29E CONFÉRENCE INTERNATIONALE DES COMMISSAIRES
À LA PROTECTION DES DONNÉES ET DE LA VIE PRIVÉE

LES HORIZONS DE LA PROTECTION DE LA VIE PRIVÉE

TERRA INCOGNITA

PRIVACY HORIZONS

29TH INTERNATIONAL CONFERENCE OF
DATA PROTECTION AND PRIVACY COMMISSIONERS

Atelier
Dragon : *Quand la loi rencontre la technologie*
Protéger la vie privée au moyen de la dépersonnalisation : réalité ou illusion?

Workshop
"Law Meets Technology" Dragon
Protecting Privacy Through De-Identification: Reality or Fallacy?

26 septembre/September 26
13h30 – 16h

Série Terra Incognita, cahier de travail # 7/Terra Incognita, workbook series # 7

Office of the Privacy Commissioner of Canada
Commissariat à la protection de la vie privée du Canada

Table des matières / Table of contents

Biographies

Présidente : M^me^ Ann Cavoukian, Ph. D.

Nommée au poste de commissaire à l'information et à la protection de la vie privée de l'Ontario en 1997, Ann Cavoukian est la première commissaire à laquelle on confie un deuxième mandat. Elle est reconnue comme une des principales spécialistes en matière de protection de la vie privée à l'échelle mondiale. D'aucuns la considèrent comme une conférencière distinguée : elle est d'ailleurs souvent invitée à faire des exposés lors d'importants colloques aux quatre coins du monde. M^me^ Cavoukian a reçu de nombreux prix, notamment de l'Association du Barreau de l'Ontario, de l'Association de psychologie de l'Ontario, et de l'*International Association of Privacy Professionals*, pour souligner le leadership et l'esprit novateur dont elle fait preuve dans le domaine de la protection de la vie privée. Bien connue à cause de son ouvrage précurseur de 1995 sur les technologies permettant d'accroître le respect de la vie privée, elle cherche toujours à incorporer la protection de la vie privée aux spécifications de projets de technologie, assurant ainsi les meilleures mesures de protection. Au nombre de ses œuvres publiées, il y a notamment *Who Knows: Safeguarding Your Privacy in a Networked World* (1997), écrit avec Don Tapscott, et *The Privacy Payoff: How Successful Businesses Build Customer Trust* (2002), écrit avec Tyler Hamilton.

Présidente : M^me^ Debra Grant, Ph. D.

Debra Grant occupe le poste de spécialiste principale de la protection de la vie privée en matière de santé au Bureau du commissaire à l'information et à la protection de la vie privée de l'Ontario, l'organisme indépendant qui surveille l'application des lois provinciales suivantes : la *Loi sur l'accès à l'information et la protection de la vie privée*, la *Loi sur l'accès à l'information municipale et la protection de la vie privée* ainsi que la nouvelle *Loi sur la protection des renseignements personnels sur la santé*. M^me^ Grant a obtenu un doctorat en psychologie sociale de l'Université York en 1991. Au cours des 17 dernières années, elle a œuvré au sein du Bureau du commissaire à l'information et à la protection de la vie privée de l'Ontario dans les domaines de la recherche et de l'élaboration de politiques sur des questions d'accès à l'information et de protection de la vie

Biographies

Chair : Dr. Ann Cavoukian

Dr. Ann Cavoukian was appointed Ontario's Information and Privacy Commissioner in 1997, and is the first to be reappointed for a second term. Dr. Cavoukian is recognized as one of the foremost privacy experts in the world and widely regarded as a distinguished speaker, frequently appearing at major forums around the globe. Dr. Cavoukian is the recipient of many awards including ones from the Ontario Bar Association, the Ontario Psychological Association, and the International Association of Privacy Professionals, for privacy leadership and innovation. Noted for her seminal work on Privacy Enhancing Technologies in 1995, her mantra of "privacy by design" seeks to embed privacy into the design specifications of technology, thereby achieving the strongest protections. Dr. Cavoukian's published works include *Who Knows: Safeguarding Your Privacy in a Networked World* (1997), written with Don Tapscott, and, *The Privacy Payoff: How Successful Businesses Build Customer Trust* (2002), written with Tyler Hamilton.

Chair : Dr. Debra Grant

Debra Grant is a Senior Health Privacy Specialist for the Information and Privacy Commissioner for the Province of Ontario (IPC), the independent body that oversees the provincial *Freedom of Information and Protection of Privacy Act*, the *Municipal Freedom of Information and Protection of Privacy Act*, and the new *Personal Health Information Protection Act*. She graduated in 1991 with a Ph.D. in social psychology from York University. For the past seventeen years, she has worked for the IPC conducting research and developing policies on access and privacy issues in relation to a wide variety of topics including personal health information. She has worked on numerous submissions to the Ontario government on existing and proposed public and private sector privacy legislation and provides expert advice to government and health sector organizations on privacy

privée touchant une vaste gamme de sujets, y compris les renseignements personnels sur la santé. Elle a travaillé à de nombreuses présentations au gouvernement ontarien sur des lois en vigueur et des projets de loi en matière de protection des renseignements personnels applicables aux secteurs public et privé. Elle fournit également des conseils éclairés à des organismes gouvernementaux et du secteur de la santé sur des questions de protection de la vie privée touchant les renseignements personnels sur la santé.

