J Hum Genet (2003) 48:327–330
DOI 10.1007/s10038-003-0022-6

SHORT COMMUNICATION

**Kazuo Hara · Kazuhiko Ohe · Takashi Kadowaki
Naoya Kato · Yasushi Imai · Katsushi Tokunaga
Ryozo Nagai · Masao Omata**

Establishment of a method of anonymization of DNA samples in genetic research

Received: 22 January 2003 / Accepted: 4 March 2003 / Published online: 15 May 2003

Abstract As the number of the genetic studies has rapidly increased in recent years, there has been growing concern that the privacy of the participants in such studies can be invaded unless effective measures are adopted to protect confidentiality. It is crucial for the scientific community to establish a method to anonymize DNA samples so that the public will trust genetic researchers. Here, we present a reliable and practical method of making DNA samples used in the genetic research anonymous. It assures complete anonymity by coding samples and personal information twice. Since it does not require equipment, such as bar-code readers or a software package, its cost is nominal compared with the laboratory costs. All institutions engaged in genetic research may wish to take measures such as the one described here to ensure the privacy and confidentiality of the participants in their genetic studies.

K. Hara and K. Ohe contributed equally to this work

K. Hara · Y. Imai · R. Nagai
Department of Clinical Bioinformatics,
Graduate School of Medicine, University of Tokyo,
Tokyo, Japan

K. Hara · T. Kadowaki
Department of Metabolic Disease,
Graduate School of Medicine, University of Tokyo,
Tokyo, Japan

K. Ohe (✉)
Department of Planning, Information and Management,
University of Tokyo Hospital, 7-3-1 Hongo,
Bunkyo-ku, Tokyo 113-8655, Japan
E-mail: kohe-tky@umin.ac.jp
Tel.: +81-3-38155411

N. Kato · M. Omata
Department of Gastroenterology, Graduate School of Medicine,
University of Tokyo, Tokyo, Japan

K. Tokunaga
Department of Human Genetics, Graduate School of Medicine,
University of Tokyo, Tokyo, Japan

Introduction

Recent developments in analytical technology have enabled even small laboratories to engage in genetic research, and as a result the number of genetic studies has been increasing year by year (Kawamoto et al. 2001; Nishio et al. 2001; Watanabe et al. 2002; Yamada et al. 2002). At the same time, society has been paying more and more attention to protecting the privacy of individuals who participate in genetic investigations out of fears that governments, insurance companies, or employers might use or abuse genetic information obtained during the course of such studies (Pokorski 1998; The Japanese Society of Human Genetics, Council Committee of Ethics 2001; Wertz 2002). Another important issue to be aware of is that genetic information gathered from participants in such studies may provide genetic information about their relatives and even offspring. Since such concerns may hamper the progress of genetic research by discouraging individuals from participating in the investigations, establishing a framework for protecting the privacy of participants in the genetic study, namely anonymization, is crucial (Fuller et al. 1999). To address ethical issues in this field, a guideline for research on the human genome (''the guideline'') drafted by collaboration among the Japanese Ministry of Education, Culture, Sports, Science and Technology, Ministry of Health, Labor and Welfare, and Ministry of Economy, Trade and Industry was released, on March 29, 2001 (http://www2.ncc.go.jp/elsi/index.htm, in Japanese). All investigators engaged in genetic research in Japan are now subject to the guideline. It recommends that researchers anonymize DNA specimens obtained from the participants and remove any individual identifying information from their clinical records before starting the research. Although the necessity of anonymity is widely accepted, the methodology for accomplishing it has not been clearly described. We describe a reliable and practical method for making specimens and clinical data retrieved from

participants in genetic studies in our institution anonymous.

