OPEN ACCESS Freely available online

PLoS MEDICINE

Guidelines and Guidance

Ethical and Practical Issues Associated with Aggregating Databases

David R. Karp*, Shelley Carlin, Robert Cook-Deegan, Daniel E. Ford, Gail Geller, David N. Glass, Hank Greely, Joel Guthridge, Jeffrey Kahn, Richard Kaslow, Cheryl Kraft, Kathleen MacQueen, Bradley Malin, Richard H. Scheuerman, Jeremy Sugarman

The goal of "personalized medicine" relies upon defining the genetic variation responsible for disease susceptibility and response to therapy [1]. For most common human diseases, the contribution of a single sequence variant to disease susceptibility is typically small, and can only be detected with data from large numbers of people [2]. Practically, this necessitates collaboration among investigators who either have DNA and phenotypic information previously collected, or have access to populations from which to recruit participants. It also requires that data be shared among the collaborators. Modern bioinformatics platforms have the capacity to combine datasets and store them for re-analysis. This is scientifically advantageous since it makes possible studies with enhanced validity in a cost-effective fashion. However, this data storage can complicate the already vexing practical, scientific, and ethical issues associated with gene and tissue banks. Research participants' data may have been collected without authorization that meets today's standards for informed consent. Research participants may not have consented to participation in genetics research in general, to inclusion in genetics databases specifically, or to use of their samples in genetic analyses that were unanticipated, unknown, or nonexistent at the time samples were collected [3]. Participants who consented to the collection of their data for use in a particular study, or inclusion in a particular database, may not have consented to "secondary uses" of those data for unrelated research, or use by other investigators or third parties [4]. There is concern that institutional review boards (IRBs) or similar bodies will not approve of the formation of aggregated databases or will limit the types of studies that can be done with them, even if those studies are believed by others to be appropriate, since there is a lack of consensus about how to deal with re-use of data in this manner.

Combined databases can raise other important ethical concerns that are unrelated to the original consent process. For example, they may make it possible for investigators to identify individuals, families, and groups. Such concerns may be exacerbated in settings where there is the possibility of access to data by individuals who are not part of the original research team. For example, the National Institutes of Health (NIH) Data Sharing Policy (http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html) requires investigators receiving more than US$500,000 per year to share the final research data. For genome-wide association studies, individual records are to be deposited into dbGaP (Database of Genotypes and Phenotypes) at the National Center for Biotechnology Information. While only aggregate data will be accessible to the public, record-level data will be accessible to other investigators who agree to certain terms regarding confidentiality and data security (http://grants.nih.gov/grants/guide/notice-files/NOT-OD-07-088.html). Further, databases developed for different kinds of studies, with different methodologies and conventions for analyzing and reporting results, that are then aggregated, may not be interoperable, leading to flawed analyses. Finally,

Box 1. Recommendations

1. Determine whether initial consent and ethical approval will allow secondary research.
2. Ensure that there are appropriate data security mechanisms and review bodies to protect privacy interests in aggregated databases.
3. Informed consent should take into account the potential incorporation of data into aggregated databases.
4. Address special challenges of using data obtained from existing databases.
5. Pursue efforts directed at standardization of data.
6. Establish data sharing rules, including attribution of contributions.
7. Adopt "best practices" to avoid identifiability of the data.

The Guidelines and Guidance section contains advice on conducting and reporting medical research.

Funding: This work was supported by a contract from the National Institute of Allergy and Infectious Diseases of the United States National Institutes of Health (N01-AI40076). No separate funding from the National Institutes of Health was received for this article.

Competing Interests: The authors have declared that no competing interests exist.

