COMMENTARY

Precise Definition of Anonymization in Genetic Polymorphism Studies

Nobuyuki Hamajima, Yoshiko Atsuta, Yoshimitsu Niwa, Kazuko Nishio, Daisuke Tanaka, Kazuhito Yamamoto, Akiko Tamakoshi

Abstract

Anonymization is an essential tool to protect privacy of participants in epidemiologiocal studies. This paper classifies types of anonymization in genetic polymorphism studies, providing precise definitions. They are: 1) unlinkable anonymization at enrollment without a participant list; 2) unlinkable anonymization before genotyping with a participant list; 3) linkable anonymization; 4) unlinkable anonymization for outsiders; and 5) linkable anonymization for outsiders. The classification in view of accessibility to a table including genotype data with directly identifiable data such as names is important; if such tables exist, staff may obtain genotype information about participants. The first three modes are defined here as anonymization unaccessible to genotype data with directly identifiable information for research staff. Anonymization with a key code held by participants is possible with any of the above anonymization modes, by which participants can access to their own genotypes through telephone or internet. A guideline issued on March 29, 2001 with collaboration of three Ministries in Japan defines "anonymization in a linkable fashion" and "anonymization in an unlinkable fastion", "for the purpose of preventing the personal information from being divulged externally in violation of law, the present guidelines or a research protocol", but the contents are not clear in practice. The proposed definitions will be useful when we describe and discuss the preferable mode of anonymization for a given polymorphism study.

Key Words: Anonymization – genetic polymorphism – informed consent

Asian Pacific J Cancer Prev, **5**, 83-88

Introduction

Genetic epidemiology has added a new dimension for disease prevention taking account of individual genetic traits and reports on the association between disease risk and genetic polymorphisms, in particular, have been rapidly increasing (Hamajima et al., 2001; McLeod et al., 2003; Marnellos, 2003). Assessment of polymorphism genotypes has a potential for detecting high risk individuals and modifying unhealthy behaviors such as smoking (McBride et al., 2002). However, polymorphism studies are complicated by ethical problems concerning genotype information of study participants. Since genotypes at large are regarded as very private information which also affects participants' relatives (The Japan Society of Human Genetics, Council Committee of Ethics, 2001), tools to maintain confidentiality are essential.

Anonymization is a useful tool to protect participants' privacy. In Japan, a guideline for research on the human genome drafted with collaboration of three ministries (the Ministry of Education, Culture, Sports, Science and Technology, the Ministry of Health, Labour and Welfare, and the Ministry of Economy, Trade and Industry) was issued on March 29, 2001. The guideline recommends the anonymization of genetic data and genome samples in human genome research (5. Duties of principal investigators (6)). In addition, the statement of "14. Definition of terms, (6) Anonymization" says that anonymization is "for the purpose of preventing the personal information from being divulged externally in violation of law, the present Guidelines or a research protocol", and "(7) Personal information custodian" says the person is in charge of anonymizing private information as well as preventing the unlawful notification to outsiders. "(6) Anonymization" defines "a. anonymization in a linkable fashion" as "anonymization implemented through a method where a corresponding list of an individual and a newly-given symbol or number is maintained so that the person may be identified as necessary" and "b. anonymization in an unlinkable fashion" as "anonymization implemented through a method where no corresponding list

Department of Preventive Medicine / Biostatistics and Medical Decision Making, Nagoya University Graduate School of Medicine, 65 Tsurumai-cho, Showa-ku, Nagoya 466-8550 Japan, TEL:+81-52-744-2132, FAX:+81-52-744-2971 nhamajim@med.nagoya-u.ac.jp

aforementioned in a. is maintained, so that an individual could not be identified". Both anonymization methods are understandable conceptually, but in practice several different modes of anonymization exist in terms of anonymization timing and masking for staff.

Although focus of the guideline is placed rather on hereditary disease genes, the guideline is applicable for genetic polymorphisms. This paper aims to define anonymization modes in details, focusing on the polymorphism genotype data, for which counseling is not required. Although many modified modes are considered in practice, the prototypes are presented here. In order to describe the modes, definitions of "table" and "staff" are needed. These definitions provide the terms for discussion on the level of anonymization. The present paper does not aim to recommend an anonymization mode for a given polymorphism research, which totally depends on the situation where polymorphism studies are conducted.