Materials and methods

Coding specimens, clinical data documents, and informed consent documents

After obtaining permission from the IRB (Institutional Review Board) to proceed with a genetic study, researchers request the administrator of personal information (“the administrator”), who is appointed by the Dean of the Graduate School of Medicine of the University of Tokyo, to carry out the series of procedures described below. According to the guideline, the administrator is not allowed to be involved in any genetic research performed in our institution. First, the administrator creates sets of code numbers comprising an “S-number” to code the sample, a “C-number” to code the informed consent document, and a “W-number” to the code clinical data document (work sheet) on which the researcher fills personal identifying information (such as name and date of birth) and non-identifying information (such as body weight, body height, blood levels of some hormones) (Fig. 1a). The number is printed on an adhesive label (S-, W-, and C-label) covered with a seal that prevents the printed number from being read unless the seal is torn off (Fig. 1b). The seal is manufactured in such a manner that it cannot be used to cover the label again once it has been detached from it. In addition, the S-, W-, and C-labels are different so that the specimen cannot be easily linked to the clinical data or the name signed on the informed consent document. This guarantees the confidentiality of the temporary code numbers. Since these numbers are replaced by permanent numbers in a later procedure, we call them “temporary code numbers”. The administrator issues an individual set of the temporary code labels with a protective seal to the researcher (procedure 1, Fig. 2), and the researcher attaches “S-labels” to specimens such as blood samples, “W-labels” to clinical data forms, and “C-labels” to informed consent forms (procedure 2). The researcher then submits sets of samples, clinical data documents and informed consent documents to the administrator when a large number of sets (typically 20 sets) has accumulated (procedure 3). Through all these procedures, the researchers cannot know the temporary code numbers.

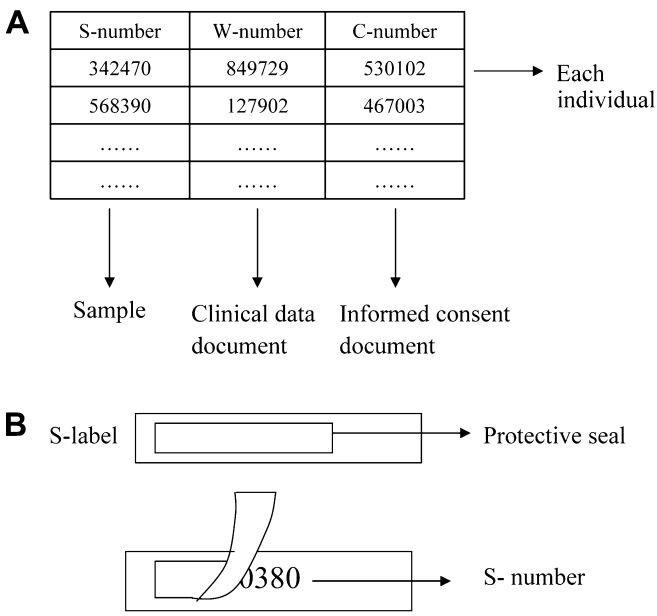

Fig. 1 The temporary code numbers and labels to protect confidentiality in genetic research

Replacement of temporary code numbers by anonymous numbers

The administrator takes off the protect seal and reads the temporary code numbers that accompany the clinical data documents and informed consent documents and creates a data file in which the personal identifying and non-identifying information are linked to the temporary code numbers (procedure 4). The administrator then assigns a permanent number (“anonymous number”) to the temporary code numbers (procedure 5). The anonymous number is a unique six-digit number created by using a random number generator. The generated database is encrypted and stored in a standalone computer that only the administrator is allowed to access (procedure 6a). The administrator then creates another data file (anonymous data file) that links the anonymous number only to the clinical parameters (non-identifying information) by removing all the personal identifying information from the linking data file (procedure 6b). The S-labels accompanying the samples are then replaced by their related anonymous number under the supervision of the administrator in a room exclusively used for this procedure (procedure 7). The S-label on which the temporary code number was printed is discarded. The administrator delivers the anonymous data file and specimens coded with the anonymous number to the researcher (procedure 8).

Genetic analysis using anonymously coded specimens and clinical data

DNA is extracted from anonymous specimens in the laboratory of the researcher. The researcher can investigate the effect of genetic variations on clinical parameters by using the anonymous data file, but cannot trace the specimen back to the subject from whom it was obtained. If the researcher needs to identify the specimen for some reason, the researcher can request the IRB to permit linking of the specimen to personal identifying information.