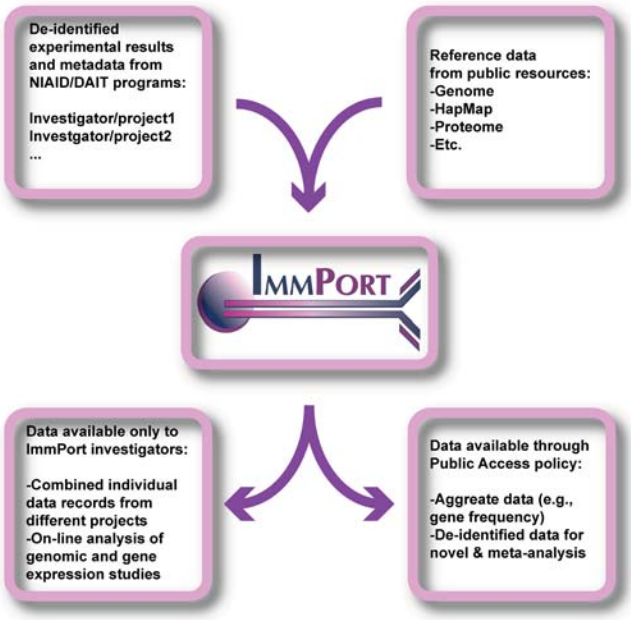
Citation: Karp DR, Carlin S, Cook-Deegan R, Ford DE, Geller G, et al. (2008) Ethical and practical issues associated with aggregating databases. PLoS Med 5(9): e190. doi:10.1371/journal.pmed.0050190

Abbreviations: DAIT, Division of Allergy, Immunology, and Transplantation; ImmPort, Immunology Database and Analysis Portal; IRB, institutional review board; NIAID, National Institute of Allergy and Infectious Diseases; NIH, National Institutes of Health

See section at end of manuscript for author affiliations

* To whom correspondence should be addressed. E-mail: david.karp@utsouthwestern.edu

Provenance: Not commissioned; externally peer reviewed



doi:10.1371/journal.pmed.0050190.g001

Figure 1. The ImmPort System

Investigators participating in major research programs supported by NIAID/DAIT use ImmPort to satisfy NIH data sharing policies. Experimental results, genotypes, and participant phenotypes are stored on secure servers accessible via the Internet. Data can be analyzed by the original investigator within the ImmPort system and can be combined with data from other DAIT-supported investigators for collaborative projects. Reference data from several public sources are integrated within the system and can be a source of comparison with experimental data. The reference data are available to the public, while experimental data are initially visible only to investigators. Ultimately, aggregated experimental data and anonymous research records can be made available to non-DAIT investigators through the mechanisms discussed in the text.

data and specimen repositories that investigators have carefully collected over decades can quickly become "public" information, thus compromising publication priority and intellectual property claims.

We convened a panel of bioethicists, scientists, and legal experts to analyze these issues and to develop guidelines for aggregating databases. Our analysis focused on ImmPort (Immunology Database and Analysis Portal; http://www.ImmPort.org/; see Figure 1), a Web-based resource being developed for the National Institute of Allergy and Infectious Diseases (NIAID). The purpose of ImmPort is to provide "…advanced information technology support in the production, analysis, archiving, and exchange of scientific data for the diverse community of life science researchers supported by NIAID/DAIT [Division of Allergy, Immunology, and Transplantation]." ImmPort provides a platform for investigators to store and share data that have been collected in NIAID-sponsored studies, including genotypes, experimental results (e.g., response to vaccination or stem cell transplant), and clinical phenotypes (e.g., healthy/affected), but excluding individual participant identifiers. The charge of the panel was to consider the ethical concerns that arise when data are shared in aggregated databases such as ImmPort. Observations about ImmPort should be relevant to other efforts directed at aggregating databases.

Methods

Two of the authors (DRK and JS) organized the two-day meeting. Potential participants who were known to the organizers either as stakeholders in ImmPort or for their work in the fields of research methodology and regulation were invited to participate. These included NIH program staff working with ImmPort, bioethicists, NIH-supported researchers who would be required to submit data to ImmPort, and content experts. We sought people with expertise in the areas of informed consent, protection of vulnerable populations, privacy protections, research regulation, and intellectual property. Lay members of the University of Texas Southwestern Medical Center IRB and research participants were also included. A total of 32 people were invited, and 20 participated. Six discussants were asked to present the issues and controversies relating to three general themes: Assuring Data Privacy and Security, Informed Consent, and Challenges of Collaboration. Audio recordings of the presentations and discussion, including controversial issues and dissenting opinions, were prepared for reference. After the presentations and discussion by all participants, the major concepts were summarized by the organizers along with a listing of problem areas and possible solutions, which were discussed in detail by the entire group. Consensus on the recommendations was achieved and refined through a collaborative writing process by all authors. The strengths of this meeting include the focus on a particular research topic, the ImmPort project, with potential wide applicability to similar initiatives, by a small group of persons with relevant expertise. The weakness of this method is also related to these aspects, as the selected participants may not have addressed all possible pitfalls in database aggregation and sharing.