Conférenciers

M. Khaled El Emam, Ph. D.

Khaled El Emam est professeur agrégé à la Faculté de médecine et à l'École d'ingénierie et de technologie de l'informatique de l'Université d'Ottawa et dirige une équipe de chercheurs à l'Institut de recherche du Centre hospitalier pour enfants de l'est de l'Ontario. Il est titulaire d'une chaire de recherche du Canada sur les données électroniques en matière de santé à l'Université d'Ottawa, et l'anonymisation des données fait partie de ses domaines de recherche. M. El Emam a été agent de recherche principal au Conseil national de recherches du Canada et, auparavant, chef du groupe sur les méthodes quantitatives au Fraunhofer Institute à Kaiserslautern (Allemagne). En 2003 et 2004, Khaled El Emam a été désigné meilleur chercheur universitaire du monde en ingénierie des systèmes et en génie logiciel par le *Journal of Systems and Software* pour sa recherche sur l'évaluation et l'amélioration des mesures et de la qualité, et s'est classé au deuxième rang en 2002 et en 2005. Il a obtenu un doctorat du département d'électricité et d'électronique de King's College, à l'Université de Londres (Royaume-Uni). Le site Web de son laboratoire peut être consulté à l'adresse suivante : http://www.ehealthinformation.ca/ (en anglais seulement).

M. William Lowrance, Ph. D.

William Lowrance est consultant en matière d'éthique et politiques relatives à la recherche en santé à Genève. Au cours des dernières années, il s'est intéressé à l'utilisation des renseignements sur la santé et des échantillons biologiques dans

issues in relation to personal health information

Speakers

Dr. Khaled El Emam

Dr. El Emam is an Associate Professor at the University of Ottawa, Faculty of Medicine and the School of Information Technology and Engineering. He is a Canada Research Chair in Electronic Health Information at the University of Ottawa. One of his areas of research is data anonymization. Dr. El Emam was a Senior Research Officer at the National Research Council of Canada and, before that, head of the Quantitative Methods Group at the Fraunhofer Institute in Kaiserslautern, Germany. In 2003 and 2004, Khaled El Emam was ranked as the top systems and software engineering scholar worldwide by the *Journal of Systems and Software* based on his research on measurement and quality evaluation and improvement, and ranked second in 2002 and 2005. He holds a Ph.D. from the Department of Electrical and Electronics, King's College, University of London (UK). His lab's web site is: http://www.ehealthinformation.ca/.

Dr. William Lowrance

Dr. William Lowrance is a consultant in health research policy and ethics, based in Geneva. His focus in recent years has been on the use of health information and biospecimens in research. In 1997 he prepared a report, *Privacy and*

le domaine de la recherche. En 1997, il a présenté un rapport intitulé *Privacy and Health Research* à la secrétaire à la Santé et aux Services sociaux des États-Unis. En 2002, un autre rapport intitulé *Learning from Experience: Privacy and the Secondary Use of Data in Health Research* a été présenté au Nuffield Trust et, en 2006, il a produit le rapport *Access to Collections of Data and Materials for Health Research* pour le Conseil de recherches médicales du Royaume-Uni et le Wellcome Trust. En outre, il a été président du groupe consultatif intérimaire sur l'éthique et la gouvernance pendant le démarrage de la biobanque du Royaume-Uni (BioBank UK). À l'heure actuelle, il termine un projet pour le National Human Genome Research Institute des États-Unis sur la protection de la vie privée, la confidentialité et l'identifiabilité dans le cadre de recherches en génomique.

M. Bradley Malin, Ph. D.

Bradley Malin est professeur adjoint d'informatique biomédicale à la faculté de médecine de l'Université Vanderbilt et il a obtenu une nomination secondaire à la faculté de génie. Il est titulaire d'un baccalauréat en biologie moléculaire, d'une maîtrise en découverte des connaissances et en exploration de données, d'une maîtrise en gestion et politique publiques, de même que d'un doctorat en informatique, diplômes qu'il a tous obtenus à l'Université Carnegie Mellon. Il a publié de nombreux articles scientifiques sur l'informatique biomédicale, l'exploration des données et des liens informatiques et la protection des renseignements personnels. Ses recherches portant sur les bases de données génétiques et sur la protection de la vie privée lui ont valu plusieurs prix de l'American Medical Informatics Association et d'autres organismes internationaux. Il a dirigé plusieurs ateliers sur la protection de la vie privée et le forage des données pour le compte du IEEE et de l'ACM. De 2004 à 2006, il a été rédacteur en chef du *Journal of Privacy Technology* (JOPT) et il est le directeur scientifique invité d'un numéro spécial sur la protection des renseignements personnels et le forage des données, qui sera publié par *Data and Knowledge Engineering*.

Health Research, for the U.S. Secretary of Health and Human Services; in 2002 a report, *Learning from Experience: Privacy and the Secondary Use of Data in Health Research*, for The Nuffield Trust; and in 2006 a report, *Access to Collections of Data and Materials for Health Research*, for the Medical Research Council U.K. and the Wellcome Trust. He chaired the Interim Advisory Group on Ethics and Governance during the startup of the U.K. Biobank. Currently he is finishing a project for the U.S. National Human Genome Research Institute on privacy, confidentiality, and identifiability in genomic research.

Dr. Bradley Malin

Bradley Malin is an Assistant Professor of Biomedical Informatics in the School of Medicine at Vanderbilt University and holds a secondary appointment in the School of Engineering. He received a bachelor's in molecular biology, a master's in knowledge discovery and data mining, a master's in public policy and management, and a doctorate in computer science, all from Carnegie Mellon University. He is the author of numerous scientific articles on biomedical informatics, data and link mining, and data privacy. His research in genetic databases and privacy has received several awards from the American and International Medical Informatics Associations. He has chaired various workshops on privacy and data mining for the IEEE and ACM. From 2004 through 2006 he was the managing editor of the Journal of Privacy Technology (JOPT) and is the guest editor for an upcoming special issue on privacy and data mining for the journal Data and Knowledge Engineering.