Definition of Tables and Participating Staff

Tables

Genotype data tables including genotypes data of each participant are classified into three types; a) "genotype table with directly identifiable information", b) "genotype table with a linkage code", and c) "genotype table with no linkage code", as shown in Table 1. The directly identifiable information means the information such as name, address, and birthday, enables to identify each individual without additional information. How to manage the "genotype table with directly identifiable information" is the target for confidentiality management. Any genotype table may include demographic data and lifestyle information (Fig 1). "Genotype table without any linkage codes and directly identifiable information" with data for analysis is the final product after unlinkable anonymization

"Linkage table" is a table with a linkage code linking information to genotype data table. The information is on participants ("linkage table including directly identifiable information"), other data such as lifestyle or clinical data (" linkage table including data other than directly identifiable information"), or other linkage codes ("linkage table including other codes"). In terms of anonymization, the first linkage table is of primary interest.

The description on the place of "linkage table" stored is important to describe the level of confidentiality. The table is classified in the relation to genotype table (Table 1). " Linkage table in the same computer system as that for genotype data table," "linkage table in a different computer system from that for genotype data table in the same facility," and "linkage table in a computer system of an independent facility not directly connected with that for genotype data table" make a difference in the level of confidentiality.

Staff

Several kinds of staff are actively engaged in genetic polymorphism research. "Enrollment staff" are personnel to register participants with or without identifying participants. Through the registration they can make a linkage table. They are at a position to know their private information on lifestyle or disease history, but do not know participants' genotypes. "Anonymization staff" is the person to anonymize sample and data in a linkable or unlinkable way, who have no chance to see genotypes. "Genotyping staff" is in charge of genotyping, for whom directly identifiable information is not available. "Record linkage staff" links the genotype data with other data, who could work as a "statistical analysis staff".

As mentioned in Introduction, the guideline for research on the human genome issued on March 29, 2001, requires the appointment of a "personal information custodian" in charge of anonymization and private information security. The person is not defined here, but he/she can be anonymization staff.

Table 1. Terms Describing the Level of Anonymization

I. Genotype data table
- a) Genotype table with directly identifiable information (e.g., genotypes and name)
- b) Genotype table with a linkage code (e.g., genotypes and registry number)
- c) Genotype table without any linkage codes and directly identifiable information (e.g., genotypes with lifestyle data; a final product through anonymization)

II. Linkage table with a linkage code linking information to genotype data table
- a) Linkage table including directly identifiable information
- b) Linkage table including data other than directly identifiable information
- c) Linkage table including only other codes

III. Place of linkage table stored, in relation to genotype data table
- a) Linkage table in the same computer system
- b) Linkage table in a different computer system
- c) Linkage table in an independent facility

IV. Staff
- a) Enrollment staff
- b) Anonymization staff
- c) Genotyping staff
- d) Record linkage staff
- e) Statistical analysis staff

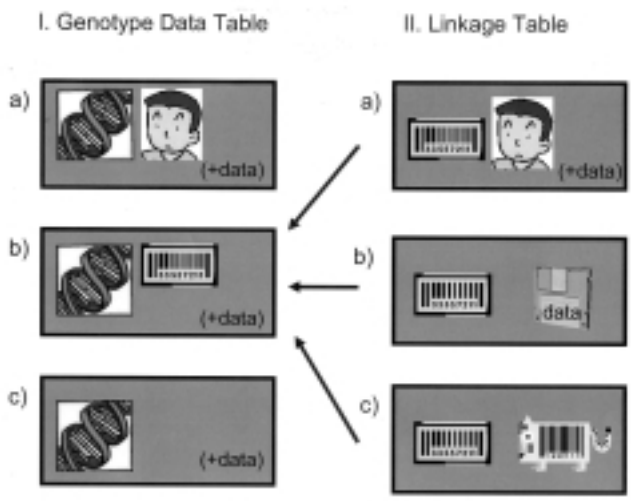

Figure 1. Genotype Data Table and Linkage Table

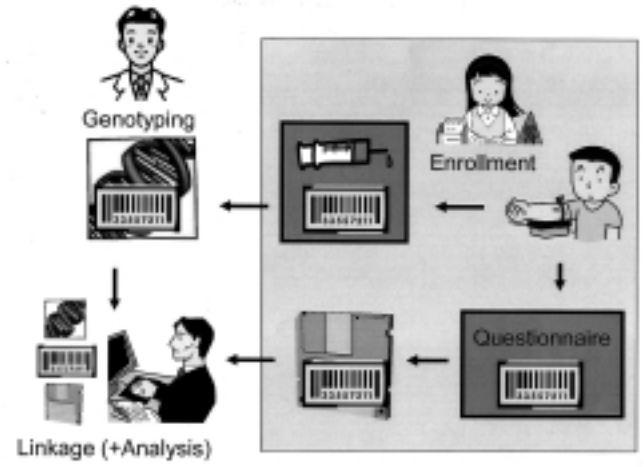

Figure 2. Unlinkable Anonymization at Enrollment with no Participant List.