Recommendations

The panel developed seven broad recommendations, which are summarized in Box 1.

1. Determine whether initial consent and ethical approval will allow secondary research. In general, this necessitates ensuring that data were collected with prospective informed consent where it was practicable to do so. If the investigator who submits data to a database did not collect the original biological specimen or clinical phenotype, then its history should be detailed, tracing the custody of the specimen and/or information to the original investigator and consented participant. The initial IRB or similar oversight body should typically be able to attest that there was approval of the protocol used to collect the information and confirm whether submission to the aggregated database is within the scope of the original consent. Did the consent address secondary uses of the data or the possibility that the participant could be re-identified? Where the submitting investigator has received "de-identified" specimens and data from another source, the responsible IRB should have assessed the propriety of acquisition and study of the samples at the time of proposed transfer to the submitting investigator. IRBs and those creating aggregated databases must recognize that some research protocols and informed consent processes specify limited use of participants' data. For example, such limitations may specify that data may be shared with other investigators doing research only on the same or related conditions.

Finally, as technology makes whole-genome analyses faster and cheaper, the need for sharing samples will be replaced by the sharing of data. One investigator's "affected" could be another investigator's "control." However, IRBs historically have been concerned with informed consent for the collection and sharing of specimens (blood, DNA, etc.). Both IRBs and investigators should anticipate sharing of data derived from specimens and be able to document participants' consent for this purpose.

2. Ensure that there are appropriate data security mechanisms and review bodies to protect privacy interests in aggregated databases. By design, the research facilitated by combination and re-analysis of datasets within de-identified databases should be exempt from complete prospective IRB review. By only accepting data that cannot be directly linked to living persons, studies in these databases do not constitute human participants research under US regulations (http://www.hhs.gov/ohrp/humansubjects/guidance/cdebiol.pdf). This may lead to a false sense of security, given that complex phenotypes and extensive genotypes may be used to re-identify individuals. Re-identification should be prevented using technology that does not significantly degrade the utility of the data (see Recommendation 7) and policies that address the appropriate uses of those data. Only data in de-identified, aggregated form should be accessible to all users. Access to record-level data must be limited, either by keeping all data within the confines of the secure database and providing users with sophisticated analytical capacity, or by requiring legal agreements to prevent attempted re-identification of participants. Downloading and re-distribution of data should be carefully controlled. If allowed at all, extraction of record-level data should only be permitted for specified, pre-approved purposes. Database stewards should play an active role in the periodic review of data use to confirm that data integrity is not compromised and that any proposed data use is consistent with the initial approval by the IRB and consent of the participants. Ideally, this oversight should come from the organization (e.g., NIH Institute) holding the aggregated data, and not from the multiple IRBs that approved the original data collection.

3. Informed consent should take into account the potential incorporation of data into aggregated databases. Research into the attitudes of participants suggests that the majority of them are willing to allow their de-identified data to be used for future research by either the investigator who recruited them or by other investigators in the future [5]. However, it seems likely that very few consent processes have fully anticipated the possibility that data would be included in databases controlled by private corporations, the federal government, or international consortia. Surveys have found that a substantial fraction of participants may not want their data used in this manner. In one report, the overwhelming majority (more than 80%) of survey respondents felt that new consent was required for each use of DNA samples for research [6]. Some may wish to limit the use of their data (e.g., to studies of their own disease or condition). Some may limit the use of their data to exclude the study of mental illness or studies that have the potential to stigmatize their families or ethnic groups. Others may wish their data to be used for nonprofit research only, or have concerns about the custody of their data and their ability to withdraw from research. Legal action brought by the Havasupai tribal members and families with Canavan disease, along with the recent *Washington University v. Catalona* decision, all point to the importance of these considerations [7,8].