M. Donald J. Willison, Ph. D.

Donald J. Willison est professeur agrégé au Département d'épidémiologie clinique et de biostatistique de l'Université McMaster à Hamilton, Ontario. Il détient un baccalauréat en pharmacie de l'Université de Toronto, une maîtrise en conception, mesures et évaluation de l'Université McMaster et un doctorat en évaluation de programmes du Département de politique et de gestion en matière de santé, de la Harvard School of Public Health. Parmi ses intérêts en recherche, il y a notamment la gouvernance de l'utilisation de renseignements personnels pour la recherche en santé et la création de systèmes novateurs pour connaître les choix en matière de consentement en vue de la participation à de la recherche. Il a travaillé avec les Instituts de recherche en santé du Canada à l'élaboration d'un grand nombre d'initiatives stratégiques liées à l'utilisation de renseignements personnels pour la recherche en santé. La plus récente de ces politiques s'intitule Pratiques exemplaires des IRSC en matière de protection de la vie privée dans la recherche en santé (septembre 2005).

Dr. Donald J. Willison

Willison is Associate Professor, Department of Clinical Epidemiology & Biostatistics at McMaster University in Hamilton, Ontario. His training combines an undergraduate degree in pharmacy from the University of Toronto, a Master's degree in Design, Measurement, and Evaluation from McMaster University, and a Doctorate in Program Evaluation from the Department of Health Policy and Management, Harvard School of Public Health. Dr. Willison's research interests include governance over use of personal information for health research and the development of innovative systems for eliciting consent choices for participation in research. He has worked with the Canadian Institutes of Health Research on several policy initiatives related to the use of personal information for health research. The most recent of these is the CIHR Best Practices for Protecting Privacy in Health Research, September 2005.

29E CONFÉRENCE INTERNATIONALE DES COMMISSAIRES
À LA PROTECTION DES DONNÉES ET DE LA VIE PRIVÉE

LES HORIZONS DE LA PROTECTION DE LA VIE PRIVÉE

TERRA INCOGNITA

PRIVACY HORIZONS

29TH INTERNATIONAL CONFERENCE OF
DATA PROTECTION AND PRIVACY COMMISSIONERS

Directives sur la dépersonnalisation des renseignements personnels sur la santé pour l'ensemble du Canada

Pan-Canadian De-Identification Guidelines for Personal Health Information

Par/by:

Khaled El Emam, Elizabeth Jonker, Scott Sams, Emilio Neri, Angelica Neisa, Tianshan Gao et/and Sadrul Chowdhury

Avril 2007/April 2007

Office of the Privacy Commissioner of Canada

Commissariat à la protection de la vie privée du Canada

De plus en plus, on recueille, utilise et communique des renseignements personnels sur la santé à des fins commerciales, de recherche ou d'élaboration de politiques. Toutefois, la population canadienne se montre préoccupée par le nombre croissant de brèches dans la sécurité et par les pressions exercées pour échanger les renseignements personnels. L'une des façons d'atténuer les répercussions des brèches dans la sécurité entraînant la perte de renseignements personnels et de faciliter l'échange légitime de ces renseignements consiste à les rendre anonymes.

Le présent rapport expose les résultats d'une série d'études visant à guider les pratiques de dépersonnalisation des données. Ces études avaient pour objectif de déterminer la manière dont des données dépersonnalisées peuvent être repersonnalisées. Le fait de comprendre le processus de repersonnalisation et les risques associés à cette opération rend possible la mise au point de techniques de dépersonnalisation plus efficaces. Les études portaient principalement sur la repersonnalisation des données par couplage de dossiers. À l'aide des conclusions tirées de ces études, on propose un processus de dépersonnalisation des données et un outil logiciel.

Un résumé des études réalisées est présenté ci-dessous.

Constitution de base de données d'identification à partir de sources publiques

Dans cette étude, on a examiné la disponibilité de renseignements provenant de sources publiques pouvant servir à la repersonnalisation malveillante de données. Les renseignements publics tirés de sources multiples peuvent être combinés afin de créer une base de données d'identification permettant de se livrer à une telle utilisation malveillante. On a trouvé un certain nombre de sources publiques, comme le *Private Property Security Registry*, le Registre des titres de propriété, les annuaires téléphoniques et l'annuaire de recherche inverse d'adresses de Postes Canada. Ces sources publiques se rattachent à des sous-populations. Il n'a pas été possible de constituer une base de données se rapportant à une population complète qui aurait permis de se livrer à la repersonnalisation malveillante de données au moyen de couplage de dossiers (p. ex. l'ensemble du Canada ou

Personal health information is increasingly being collected, used, and disclosed for research, policy, and commerce purposes. However, the rise in security breaches and the pressure to share personal information is a concern for the Canadian public. One way to minimize the impact of security breaches that result in the loss of personal information and to facilitate its legitimate sharing is to anonymize it.

In this report, we present the results of a series of studies to inform data anonymization practices. The focus of the studies was to determine how anonymized data can be re-identified (i.e. reverse anonymization). By understanding how re-identification can be performed and what the risks of successful re-identification are, we can develop more effective techniques for anonymization. The focus was primarily on re-identification through record linkage. Based on our findings, we provide a data anonymization process and software tool. Below is a summary of the studies that were conducted.

