Title: Pseudonymization of patient identifiers for translational research

Authors:

Harald Aamot (harald.aamot@nct-heidelberg.de)
Christian Kohl (christian.kohl@med.uni-heidelberg.de)
Daniela Richter (daniela.richter@nct-heidelberg.de)
Petra Knaup-Gregori (petra.knaup@med.uni-heidelberg.de)

Version: 3 Date: 19 April 2013

Author's response to reviews: see over
Dear Mr. Aldcroft,
Dear Prof. Dr. Pommerening,
Dear Prof. Dr. Mandl and Dr. Wieland,

Please find in this letter a point-by-point response to the annotations provided by the reviewers of our initial manuscript.

We’d like to thank the reviewers for their thoughtful, relevant and encouraging comments and questions. Their feedback enabled us to revise, restructure and considerably improve the manuscript.

We believe that the revised manuscript addresses the issues raised, as detailed on the following pages of this document.

Please also find enclosed our revised manuscript as a separate document which we re-submit for publication.

Kind regards,

Harald Aamot

“Pseudonymization of patient identifiers for translational research” - revised manuscript in light of the reviewers’ comments

Heidelberg, 24.03.2013
Major Compulsory Revisions


The authors should compare their approach with the one given there. What then is innovative in the authors’ approach?

As suggested, the epidemiological pseudonymization procedure has been included in the comparison of pseudonymization approaches in our revised manuscript.

We now present our translational pseudonymization method as a variant of the existing “Pseudonyms for cancer registries” method, specifically adapted to fulfill translational research requirements.

The following innovations over the existing method have been identified:

a) The pseudonymization service provider is relieved from any legal obligation. This allows for a clean separation of duties concerning pseudonymization and re-identification.

b) No additional manual step for a permitted re-identification of a patient is required.

c) Several and flexible ombudsmen are introduced. This addresses a variety of informed consents and circumstances of concrete research projects.

The presentation of the “Pseudonyms for cancer registries” method is a chapter in the revised manuscript and would exceed this cover letter. Here, we only refer to the change in the abstract:

Therefore, both person-centricity issues and a separation of pseudonymization and re-identification stood out as a central theme for our examination. This motivated us to enhance an existing pseudonymization method regarding a separation of duties.

2. The legal position of the proposed "ombudsmen" is not clear. There is a thorough expertise on the legal requirements for trusted third parties (in German) available online from the TMF website:

http://www.tmf-ev.de/Produkte/Uebersicht/ctl/ArticleView/mid/807/articleId/296/P052011.aspx

The authors should use this to discuss the requirements for ombudsmen.

Our enhanced method separates the duties of pseudonymization (Trusted Third Party) and re-identification (Ombudsman). Therefore, legal obligations apply for the ombudsman, but not for the trusted third party (TTP).

Also, the legal opinion available on the TMF website addresses patient lists (i.e. the generation of a patient identifier), but does not apply to the derivation of a pseudonym from such a patient identifier.

Nevertheless it is necessary to discuss whether and when an ombudsman fulfills the legal requirements investigated in the TMF’s legal opinion. This is covered in the discussion section of our article:

Before her/his assignment, any ombudsman has to undergo scrutiny to ascertain that she/he fulfills legal requirements. Under German standards, for instance, a doctor
would qualify as an ombudsman, if he is treating patients in a translational research program with a corresponding informed consent.

**Minor Essential Revisions**

3. A field report describing practical experiences with the proposed approach would be welcome. It should address usability and acceptance.

A field report addressing usability and acceptance has been added to the manuscript. In the meantime, 28 translational projects with a considerable number of patients use our pseudonymization method. Pseudonymizations take place regularly while re-identifications have been restricted to certain occasions so far. Both quantities appeared sufficient to provide a field report on usability and acceptance:

> Ethically or medically indicated re-identifications have not taken place so far, but successful re-identifications were performed by a principal investigator. This was done due to suspected manual confusion of two samples during entry into the sample submission sheet. In order to track the entire submission process, the investigator, among other measures, re-identified the pseudonyms for the two samples. He was able to reconstruct the pseudonymization process and to verify the right relation between patients and samples. The NCT Trials Office as a service provider would not have noticed the activity, if they had not checked the audit logs of the pseudonymization service (“User X has performed the re-identification of patient with pseudonym Y”). This shows that both pseudonymization and re-identifications processes will function without technical support from the service provider. […]

4. Genetic data lead to a high inherent re-identification risk, see Lin/Owen/Altman, Science 305 (2004). The authors should address this problem.

As this re-identification risk is inherent indeed, genetic data require appropriate authentication and authorization mechanisms for access, as implemented throughout the projects at the Heidelberg Center for Personalized Oncology. Encryption of the data may provide a solution, but raises issues regarding performance and practical usability for large data sets (e.g. whole genome sequencing data).

We address this problem briefly in the limitations section:

[…] In addition, genetic data bears an inherent re-identification risk. This could be addressed with a symmetric cryptographic approach as proposed by Cassa et al. [37]. In general, access to genetic data has to be secured by appropriate authentication and authorization methods to control the inherent re-identification risk.