Ethically and logically, one cannot give informed consent to unspecified actions that may or may not occur at some time in the future. As pointed out by Arnason in a discussion of Icelanders' participation in the deCODE Genetics databases, participants were given a choice of consent forms to sign—one that allowed only the current research, and one that allowed future research on stored samples [9]. Nonetheless, it is not practicable to incorporate all future research possibilities into a research protocol or consent process. Arnason suggests that informed consent may not be possible in this context. At best, participants can grant permission to unspecified research by indicating their understanding of the scope of future uses, the degree of de-identification/anonymity entailed, whether they will be re-contacted or informed of future findings, the ability (or not) and procedure required to withdraw from the database, the potential users of their data, and measures taken to regulate the database and ensure their privacy is protected. As Caulfield points out [10], this variation on the traditional consent model actually gives participants more autonomy over the use of specimens or data derived there from.

4. Address special challenges of using data obtained from existing databases. Many current studies are being conducted with information and specimens collected for another purpose at substantial expense and labor [11]. Therefore, it seems inefficient and contrary to the public good to let these resources remain unused and attempt to re-establish similar registries or biobanks at public expense. However, it may never be possible to determine the specific wishes of the participants with regard to the re-use of their personal information and specimens. Given such situations, stewards of aggregated databases may consider different options. One involves oversight that includes assurance by contributing investigators that consent was obtained and that, beyond the verbatim license granted for the primary research, broader future uses by different investigators are permitted. Given the size and scope of the individual database, this oversight may come from the IRB or an institutional committee within a single university, or may involve governmental representatives, and an outside group. Second, it may be possible to address the problem of a lack of anticipated secondary use of data or specimens by consulting with legitimate representatives of the donor group(s) and obtaining approval for the new studies [12]. Typically, this would not be a single IRB, but could include patient support groups, veteran's associations, leaders of indigenous people, etc. A third, though costly [13], option would be to re-contact individual participants and ask permission for the new study. There is no clear consensus on which of these (or other) methods to invoke in a particular instance. Nevertheless, prudence suggests establishing a data-use committee to make such decisions. However, if new research is contemplated that carries a risk of individual or group stigmatization, then re-consent may be the appropriate choice regardless of who is responsible for making such decisions.

5. Pursue efforts directed at standardization of data. Realizing the goals of database aggregation requires interoperability among the primary databases. Accordingly, standard data protocols should be developed and used.

For example, participant demographic characteristics and medical conditions should be described using controlled vocabularies allowing careful matching of participants across studies. For example, the PhenX project (http://www.phenx.org/) supported by the National Human Genome Research Institute is developing sets of standardized measures and surveys to be used with genome-wide association studies. The Public Population Project in Genomics (http://www.p3gconsortium.org/) is a multinational collaboration to catalog and share methods, surveys, and other tools needed to do large-scale genetic research [14]. Currently, over 100 studies with planned enrollment of over 11 million participants are part of the effort. The Organisation for Economic Co-operation and Development has recently posted draft guidelines for the operation of genetic databanks that include best practices for data interchange (http://www.oecd.org/document/50/0,3343,en_2649_34537_37646258_1_1_1_1,00.html). The Clinical Data Standards Interchange Consortium has designed standards for the interoperability of clinical trial data submitted to the US Food and Drug Administration as well as globally (http://www.cdisc.org/standards/index.html). Finally, the Open Biomedical Ontologies Foundry project (http://www.obofoundry.org/) is a multinational collaboration that is attempting to coordinate the development of ontologies of biomedical interest, including a sequence ontology for describing genetic loci and sequence variations, a phenotype ontology for describing participant characteristics, and an investigation ontology for describing laboratory methodologies, etc., that will provide the standardized vocabulary necessary for true data interoperability [15].

6. Establish data sharing rules, including attribution of contributions. Investigators spend considerable time, effort, and money collecting data and specimens. Consequently, there is rightfully a sense of entitlement felt by investigators who may not believe it is appropriate to permit other scientists to analyze "their" data before they have completely finished with them. Large sums of public and private money have been spent to build and maintain investigative teams that are conducting original research primarily for the public benefit. These investigators may see little reward and much risk in participating in aggregated databases, despite the potential public good from sharing their data. There also may be concerns regarding the intellectual property created by the original investigator. Further, funding agencies or the investigator's employers may limit the sharing of research data. The question of data ownership will be important not only for publication, but also with regard to grant applications, promotion, etc. There are currently no firm guidelines that protect investigators with regard to ownership and sharing of data. Therefore, stewards of aggregated databases should develop rules that establish the length of time or other conditions determining when an investigator can have absolute control over the data s/he has contributed, even after submission in the aggregated database. Publications and grant applications must clearly describe the origin of data taken from public or quasi-public sources (investigator/repository), and the original contributors must be given appropriate credit.