Construction of Identification Databases from Public Sources

In this study we examined the availability of public information that can be used for re-identification attacks. Public information from multiple sources can be combined to create an identification database suitable for such an attack. A number of public sources were identified such as the Private Property Security Registry, Land Registry, telephone directory, and Canada Post address reverse lookup directory. These public sources pertain to sub-populations. It was not possible to construct a full population database suitable for a re-identification attack using record linkage (e.g., all of Canada or all of Ontario). However, it was possible to construct identification databases for professionals whose associations or employers publish their membership lists (i.e. the College of Physicians and Surgeons of Ontario, the Law Society of Upper Canada and the Government Electronic Directory Service) and for home owners.

Inference Attacks

When there is insufficient public information to launch a re-identification attack on a database, it may be possible to infer some of the missing information. We examined different types of inference attacks for some common types of variables. We found that it was possible to predict gender and

l'ensemble de l'Ontario). Toutefois, il a été possible de constituer des bases de données d'identification relatives à des professionnels dont les associations ou les employeurs publient la liste de leurs membres (p. ex., l'Ordre des médecins et chirurgiens de l'Ontario, le Barreau du Haut-Canada et les Services d'annuaires gouvernementaux électroniques), ainsi que des bases de données relatives aux propriétaires d'habitation.

Inférence malveillante de données

Lorsque les renseignements personnels obtenus de sources publiques sont insuffisants pour repersonnaliser de façon malveillante les données d'une base de données, il est parfois possible d'obtenir les renseignements manquants par inférence. On a examiné différents types d'inférence malveillante de données pour certains types de variables répandues. On a constaté qu'il était possible de prédire le sexe et l'année de naissance d'une personne à partir des prénoms et de l'année où elle a obtenu son diplôme, respectivement. Il n'était pas possible de prédire avec précision les codes postaux à partir d'autres codes postaux dans un dossier, mais il était possible de le faire de manière relativement précise dans le cas des codes postaux ruraux. Ces résultats laissent croire que l'inférence malveillante permet de trouver les renseignements manquants pour certains types de variables; cette réalité devrait être prise en considération au moment de déterminer l'utilité d'une base de données d'identification dans le contexte d'une repersonnalisation malveillante de données.

Évaluer le risque associé à la repersonnalisation

Une fois que l'on a des sources publiques de renseignements, en plus des renseignements supplémentaires obtenus par inférence malveillante, quelle est la probabilité qu'une personne soit réellement en mesure de réussir à repersonnaliser de façon malveillante les données d'un ensemble de données canadien? L'objectif consiste à évaluer les risques associés à la repersonnalisation par le truchement de variables d'identification indirectes, ou quasi-identifiants (c.-à-d. le sexe, la date de naissance, le code postal, etc.). On a constaté que seul un petit sous-ensemble des quasi-identifiants représentait un faible risque de repersonnalisation, de manière

year of birth from the first names and graduation year respectively. It was not possible to accurately predict postal codes from other postal codes in a record, but it was possible to do so relatively accurately for rural postal codes. These results suggest that inference attacks can fill in the gaps for some types of variables, which should be taken into consideration when deciding on the utility of an identification database for a re-identification attack.

Measuring the Risk of Re-identification

Once we have public sources of information, augmented with additional information from inference attacks, what is the probability of someone actually being able to launch a successful re-identification attack on a Canadian data set ? We focus here on the risk of re-identification via quasi-(indirect) identifying variables (i.e. gender, date of birth, postal code, etc.). We found that only a small subset of the quasi-identifiers represented a consistently low risk of re-identification across both sample size changes and data set changes. Most quasi-identifiers were not stable and therefore presented an unsafe risk of re-identification. Consequently, the success rate for re-identification was found to be quite high under certain circumstances.

Personal Information on the Web

The risk quantification above indicates that the re-identification risks are not trivial; however, people tend to be willing to trade their privacy for some personal benefit. We examine what type of personal data Canadian job seekers are willing to expose on the public web. Are they willing to expose the type of information that is needed for an identification attack? The answer is yes. Job seekers post sufficient personal information about themselves on the public web for some simple re-identification attacks.

Personal Information and Data Remnants

We then examine the kind of data that Canadians leave on their computer disk drives when they non-destructively de-commission them. The study collected 60 second hand disk drives across the country and extracted their data remnants. We found that the majority of drives had some personal information about their owners, but little personal health information was found about the owners. However, more personal health information

uniforme, pour toutes les tailles d'échantillon et tous les types d'ensembles de données. La plupart des quasi-identifiants n'étaient pas stables et représentaient donc un risque dangereux de repersonnalisation. Par conséquent, on a trouvé que le taux de réussite de la repersonnalisation pouvait être assez élevé dans certaines circonstances.

Renseignements personnels sur le Web

La quantification des risques ci-dessus indique que les risques associés à la repersonnalisation ne sont pas négligeables; toutefois, les gens sont enclins à révéler des renseignements personnels en échange d'un avantage quelconque. On a examiné le type de renseignements personnels que les Canadiennes et les Canadiens à la recherche d'un emploi sont disposés à publier sur le Web public. Sont-ils disposés à divulguer le type de renseignements nécessaires à une utilisation malveillante de données d'identification? Oui. Les personnes à la recherche d'un emploi affichent suffisamment de renseignements personnels les concernant sur le Web public pour procéder à une repersonnalisation malveillante.