**Discretionary Revisions**

5. In section "Background/ Identity Management" the reference to [9] is inadequate.

We agree that for a complete identity management the reference is inadequate. For pseudonymization purposes, we consider it essential to create an algorithm that helps automate major portions of the patient identity management. This is why we keep referring to the algorithm for automatic PID generation proposed by Faldum et. al.:
Faldum et. al. [10] introduced an algorithm for automatic PID generation with optimal properties for error detection and correction which semi-automates identity management.
Review Report 2 by Prof. Dr. Mandl

Major Compulsory Revisions

1. Although this appears to be a logical and effective pseudonymization method, the organization of the paper makes it difficult to understand. For example:
   a) • Within the abstract, the development of a new pseudonymization method should be listed either as a method or as a result, but not both. We list the developed method as a result within the abstract of the revised manuscript. This also prompted us to restructure the sections in the article (pseudonymization method presented in the results section)
   b) • In the introductory section entitled “Translational pseudonymization requirements,” it is suggested that the new method was developed to meet the need for an unambiguous and re-identifiable pseudonym, but it is actually the desire for a separation of pseudonymization and re-identification duties that motivates the new method and distinguishes it from prior methods. We agree that the separation of duties is what distinguishes it from prior methods, but the search for an unambiguous and re-identifiable pseudonym was a secondary key motivator for developing our method. We have clarified this in the revised paper: This motivated us to enhance an existing pseudonymization method regarding a separation of duties.
   […]

2. Generally speaking, the two core requirements for comprehensive pseudonymization in translational research are a repetitively unambiguous PSN for a given PID and a separation of duties regarding pseudonymization and re-identification of a patient. If warranted, ombudsmen should be able to re-identify a patient by a given PSN in a streamlined manner without further human interaction.

3. c) • Notwithstanding the methodical literature search, the discussion of prior approaches and a table comparing them is confusing within the results section. The comparison of prior approaches is now part of the methods chapter in accordance with the rewritten abstract.

4. d) • In the section entitled “Translational pseudonymization method,” the authors write, “we chose not to use an encrypted PID directly as a pseudonym,” but after discussing the problem of repeated pseudonymization, they ultimately generate the PSN by encrypting the PID if the PBKDF2 derived key is not already present in the database. The list of PID-PSN relations is discussed prior to the derivation of the PSN. The initial step of the algorithm is described in the final paragraph of the section. The section would be clearer if it presented the steps in the algorithm beginning with the PID in a clinical setting and ending with re-identification by the ombudsman.

In the revised manuscript we present the algorithm beginning with a PID in a clinical setting and ending with a re-identification by an ombudsman. This portion of our manuscript has undergone some extensive editing which would exceed the space constraints of this letter.
e) • The sections describing the new pseudonymization method and its implementation (“Translational pseudonymization method” and “Pseudonymization Service of the NCT Trial Center”) are interrupted by a discussion of privacy threats. The “Translational pseudonymization method” and the “Pseudonymization Service of the NCT Trial Center” are now introduced consecutively. The threat analysis has been moved to the discussion section in the revised manuscript. Again, this portion of our manuscript has undergone some extensive editing which would exceed the space constraints of this letter.

f) • The mention that this method is currently in use for two research projects belongs with the description of the implementation.
We shifted this statement to the implementation section (“Pseudonymization Service of the NCT Trial Center”) as suggested.

g) • The argument that other techniques to achieve pseudonymization are unlikely to exist would be interesting in the discussion section, but it is not a limitation of the method presented.
We moved this argument to the discussion section in the revised manuscript.

The manuscript needs to be heavily edited throughout for spelling, grammar and clarity. There are numerous errors, including:
<<numerous errors>>
The final manuscript was edited and revised by an English language professional.

Minor Essential Revisions
2. On page 8, “One-way pseudonyms are not applicable for translational research.” There are almost certainly translational research projects in which one-way pseudonyms could be used, even if their use would be suboptimal.
We concede that one-way pseudonyms are applicable for research projects with a specific research question. We changed the statement in the revised manuscript: *One-way pseudonyms are applicable for translational research with a specific scientific question. They are not appropriate, however, if personalized treatment decisions based on systematic or incidental findings [5] are considered.*

Discretionary Revisions
3. While much attention is given to the “core” properties of unambiguity and reversibility, little motivation for separating encryption and decryption duties is given. Since it is the separation of duties that defines this method, exploring the need for this in greater depth would help the reader evaluate the significance of this approach. For a small study it seems plausible that the pseudonymization and re-identification duties might be assigned to a single person or group even if they could be separated. Moreover, in the two studies using the authors’ software prototype, there have been no re-identifications necessary and thus no benefit to date from this new technology, while the computational cost has been increased compared to symmetric encryption schemes without a separation of duties.
We agree that certain kinds of studies do not require a separation of duties. Likewise, an organizational separation of duties instead of a technical solution has its merits. In
comparison with symmetric approaches our method admittedly increases the computational cost, making it inefficient for high-volume payload data. For minor data volume, like patient identifiers, however, the increased computational cost of an asymmetric encryption scheme is negligible. This is addressed in the limitations section:

*The asymmetric encryption scheme we use implies a high computational cost. Therefore, symmetric schemes for encryption of the payload data appear more reasonable, especially for large quantities of data (big data).*

4. In the study limitations, it would be appropriate to include the threat analysis currently in the results section, and it would also be helpful to include the computational cost compared to symmetric methods. As addressed in 1e), the threat analysis is now part of the discussion section. Beyond doubt, we agree that asymmetric schemes are less performing than symmetric schemes. This is clarified in the limitations section without a computational cost comparison:

*The asymmetric encryption scheme we use implies a high computational cost.*

It is rather interesting to assess/compare the performance of different symmetric schemes used in the encryption of high-throughput gene expression data. For audio and video files, for instance, this has already been put into practice, see:

*Evaluating the Performance of Symmetric Encryption Algorithms*  

Blind signatures and Paillier cryptosystems (additive homomorphic property) do allow for innovative applications such as anonymous ballot casts. The RSA cryptosystem has a multiplicative homomorphic property which permits a separation of duties, for instance. The future will show whether or not these two, or further properties, can be instrumentalized to solve data privacy problems related to genetic research.