Results and discussion

After discussing how to anonymize samples and clinical data obtained from participants in the genetic research for a whole year, we concluded that we should adopt the procedure described in this paper. We believe that the researchers in our institution can fulfill their duty to protect the privacy of participants by using our procedure. It is characterized by complete anonymity being achieved in two steps (procedure 2 and 7). The specimen and clinical data are first anonymized with a temporary code number issued by the administrator (procedure 2), but since this procedure alone fails to guarantee full protection of privacy, because it is possible for a malicious researcher to link temporary code numbers to personal identifying information, the temporary code number must be replaced by the second, permanent number (procedure 7). Once this procedure is carried out under the supervision of the administrator, no one but the administrator can link the specimen to the identifying information.

The Japanese guideline for the genetic research recommends that researchers anonymize DNA specimens obtained from the participants and remove individual identifying information from their clinical records before starting their research. According to the guideline, anonymity status is classified into two categories, “linked”

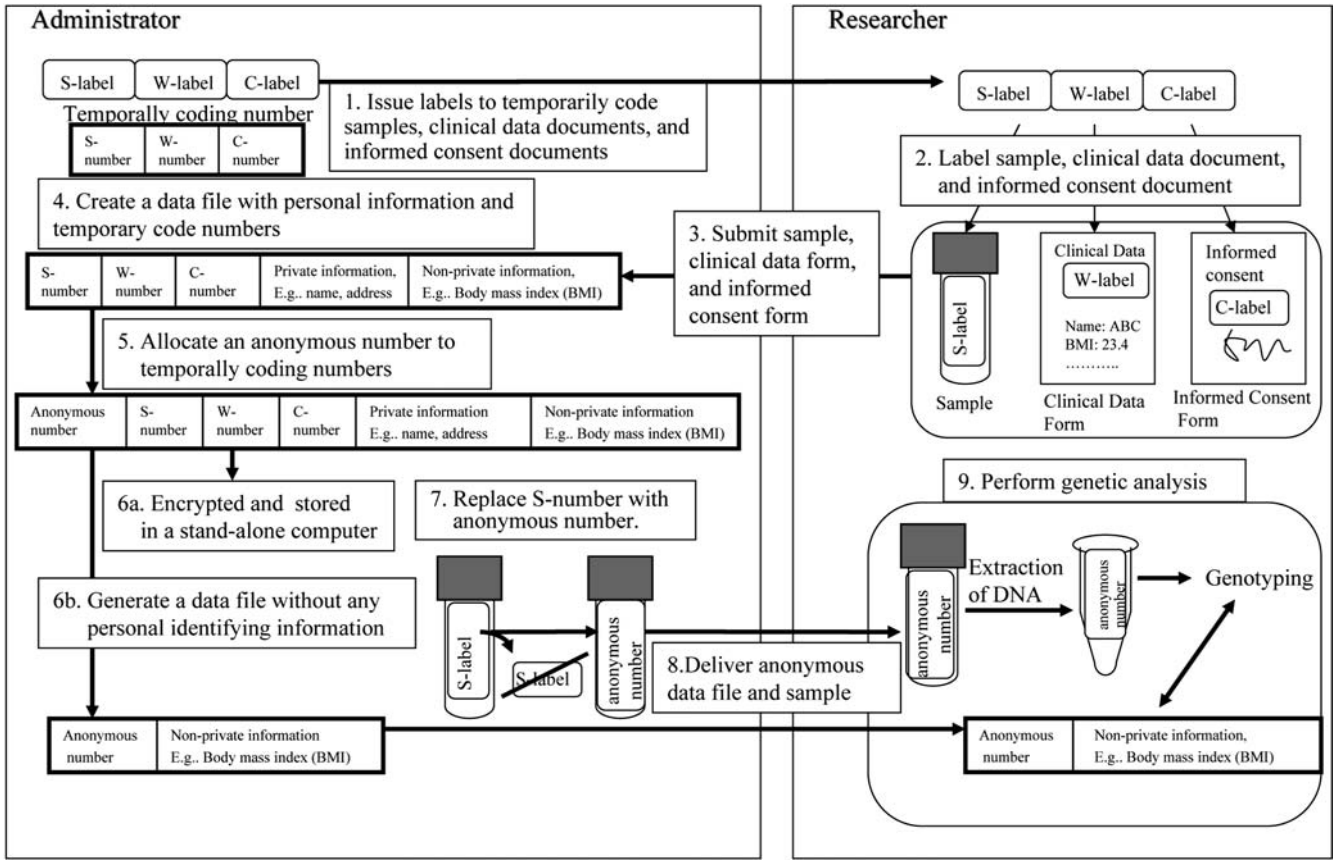

Fig. 2 Framework for protecting the privacy and confidentiality of participants in genetic research