Renseignements personnels et restes de données

On a ensuite examiné le type de données que les Canadiennes et les Canadiens laissent sur les disques durs de leurs ordinateurs quand ils cessent de les utiliser, sans les détruire. Dans le cadre de l'étude, on a recueilli 60 disques usagés dans tout le pays et on a extrait les restes de données qui s'y trouvaient. On a constaté que la plupart des disques durs contenaient certains renseignements personnels concernant leurs propriétaires, mais que seule une infime partie de ces renseignements personnels étaient relatifs à la santé. Toutefois, on a trouvé davantage de renseignements personnels sur la santé qui se rapportaient à des personnes autres que les propriétaires des disques durs. On conclut qu'il est nécessaire pour les organisations et les personnes de prendre des mesures afin de réduire le risque associé aux fuites de données personnelles provenant de disques durs usagés.

Recommandations

D'après les conclusions tirées de ces études, nous avons mis au point un processus de dépersonnalisation des données concret appuyé

was found on the drives that pertained to people other than the owners of the drives. We conclude that there is a need for organizations and individuals to take actions to reduce the risk of personal data leakage from second hand disk drives.

Recommendations

Based on our findings in these studies, we have formulated a concrete data anonymization process, with some automated tool support. Following this process will allow the end-user to manage re-identification risks in their data releases. We have also included some more general recommendations for protecting personal health information privacy, as well as suggestions for future research in this area.

The Report

The full report can be downloaded from the following web site:
http://www.ehealthinformation.ca/documents/OPCReportv11.pdf

par des outils automatisés. L'utilisateur final qui suivra ce processus sera en mesure de gérer les risques associés à la repersonnalisation dans le cadre de la communication de données. On a également formulé des recommandations d'ordre plus général pour la protection des renseignements personnels sur la santé, de même que des suggestions pour de futures recherches dans ce domaine.

Rapport

La version intégrale du rapport peut être téléchargée à l'adresse suivante : http://www.ehealthinformation.ca/documents/OPCReportv11.pdf

29E CONFÉRENCE INTERNATIONALE DES COMMISSAIRES
À LA PROTECTION DES DONNÉES ET DE LA VIE PRIVÉE

LES HORIZONS DE LA PROTECTION DE LA VIE PRIVÉE

TERRA INCOGNITA

PRIVACY HORIZONS

29TH INTERNATIONAL CONFERENCE OF
DATA PROTECTION AND PRIVACY COMMISSIONERS

Vie privée, confidentialité et identifiabilité dans la recherche en génomique

Privacy, Confidentiality, and Identifiability in Genomic Research

Par/by:

William W. Lowrance, Ph. D.

Octobre 2006/October 2006

Office of the Privacy Commissioner of Canada
Commissariat à la protection de la vie privée du Canada

1. Les défis

Protéger la vie privée *et* encourager la recherche

Pour des raisons qu'il n'est pas nécessaire de reprendre ici, il faut protéger les renseignements personnels qui sont recueillis aux fins des soins de santé, des paiements ou de la recherche et respecter la dignité et les droits personnels des sujets sources des données et des sources d'échantillons biologiques. Cette obligation prime dans la plupart des pays et elle est consacrée dans l'éthique professionnelle et en droit.

(Une remarque concernant la portée. Notre projet est axé sur la situation aux États-Unis, mais les questions que nous abordons devraient se poser de façon similaire partout où s'effectue la recherche en génomique.)

Aux États-Unis, les droits relatifs aux renseignements personnels manipulés aux fins des soins de santé ou de la recherche sont régis par deux lois d'ordre général, le Common Rule on Protection of Human Subjects (ci-après le « Common Rule »)[1] et le Privacy Rule de la *Health Information Portability and Accountability Act* (le « Privacy Rule de la HIPAA »)[2,3]. Il est aussi question de ces droits dans nombre de lois et règlements des États, dont certains portent spécifiquement sur les données génétiques. Les renseignements personnels détenus par le gouvernement fédéral sont protégés par la *Privacy Act* et les lois d'applications de certains organismes.

Dans l'Union européenne, la protection des renseignements personnels est garantie par des lois nationales qui transposent la directive générale de l'UE, appelée Directive relative à la protection des données.[4] Par exemple, il y a la *Data Protection Act* au Royaume-Uni.

Au-delà du respect de la loi et des règles éthiques, bien sûr, les chercheurs doivent toujours tâcher d'« agir correctement », par respect de la personne humaine et pour gagner et conserver la confiance de ceux qui, dans la population, consentent librement à ce que leur personne, des échantillons biologiques prélevés d'eux ou des données à leur sujet fassent l'objet de la recherche.

1. The challenges

Protect privacy *and* foster research

For reasons that need not be rehearsed here, personal information collected in the course of health care, payment, or research must be protected, and the personal rights and dignity of data-subjects and biospecimen sources must be respected. This obligation prevails in most countries and is enshrined in professional ethics and in law.

(A note regarding scope. This project is oriented to the U.S. situation, but similar issues should be of concern everywhere genomic research is pursued.)

In the U.S., rights relating to personal information handled in health care and/or research are governed by two omnibus regulations, the Common Rule on Protection of Human Subjects (hereafter, "Common Rule")[1] and the Privacy Rule under the Health Information Portability and Accountability Act ("HIPAA Privacy Rule").[2,3] They are also addressed by many State statutes and regulations, some of which focus specifically on genetic data. Personal data held by the Federal government are protected by the Privacy Act and the enabling statutes of some agencies.

In the European Union, informational privacy is guaranteed by national laws transposing the broad EU Data Protection Directive.[4] An example of such a law is the U.K. Data Protection Act.

Beyond conforming to law and ethical strictures, of course, the research community must always try to "do the right thing," both for reasons of basic decency and to earn and maintain the trust of members of the public who voluntarily allow themselves, data about themselves, or biospecimens from themselves to be studied in research.