Anonymization

Although the guideline issued on March 29, 2001 focuses the anonymization for outsiders, the present paper classifies it in view of the timing of anonymization in connection with genotype information restriction against staff. In addition, a mode to allow participants to access their own genotypes anonymously is described.

1. Unlinkable anonymization at enrollment with no participant list (Fig 2)

Participants can provide blood and questionnaire without identifying themselves at enrollment in this design. Since no table is made to link a linkage code to each participant, anonymization is complete, but duplicated participation cannot be detected. In this anonymization method, informed consent forms are not necessary. If we ask participants their signature on an informed consent form, anonymous participation could become meaningless in one sense. Of important is the situation assuring that blood donation and questionnaire administration are regarded as their consent. Acquaintances of enrollment staff may participate in study, but no records are maintained. A complicated linkage code is recommended to avoid easy memorization by the enrollment staff. The linkage code is used only for connecting blood sample with questionnaire data, but not with participants. In Fig 2, the same code number is used for blood sample and questionnaire, but it is not necessary when a table is made to link blood sample and questionnaire. The linkage can be conducted either in or out of the facility. Statistical analysis can be performed at any place, because the data is completely anonymous from the enrollment.

2. Unlinkable anonymization before genotyping with a participant list (Fig 3)

As shown in Fig 3, this system uses a list of participants with their blood sample and questionnaire. Anonymization could be done either routinely or at one time by anonymization staff before genotyping, using a linkage code. Anonymization at one time would be a convenient method in case that the participation is less frequent and a routine enrollment system is not established. Since it is unlinkable anonymization, the table linking linkage codes with participants should be deleted before genotyping if the linkage table is made. However, the list of participants can be maintained by the research group. It could work for checking the duplicated participation. Informed consent forms are preferably stored before anonymization, and so after anonymization if the participant list is maintained.

If genotyping is conducted after the collection of all follow-up data, this mode can be used for such follow-up studies.

3. Linkable anonymization (Fig 4)

In a usual follow-up study, follow-up data may be added to the dataset after genotyping. A list of participants and

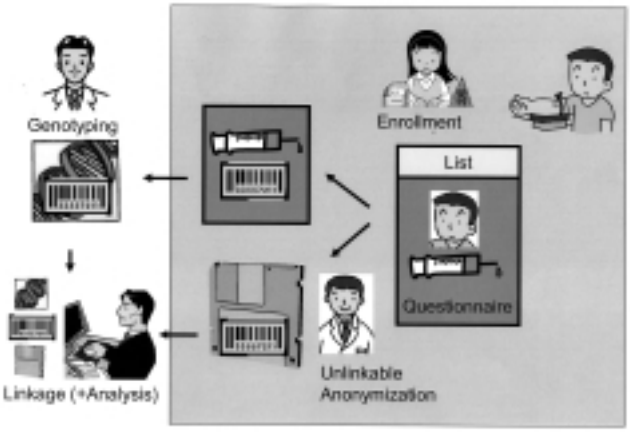

Figure 3. Unlinkable Anonymization before Genotyping with a Participant List.

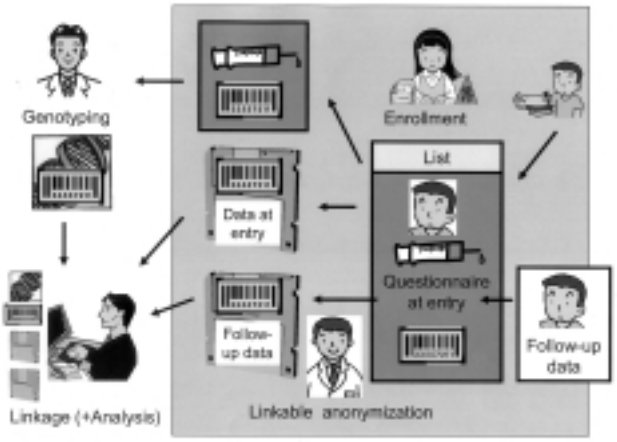

Figure 4. Linkable Anonymization before Genotyping.

