using the AES algorithm with a block length of 128bit [17]. The method has been implemented and adapted to the scenario in java by using the packages java.security.* as well as javax.crypto.*. The key to decrypt the MDAT is only shared by the study database and the center of analysis. *K*-anonymity for both the study and registry MDAT has not been done since there were no quasi-identifiers being strong enough to violate the concept of anonymity to a disconcerting degree. At the location of the data trustee and according to the model, a new random ID replaces the old ID within the study's IDAT, MDAT and consequently within the generated links. To accommodate to the sixth requirement the ID within the links is symmetrically encrypted using the same technology as previously explained before sent to the MCR. Once again the key is only given to the center of analysis.

For most transported data, a hybrid encryption system, which is a variant of asymmetrical encryption, has been implemented using the RSA/AES algorithm [18]. The used key length is 2048 bits. The technology has been adapted to the scenario and written in java.

3. Discussion

The presented model can be used as a template for studies which base their analysis on matching study data to data of medical registries. It is very strict in its requirements and therefore eliminates possible violations endangering the privacy of the patients' data. A positive approval in regard of law and ethics by the LMU's ethical review committee has been given in December 2012.

Due to the strictness of the model, comprehensive logistic and organizational efforts are needed. Therefore this model should be best suited for studies of larger scale.

As presented in the results section of this work, the requirements could be successfully applied to the study of family based CRC in Germany. One major task was maintaining the family structure throughout the whole process which could be achieved by modeling the family structure as medical data. Based on the family structure, the center of analysis is capable of rebuilding all genealogical relations between MDAT of patients and their relatives.

Because of the replacement of IDs during the whole process, it is important to note that this model is not applicable to cumulative record linkage. This means if new patients are recruited, the record linkage has to be performed on the whole dataset and not only on the new fraction of patients.

References

[1] Schnell R, Bachteler T, Reiher J. Privacy-preserving record linkage using Bloom filters. *BMC Medical Informatics and Decision Making*. 2009; 9:41

[2] Mansmann U, Stausberg J, Engel J, Heussner P, Birkner B, Maar C. Familien schützen und stärken – Umgang mit familiärem Darmkrebs. *Gastroenterologe*. 2012; 161-162.

[3] Watson AJ, Collins PD. Colon cancer: a civilization disorder. *Digestive diseases*. 2011;29(2):222-8.

[4] Half E, Bercovich D, Rozen P. Familial adenomatous polyposis. *Orphanet journal of rare diseases*. 2009 Oct 12;4:22.

[5] Slattery ML, Levin TR, Ma K, Goldgar D, Holubkov R, Edwards S. Family history and colorectal cancer: predictors of risk. *Cancer Causes Control*. 2003 Nov;14(9):879-87.

[6] Pommerening K, Drepper J, Ganslandt T, Helbing K, Müller T, Sax U, Semler S, Speer R. Das TMF-Datenschutzkonzept für medizinische Daten-sammlungen und Biobanken. Paper presented at: Proceeding of: Informatik 2009: Im Focus das Leben, Beiträge der 39. Jahrestagung der Gesellschaft für Informatik e.V. (GI). 2009;

[7] Christen P. *Data Matching – Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Heidelberg: Springer; 2012.

[8] Kijsanayotin B, Speedie SM, Connelly DP. Linking patients' records across organizations while maintaining anonymity. *AMIA Annu Symp Proc.* 2007; 1008.

[9] Blakely T, Salmond C. Probabilistic record linkage and a method to calculate the positive predictive value. *International Journal of Epidemiology*. 2002 Dec; 31(6):1246-52.

[10] L Sweeney. K-anonymity: A model for protecting privacy. *International journal of uncertainity, fuzziness and knowledge-based systems*. 2002; 10(5).

[11] BSI - Technische Richtlinie, Kryptographische Verfahren: Empfehlungen und Schlüssellängen, BSI TR-02102. 2013;

[12] Felligi I, Sunter A. A theory of Record Linkage. *American Statistical Association Journal*. 1969; 64:1183-1220.

[13] Meyer M. Kontrollnummern und Record Linkage. *Das Manual der epidemiologischen Krebsregistrierung.* Hentschel S, Katalinie A, editor. Zuckschwerdt. 2011;57-68.

[14] Nasseh D, Stausberg J. Impact of variations in Anonymous Record Linkage on Weight Distribution and Classification. Poster presented at: 14th World Congress on Medical and Health Informatics, Kopenhagen. 2013;

[15] Hinrichs H. Bundesweite Einführung eines einheitlichen Record Linkage Verfahrens in den Krebsregistern der Bundesländer nach dem KRG, Abschlussbericht, Projekt Deutsche Krebshilfe. Antragsnummer 70-2043-Ap I. OFFIS. Oldenburg; 1999

[16] Gilbert H, Handschuh H. Security Analysis of SHA-256 and Sisters. *Selected Areas in Cryptography*. 2003; 175–193

[17] Daemen J, Rijmen V. AES Proposal: Rijndael. 1999;

[18] Palanisamy V, Jeneba M. Hybrid cryptography by the implementation of RSA and AES. *International Journal of Current Research.* April 2011;33(4): 241-44.