The challenge is to protect privacy and foster research at the same time. A principal strategy for achieving this, especially in database-centered research, is to shield the identities of the people the data are about by blurring, removing, destroying or otherwise altering information that could lead to identification of the subjects. This kind of research is about cases and categories, not people.

Le défi consiste à protéger la vie privée tout en favorisant la recherche. Une des principales stratégies pour y parvenir, surtout dans la recherche centrée sur des bases de données, est de protéger l'identité des personnes sur lesquelles portent les données en brouillant, enlevant, détruisant ou modifiant autrement les informations qui pourraient permettre d'identifier les sujets. Ce genre de recherche porte sur des cas et des catégories et non sur des personnes.

Repenser la libre diffusion des données?

Le contexte du projet nous vient en partie de l'habitude de diffusion rapide et libre, ou pratiquement libre, des données génomiques qu'ont adoptées les scientifiques et les établissements intéressés depuis les débuts de l'entreprise génomique. Voilà un trait déterminant de la mentalité dans le domaine. L'habitude a grandement facilité la recherche par delà les frontières politiques, sectorielles et disciplinaires, allant de pair et même favorisant l'importance croissante donnée à la mise en commun des données scientifiques et la libre publication, de façon générale.

Dès 1991, le National Center for Human Genome Research (prédécesseur du NHGRI) et le département de l'Énergie ont demandé que les données de séquençage soient diffusées dans les six mois. Puis, en 1996, le Consortium international pour le séquençage du génome humain a adopté les « principes des Bermudes », pour encourager l'entrée rapide dans le domaine public des séquences assemblées d'une ou deux kilobases ou plus[5]. En 1997, le NHGRI a exigé que ceux à qui il donnait des subventions diffusent dans les 24 heures les séquences assemblées. En 2000, l'Institut a revu sa politique pour exiger la publication hebdomadaire des tracés de séquences brutes[6]. Les NIH et d'autres établissements ont constitué les bases de données nécessaires pour recevoir et diffuser les données. La démarche a si bien servi la recherche en rapide évolution que les intéressés ont établi les « principes de Fort Lauderdale », en 2003, pour préciser les rôles des organismes subventionnaires, des producteurs de données et des utilisateurs de données[7].

HapMap est un exemple de projet qui respecte les principes[8] :

Re-think open release of data?

Part of the backdrop to this project is the cultural habit of rapid, open release, or at least fairly open release, of genomic data that has been adopted by the involved scientists and institutions since the beginning of the genomic endeavor. Sociologically it is one of the defining features of the field. It has greatly facilitated research across political, sectoral, and disciplinary boundaries, and it has been consonant with, indeed has stimulated, the growing emphasis on scientific data-sharing and open publication generally.

As early as 1991 the National Center for Human Genome Research (the predecessor to NHGRI) and the Department of Energy required that sequence data be released within six months. Then in 1996 the International Human Genome Sequencing Consortium adopted so-called "Bermuda Principles," which encouraged rapid release into the public domain of sequence assemblies of 1-2 kilobases or greater.[5] In 1997 NHGRI required grantees to release sequence assemblies within 24 hours, and in 2000 it extended the policy to require weekly publication of raw sequence traces.[6] NIH and other institutions set up the necessary databases to receive and distribute the data. The approach served the fast-moving research so effectively that in 2003 the community developed more detailed "Fort Lauderdale Principles," outlining roles for funding organizations, data producers, and data users.[7]

HapMap is an example of a project that follows the Principles:[8]

> As is now standard practice in large-scale genomic research projects, the International HapMap Consortium follows a policy of releasing data as quickly as possible, anticipating that they will be useful for many investigators. The Consortium anticipates that the Project's data will be used in many ways, such as in developing new analytical methods, in understanding patterns of polymorphism, linkage disequilibrium, and haplotype associations, and in guiding selection of markers to map genes affecting specific diseases. Thus, the Consortium recognizes that the data are available to all users for any purpose.

> Comme il est maintenant pratique courante en recherche génomique à grande échelle de rendre les données publiques dès que possible, le Consortium international HapMap suit cette politique, en prévoyant que les données seront utiles pour beaucoup de chercheurs. Le Consortium prévoit que les données du projet seront utilisées de plusieurs manières, par exemple, pour développer de nouvelles méthodes analytiques, pour améliorer la compréhension des modèles de polymorphisme, du déséquilibre de liaison et des associations d'haplotypes, ainsi que pour guider le choix des marqueurs afin de cartographier les gènes de susceptibilité à des maladies spécifiques. Par conséquent, le Consortium reconnaît que les données sont à la disposition de tous les utilisateurs et dans n'importe quel but.

De l'avis des NIH, de façon générale, la diffusion des données devrait être aussi large et libre que possible tout en préservant la vie privée des participants et en protégeant les données exclusives. Par conséquent, les Instituts exigent que ceux qui demandent des subventions de plus de 500 000 $ incluent dans leur demande un plan de mise en commun des données définitives de recherche ou produisent une justification claire de la non mise en commun[9]. Les raisons invoquées par les NIH sont les suivantes :

> [traduction]
> Le partage des données renforce la libre investigation scientifique, encourage la diversité d'analyse et d'opinion, stimule de nouvelles recherches, permet de vérifier des hypothèses originales ou différentes et d'expérimenter des méthodes d'analyse, fonde des études sur les méthodes de collecte de données et les mesures, facilite la formation des nouveaux chercheurs, permet d'étudier des aspects non envisagés par les premiers chercheurs et autorise la création de nouveaux ensembles de données par la réunion de données de sources multiples.