7. Adopt "best practices" to avoid identifiability of the data. Clinical research data can be collected and stored with several levels of confidentiality. The terminology in this area is confusing, but in general, data can be fully identified, de-identified (coded) but linkable to identifiers, or de-identified and un-linkable to a code (i.e., "anonymized") [16]. Participants may desire that their de-identified information be treated as if it were anonymous, and many conflate anonymity and de-identification, but it is relatively rare for research to be conducted under conditions of true anonymity—blood samples without names or numbers on the tubes, or clinical phenotypes that never contained explicit identifiers, for example. But the general concept of anonymity/identification is not absolute [17].

Genome-wide analyses by their very nature produce uniquely identifying information. It is estimated that between 75–100 single nucleotide polymorphisms or fewer than 20 microsatellite markers can unambiguously identify a single individual [18]. Likewise, a highly detailed clinical phenotype with participant demographic characteristics can be used to re-identify an individual. The most disconcerting scenarios regarding re-identification assume that there is some other database of identified participants to compare with. While it may seem unlikely that an individual's genotype and phenotype would exist in more than one place, techniques such as "trail re-identification" have shown that this is often the case. That is, patients or research participants deposit seemingly de-identified data in multiple hospitals or other institutions that share information, revealing a visit pattern or "trail" that can be linked to individual identities [19]. The threat of sample re-identification will grow as more data are collected and aggregated. The concern extends for database users outside the academic research arena, including pharmaceutical companies, insurance companies, and law enforcement agencies with access to rich storehouses of public and private information to re-identify the phenotypes stored by researchers.

Technical approaches exist that can mask the identities of participants in large databases beyond the ad hoc approaches of removing links between data and identifiers as prescribed in the HIPAA (Health Insurance Portability and Accountability Act) safe harbor provision. First, data can be generalized. For example, in storing genetic sequence information, AGG could become AG(G or C) at polymorphic residues [20,21]. Second, data can be encrypted at the source, and stored in encrypted form in the database server. A researcher would only be able to query a processing engine that would interact with the database and report the results. Analysis of record-level information would still be possible, but that analysis would be done by the database host, and not by downloading the data to client computers. Third, data can be manipulated so that each record that is shared with other users is linked or mapped to a prespecified number of potential identities in the database. Thus the original data retain utility, but now have a much lower chance of being linked to a specific individual. While the computational methodologies necessary to protect data in this manner exist, their refinement and practical application require further development.

Concluding Comments

While there is considerable activity focused on providing public access to clinical trial data and on merging multiple databases, a set of best practices for this type of research is clearly needed. It is hoped that the recommendations offered here will facilitate scientifically and ethically sound research.

Since it is unlikely that the issues and proposed solutions described here will be static over time, they will need to be reviewed periodically. The consistency of data sharing and data protection must be evaluated and maintained. Data protection standards will evolve, and a methodology that was appropriate at one time may not be appropriate later. Equally likely to change are standards imposed by the user community and donor-participant community, and privacy protection methods will need to reflect those changes in order to retain the trust of all stakeholders. This is necessary to help realize the potential for personalized medicine. ■

Acknowledgments

This article summarizes a workshop organized by the University of Texas Southwestern Medical Center at Dallas as part of the Bioinformatics Integration Support Contract. CK was a member of the NIAID Program Staff at the time of the meeting. The authors are listed in alphabetical order, except for DRK and JS. Each author attended the meeting, took part in the discussions, and contributed to the text.