and ''unlinked''. The former means coding a specimen with a random number with a key that links that number to personal identifying information preserved by the administrator. The latter means removing all the personal identifying information from a specimen without retaining any key. Consequently, there is no way the specimen can be traced back to the person from whom it was obtained. Control samples obtained from healthy subjects should be treated with unlinked anonymity, whereas samples from patients can be treated with linked anonymity. Unlinked anonymity is likely to be more secure than linked anonymity, but potential benefits may be lost. If unlinked, measurements of the levels of a protein encoded by a newly discovered susceptibility gene or collection of related medical information from participants is impossible, and that may reduce the long-term value of the research. Another possibility is that a serious genetic disease may be unexpectedly found in subjects enrolled even in genetic research on multifactorial disorders. If a treatment or prophylaxis of the genetic disease has been established, carriers among family members of participants would obtain enormous benefit from such genetic information. Therefore, protection of privacy and confidentiality must be weighed against the need to perform genetic research to develop novel and etiology-based methods of treatment of diseases in which genetic factors are involved.

There is a report of DNA sample anonymization by using an encryption software package manufactured by a third party (Gulcher et al. 2000), but that allows malicious personnel in the clinic and the laboratory to cooperate in generating a key table linking temporary code numbers to anonymous numbers, which would destroy the anonymization system. Our system prevents researchers from generating such keys and makes the system more secure. Moreover, the cost of anonymizing a specimen (approximately 40 cents per specimen) by our procedure is nominal compared with the laboratory costs and is smaller than the cost of using encryption software manufactured by a third party. In any event, it is no exaggeration to say that the security of privacy depends not just on the anonymizing procedure but also on the ethics of researchers and how carefully they handle specimens and genetic data.

Acknowledgements We thank Toppan Label Co. for manufacturing all the labels used in the anonymizing procedure describe in the present manuscript.

References

Fuller BP, Kahn MJ, Barr PA, Biesecker L, Crowley E, Garber J, Mansoura MK, Murphy P, Murray J, Phillips J, Rothenberg K, Rothstein M, Stopfer J, Swergold G, Weber B, Collins FK, Hudson KL (1999) Privacy in genetic research. Science 285:1359–1361

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 8:739–742

Kawamoto R, Kohara K, Tabara Y, Miki T, Doi T, Tokunaga H, Konishi I (2001) An association of 5,10-methylenetertrahydrofolate reductase (MTHFR) gene polymorphism and common carotid atherosclerosis. J Hum Genet 46:506–510

Nishio Y, Noguchi E, Ito S, Ichikawa E, Umebayashi Y, Otsuka F, Arinami T (2001) Mutation and association analysis of the interferon regulatory factor 2 gene (IRF2) with atopic dermatitis. J Hum Genet 46:664–667

Pokorski RJ (1998) A test for the insurance industry. Nature 391:835–836

The Japanese Society of Human Genetics, Council Committee of Ethics (2001) Guidelines for genetic testing. J Hum Genet 46:163–165

Watanabe H, Hamada H, Yamada N, Sohda S, Yamakawa-Kobayashi K, Yoshikawa H, Arimnami T (2002) Association analysis of nine missense polymorphisms in the coagulation factor V gene with severe preeclampsia in pregnant Japanese women. J Hum Genet 47:131–135

Wertz DC (2002) Genetic discrimination – an overblown fear? Nat Rev Genet 3:496

Yamada Y, Fujisawa M, Ando F, Niino N, Tanaka M, Shimokata H (2002) Association of a polymorphism of the transforming growth factor-beta1 gene. J Hum Genet 47:243–248