(Une justification moins souvent invoquée, du moins jusqu'à récemment, est que leur diffusion hâtive et libre établit l'antériorité des données et en fait des « connaissances préalables » évidentes, empêchant ainsi qu'elles fassent l'objet de prétentions exclusives dans des brevets, ce qui en limiterait la disponibilité pour la recherche.)

NIH generally believes that "data should be made as widely and freely available as possible while safeguarding the privacy of participants and protecting confidential and proprietary data." Accordingly, it requires that applicants for grants exceeding $500,000 include in their application a plan for sharing final research data, or a clear justification of not sharing.[9] NIH's rationale is that:

> Sharing data reinforces open scientific inquiry, encourages diversity of analysis and opinion, promotes new research, makes possible the testing of new or alternative hypotheses and methods of analysis, supports studies on data collection methods and measurement, facilitates the education of new researchers, enables the exploration of topics not envisioned by the initial investigators, and permits the creation of new datasets when data from multiple sources are combined.

(Less often mentioned as a rationale, at least until recently, is that early open release of data establishes their precedence as obvious "prior knowledge," thereby preventing their being claimed as proprietary know-how in patents and made less accessible for research use.)

There is no question about the research advantages of such principles or policies. Nor is there question about flexibility, as the Fort Lauderdale Principles and the policies based on them have never been construed as being absolute or encouraging the transgression of people's rights.

However. Part of the remit of this project is to examine whether in the future, genomic data, with various clinical or other associated data, will have to be modified to reduce identifiability and/or held back for controlled release, to a greater extent than has been done up to now.

A cautionary remark about language. Experts in this arena often speak of "public" data or "open" publication. But the usages are sloppy, and may refer either to data that truly are in the public domain, as when posted on a freely accessible website, or, quite differently, to data that only the "professional public" outside the custodian organization may apply to use, under conditional terms. Care should be taken with "public" and "open," as either can justifiably be understood by lay audiences as implying that data are being exposed to plain public view.

De tels principes et les politiques qui en découlent ont d'indéniables avantages pour la recherche. Leur souplesse ne fait pas de doute non plus, puisque les principes de Fort Lauderdale et les politiques qu'ils fondent n'ont jamais été interprétés comme étant absolus ou comme encourageant la transgression de droits des gens.

Cela dit, notre projet vise notamment à examiner s'il faudra modifier les données génomiques, et diverses données cliniques ou autres connexes, pour les rendre moins identifiables ou s'il faudra en limiter la diffusion plus qu'on ne l'a fait jusqu'à maintenant.

Un mot d'avertissement au sujet de la terminologie. Les spécialistes du domaine parlent souvent de données « publiques » ou de publication « libre ». Or, l'usage est relâché, car on désigne ainsi tantôt les données qui sont véritablement du domaine public, par exemple celles affichées sur un site Web accessible à tous, tantôt des données que seuls des spécialistes peuvent demander d'utiliser, à certaines conditions, à l'extérieur de l'organisme qui en a la garde. Il faut employer avec prudence les qualificatifs « public » et « libre », car les profanes peuvent légitimement penser que ces termes supposent que les données sont à la portée de tous.

Problèmes qui surgissent alors que la génomique parvient à maturité

De nombreuses initiatives récentes d'envergure – y compris le Genetic Association Information Network (GAIN), The Cancer Genome Atlas (TCGA), le Genes and Environment Initiative (GEI), la biobanque du Royaume-Uni, le programme de séquençage médical du NHGRI et bien d'autres projets – vont[10] :

- produire des données élaborées, qui couvrent une grande partie du génome et qui sont propres à une personne;
- catégoriser de nombreuses données se rattachant à des gènes pathologiques ou aux diagnostics;
- maintenir des liens, du moins indirectement, avec les données cliniques, familiales, sociales et démographiques;
- faire tout ce qui précède à partir d'un matériel

Issues arising as genomics matures

Many big new initiatives − including the Genetic Association Information Network (GAIN), The Cancer Genome Atlas (TCGA), the Genes and Environment Initiative (GEI), U.K. Biobank, the NHGRI Medical Sequencing Program, and many other projects − will:[10]

- generate data that are fine-grained, of wide genome coverage, and person-specific
- categorize many data with respect to disease-related genes or disease diagnosis
- maintain links, at least indirectly, to clinical, family, social, and demographic data
- and do all this on material from very large numbers of people.

And they will release the resulting data into a world that will:

- continue to assemble identified, or at least circumstantially characterized, police, military, and other DNA and genomic reference collections
- increasingly integrate genomic data with personal medical records
- as genotyping costs drop and knowledge increases, become more routinely capable of matching data to reference collections and inferring probabilistic implications for physical appearance, mental health, and illness risks
- continue to amass databases on most aspects of people's lives, with incentives to link those databases with genomic data for research, healthcare, public health, administrative, marketing, forensic, and other purposes
- continue to worry about the risks of erroneous or malicious identity disclosure and consequent embarrassment, blackmail, group stigmatization, financial fraud, or negative discrimination for health or life insurance, employment, promotion, mortgages, or loans.

Recently several observers have served serious notice that genomic data are becoming more identifiable.

provenant de très grands nombres de populations.

Et les résultats seront diffusés dans un monde qui va :

- continuer de constituer des collectes de référence d'ADN et de ressources génomiques qui sont identifiées ou indirectement caractérisées et qui servent aux policiers, aux militaires ou à d'autres;
- intégrer de plus en plus les données génomiques aux dossiers médicaux personnels;
- le prix du génotypage baissant et les connaissances augmentant, pouvoir de façon plus habituelle apparier les données aux collectes de référence et déduire par inférence probabiliste l'apparence physique, la santé mentale et les risques de maladie;
- continuer de garnir des bases de données sur la plupart des aspects de la vie des gens, avec des incitations à rapprocher ces ensembles des données génomiques pour la recherche, les soins de santé, la santé publique, l'administration, le marketing, l'analyse judiciaire et d'autres fins;
- continuer de s'inquiéter des risques de divulgation erronée ou malveillante de l'identité et de l'embarras qui en résulte et de craindre le chantage, la stigmatisation de groupe, la fraude financière ou la discrimination à l'égard de certains demandeurs d'assurance vie ou santé, d'emploi, de promotion, d'hypothèque ou de prêt.