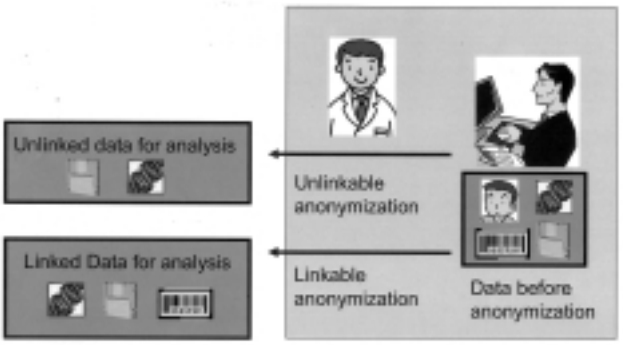

Figure 5. Unlinkable and Linkable Anonymization for Outsiders.

linkage codes are referred to add follow-up data. Anonymization staff puts the linkage code on the follow-up data, and send it to linkage staff. This process prevents anonymization staff to know genotype data and linkage staff to know the participants. If anonymization staff and linkage staff are the same person, the person is at a position to know participants and their genotypes at the same time. If so, the mode is rather close to the below mode.

4. Unlinkable and linkable anonymization for outsiders (Fig 5)

In this mode, a table including genotypes and directly identifiable information, as well as linkage code and other data, is made in the research group. The dataset is anonymized when it is provided for outsiders. The risk exists that the dataset before anonymization is unlawfully moved out. In addition, the staff may become known the genotypes of acquaintance through genotype table with directly identifiable information by chance. It hurts credibility of the research team.

For outsiders to analyze the data, neither directly identifiable information nor linkage codes are necessary. Accordingly, unlinkable anonymization makes no problems, except in the case that the record errors are found and correction using a linkage code is required. In practice, it is quite often that the errors are found at a statistical analysis stage, and several times of data cleaning processes between analysis staff and data cleaning staff are not rare. In this case, linkable anonymization is absolutely convenient.

5. Anonymization with a key code held by participants

This mode dose not defined in terms of the timing of anonymization, but accessibility for participants. When a key code connecting data is provided for participants, they can access to their genotype, independent on the modes defined above. Imagine that anonymized genotypes are stored with a key code. The participants holding the key code are accessible through telephone questioning or internet. Those interested in their genotypes could know their genotypes under a situation anonymous for all research staffs. Asking the genotypes is totally dependent on participants' autonomy. Fig 6 demonstrates a case for unlinkable anonymization at enrollment with a key code held by participants.

Independence among Staffs

The above anonymization modes require independent roles among the staffs. All the modes except "anonymization for outsiders" pay attention to the point that even research staff should not know participant's genotypes in the process to make an anonymized dataset. Accordingly, "enrollment", "anonymization", "genotyping" and "linkage" should be conducted by different persons. "Linkage staff" and "statistical analysis staff" can be the same person, because both are apart from directly identifiable information.

Checking List on Anonymization

Table 2 shows a list to delineate the level of anonymization. Through filling out the questions, we can characterize the mode actually used for anonymization. We

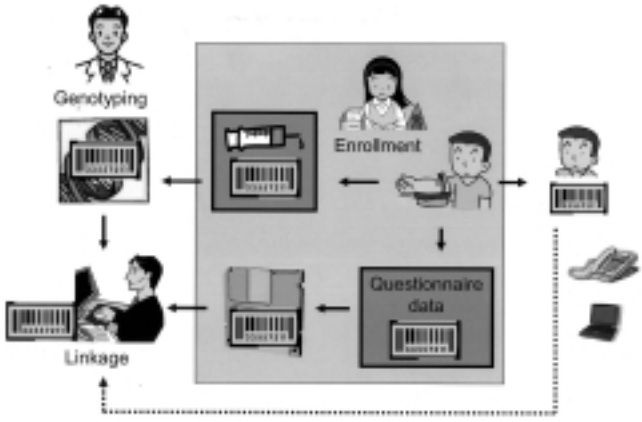

Figure 6. Unlinkable Anonymization at Enrollment with a Key Code Held by Participants.

have already learned that the simple terms, "unlikable" or " linkable" cannot describe the real situation of anonymization in a given polymorphism study.