Author Affiliations: David R. Karp is at the Department of Internal Medicine, University of Texas Southwestern Medical Center at Dallas, Dallas, Texas, United States of America. Shelley Carlin is at the Institutional Review Board, University of Texas Southwestern Medical Center at Dallas, Dallas, Texas, United States of America. Robert Cook-Deegan is at the Institute for Genome Sciences and Policy, Duke University, Durham, North Carolina, United States of America. Daniel E. Ford is at the School of Medicine, Johns Hopkins University, Baltimore, Maryland, United States of America. Gail Geller is at the Berman Institute of Bioethics and Department of Medicine, Johns Hopkins University, Baltimore, Maryland, United States of America. David N. Glass is at the Department of Pediatrics, University of Cincinnati, Cincinnati, Ohio, United States of America. Hank Greely is at Stanford Law School, Stanford, California, United States of America. Joel Guthridge is at the Arthritis and Immunology Research Program, Oklahoma Medical Research Foundation, Oklahoma City, Oklahoma, United States of America. Jeffrey Kahn is at the Center for Bioethics, University of Minnesota, Minneapolis, Minnesota, United States of America. Richard Kaslow is at the School of Public Health, University of Alabama at Birmingham, Birmingham, Alabama, United States of America. Cheryl Kraft is at the Division of Allergy, Immunology, and Transplantation, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Rockville, Maryland, United States of America. Kathleen MacQueen is at Family Health International, Research Triangle Park, North Carolina, United States of America. Bradley Malin is at the Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, Tennessee, United States of America. Richard H. Scheuerman is at the Department of Pathology and Division of Biomedical Informatics, University of Texas Southwestern Medical Center at Dallas, Dallas, Texas, United States of America. Jeremy Sugarman is at the Berman Institute of Bioethics, Department of Medicine, and Department of Health Policy and Management, Johns Hopkins University, Baltimore, Maryland, United States of America.

References

1. Guttmacher AE, Collins FS (2005) Realizing the promise of genomics in biomedical research. JAMA 294: 1399-1402.
2. Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661-678.
3. Dalton R (2004) When two tribes go to war. Nature 430: 500-502.
4. Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, et al. (2007) Toward a national framework for the secondary use of health data: An American Medical Informatics Association White Paper. J Am Med Inform Assoc 14: 1-9.
5. Wang SS, Fridinger F, Sheedy KM, Khoury MJ (2001) Public attitudes regarding the donation and storage of blood specimens for genetic research. Community Genet 4: 18-26.
6. Caulfield T (2007) Biobanks and blanket consent: The proper place of the public good and public perception rationales. Kings Coll Law J 18: 209-226.
7. Andrews L (2006) Who owns your body? A patient's perspective on Washington University v. Catalona. J Law Med Ethics 34: 398-407.
8. Hakimian R, Korn D (2004) Ownership and use of tissue specimens for research. JAMA 292: 2500-2505.
9. Arnason V (2004) Coding and consent: Moral challenges of the database project in Iceland. Bioethics 18: 27-49.
10. Caulfield T, Upshur RE, Daar A (2003) DNA databanks and consent: A suggested policy option involving an authorization model. BMC Med Ethics 4: E1.
11. Maschke KJ (2005) Navigating an ethical patchwork—Human gene banks. Nat Biotechnol 23: 539-545.
12. Lavori PW, Krause-Steinrauf H, Brophy M, Buxbaum J, Cockroft J, et al. (2002) Principles, organization, and operation of a DNA bank for clinical trials: A Department of Veterans Affairs cooperative study. Control Clin Trials 23: 222-239.
13. Vates JR, Hetrick JL, Lavin KL, Sharma GK, Wagner RL, et al. (2005) Protecting medical record information: Start your research registries today. Laryngoscope 115: 441-444.
14. Knoppers BM, Fortier I, Legault D, Burton P (2008) The Public Population Project in Genomics (P(3)G): A proof of concept? Eur J Hum Genet 16: 664-665.
15. Smith B, Ashburner M, Rosse C, Bard J, Bug W, et al. (2007) The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 25: 1251-1255.
16. Knoppers BM, Saginur M (2005) The Babel of genetic data terminology. Nat Biotechnol 23: 925-927.
17. Kohane IS, Altman RB (2005) Health-information altruists—A potentially critical resource. N Engl J Med 353: 2074-2077.
18. Lin Z, Owen AB, Altman RB (2004) Genomic research and human subject privacy. Science 305: 183.
19. Malin BA (2005) An evaluation of the current state of genomic data privacy protection technology and a roadmap for the future. J Am Med Inform Assoc 12: 28-34.
20. Lin Z, Hewett M, Altman RB (2002) Using binning to maintain confidentiality of medical data. Proc AMIA Symp: 454-458.
21. Malin BA (2005) Protecting genomic sequence anonymity with generalization lattices. Methods Inf Med 44: 687-692.