Comments

Until a few years ago, majority of medical professions and investigators were not aware of the importance of the confidentiality of private information. Complicated anonymous processes were simply regarded as useless one or an obstacle to cause mistakes. When genotype researches became popular, a dramatic change occurred in Japan as symbolized by the guideline issued on March 29, 2001, which made researchers to realize anonymization necessary. Now, in Japan, many studies including large-scale follow-up studies are planed or on going under the guideline.

Anonymization means the process to make an anonymous condition not only for outsiders and participants, but also for research staffs. The process to make samples and data anonymous for staffs requires research systems described here. The establishment of such systems burden researchers, but now, it has to be cleared.

Under some special circumstances, participants may wish to know test results. For example, an anonymous HIV test for those anxious about their infectious status is conducted in a way anonymous for health professionals, not for participants. Similarly, there are occasions participants wish to know their genotypes anonymously for research staffs. Anonymization with a key code held by participants is used for such occasions.

There may be many modifications relating to anonymization process; how to issue labels of linkage codes, when and who inputs data, when and how data send, and so forth. A system using different codes for samples, data, and informed consent document was reported (Hara et al., 2003). The system prevents staff to memorize the codes for a certain participant, resulting in the attainment of complete anonymization. Since this process is relating to labels used for linking codes, it does not contradict the classification proposed in this paper. An anonymization system with third-party encryption was proposed in Iceland (Gulcher et al., 2000). In Japan, a millennium project of Ministry of Education, Culture, Sports, Science and Technology, established a center to distribute linkage code labels to participating hospitals.

Anonymization is an important factor of the contents of informed consent. However, the characterization is not precisely described even among researchers in the field of polymorphism studies. This paper defined the modes of anonymization, which can be used for informed consent process in polymorphism studies. In addition, the above categorization provides a common framework for discussion what kind of anonymization is suitable for a given study.

Acknowledgements

This work was supported in part by a Grant-in-Aid for Scientific Research on Special Priority Areas of Cancer from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

Table 2. Checklist for Anonymization

1. Which design does the study use ?
   1) Anonymous cross-sectional study
   2) Cross-sectional study or case-control study
   3) Follow-up study
2. Which genotype data tables does the study use ?
   1) With directly identifiable information
   2) With a linkable code
   3) Without any linkable codes and directly identifiable information
3. Where is the linkage table including directly identifiable information stored ?
   1) In the same computer system as that for genotype data tables
   2) In a different computer system from that for genotype data tables in the same facility
   3) In an independent facility
   4. Are enrollment staff, anonymization staff, genotyping staff, and record linkage staff, different persons ?
      1) Yes 2) No
      If yes, which roles do they play ?
      _____ and _____ _____ and _____
      _____ and _____ _____ and _____
5. Which anonymization does the study use ?
   1) Unlinkable anonymization at enrollment with no participant list
   2) Unlinkable anonymization before genotyping with a participant list
   3) Linkable anonymization before genotyping
   4) Unlinkable anonymization for outsiders
   5) Linkable anonymization for outsiders
6. Does the study use a key code held by participants ?
   1) Yes 2) No

References

Gulcher JR, Kristjansson K, Gudbjartsson H, Stefansson K (2000). Protection of privacy by third-party encryption in genetic research in Iceland. *Eur J Hum Genet*, **8**, 739-42.

Hamajima N, Matsuo K, Saito T, et al (2001). Gene-environment interactions and polymorphism studies for cancer risks in the Hospital-based Epidemiologic Research Program at Aichi Cancer Center II (HERPACC-II). *Asian Pacific J Cancer Prev*, **2**, 99-107.

Hara K, Ohe K, Kadowaki T, et al (2003). Establishment of a method of anonymization of DNA samples in genetic research. *J Hum Genet*, **48**, 327-30.

Marnellos G (2003). High-throughput SNP analysis for genetic association studies. *Curr Opin Drug Discov Devel*, **6**, 317-21.

McBride CM, Bepler G, Lipkus IM, et al (2002). Incorporating genetic susceptibility feedback into a smoking cessation program for African-American smokers with low income. *Cancer Epidemiol Biomarkers Prev*, **11**, 521-8.

McLeod HL, Yu J (2003). Cancer pharmacogenomics: SNPs, chips, and the individual patient. *Cancer Invest*, **21**, 630-40.

The Japan Society of Human Genetics, Council Committee of Ethics (2001). Guidelines of genetic tests. *J Hum Genet*, **46**, 163-5.