Récemment, plusieurs observateurs ont servi une sérieuse mise en garde quant à l'identifiabilité grandissante des données génomiques.

Malin et Sweeney ont montré que des séquences d'ADN sans informations démographiques ou identificateurs, si elles faisaient l'objet d'une interprétation à l'égard de certains gènes pathologiques et d'un criblage probabiliste d'après des données publiques (p. ex. les données détaillées de sorties des hôpitaux qui sont accessibles au public dans certains États), peuvent parfois être rapportées à quelques personnes[11].

Malin and Sweeney showed that DNA sequences unlabelled as to demographics or identifiers, if interpreted for some common disease genes and probabilistically screened against publicly available data (such as the detailed hospital discharge data that are publicly accessible in some States), can sometimes be narrowed-down to a few individuals.[11]

In a different approach, Malin argued that DNA sequences can be mapped against family disease-incidence patterns, hospital visit patterns, and other data, and be identified by "trail analysis."[12]

Concerned that "genome-wide association studies now routinely use more than 100,000 SNPs to genotype individuals" and that current protections are inadequate, McGuire and Gibbs have recommended that sequencing research be clearly designated human-subjects research − thus requiring more elaborate consent and closer scrutiny by Institutional Review Boards (IRBs) − and that tiered approaches, in which the data-subjects have more say in determining how data are released and by whom they can be used, should be adopted for release of genomic data.[13]

Getting things right in this area, Foster and Sharp have remarked, will have implications broader than (just) facilitating genomic research:[14]

> The challenges of using linked genotype−phenotype data for medical-sequencing projects prefigures issues that will arise in future uses of many existing biological samples linked to phenotypic information, including disease registries, hospital-based tissue collections and prospective cohort studies. Thus, developing effective strategies for addressing these challenges in medical-sequencing research can inform a much broader set of issues in the ethical conduct of research.

And finally, an observation by this author about subject selection and a question about consent. The Human Genome and HapMap projects have genotyped DNA from only a few very carefully selected people who have consented to the analysis and open publication only after thorough explanation and discussion. But such painstaking selection and consent negotiation cannot as a general matter be expected for future projects involving many people. As an ethical matter,

Dans une démarche différente, Malin a soutenu que les séquences d'ADN peuvent être cartographiées en fonction de l'incidence de maladies familiales, des passages dans les hôpitaux et d'autres données et être identifiées au moyen d'une « analyse des pistes »[12].

Inquiets des études d'association à l'échelle du génome qui utilisent maintenant couramment plus de 100 000 polymorphismes de nucléotide simple (SNP) pour génotyper des individus et que les protections actuelles sont insuffisantes, McGuire et Gibbs ont recommandé que la recherche sur le séquençage soit clairement désignée comme recherche avec des êtres humains – de façon à ce qu'elle doive faire l'objet d'une procédure plus approfondie de consentement et d'un examen plus minutieux par les comités d'examen institutionnels (CEI) – et que des approches à plusieurs niveaux soient adoptées à l'égard de la diffusion des données génomiques, approches où les sujets sources des données ont davantage leur mot à dire sur la façon dont les données sont diffusées et par qui elles peuvent être utilisées[13].

Foster et Sharp ont fait remarquer que de mettre les choses en règle dans ce domaine fera plus que de (simplement) faciliter la recherche en génomique[14] :

> [traduction]
> Les difficultés que présente l'emploi de données couplées de génotype et phénotype dans les travaux de séquençage médical préfigurent les problèmes qui surviendront dans les utilisations futures de nombreux échantillons biologiques existants couplés à des informations phénotypiques, y compris les registres de maladie, les collectes de tissus des hôpitaux et les études prospectives de cohorte. Par conséquent, établir des stratégies efficaces pour régler ces difficultés dans la recherche de séquençage médical peut éclairer un ensemble beaucoup plus vaste de questions de conduite éthique en recherche.

Pour finir, l'auteur soumet une observation de son cru à propos du choix des sujets et une question à propos du consentement. Le projet du génome humain et celui de HapMap ont génotypé l'ADN de quelques personnes triées sur le volet qui ont consenti à l'analyse et à la publication à grande diffusion uniquement après qu'on leur ait bien expliqué de quoi il s'agissait et après en avoir

should consent be relied upon to justify open publication of data that are potentially identifiable?

There is ample cause for concern, and work to be done.

2. Identifiability and identifiers

"Identifiability"

Identifiability is the potential associability of data with persons. Identifiability runs a spectrum, from overtly identified, to possibly deductively identifiable, to absolutely unidentifiable.

In legal regimens, indirect identifiability is as important as direct. For instance, the HIPAA Privacy Rule applies to "individually identifiable health information," i.e. "information that identifies an individual; or with respect to which there is a reasonable basis to believe the information can be used to identify the individual" (§160.103).

Similarly, the U.K. Data Protection Act applies to all "personal data," "data which relate to a living individual who can be identified – (a) from those data, or (b) from those data and other information which is in the possession of, or is likely to come into the possession of, the data controller," the person legally responsible for determining the purposes for which, and manner in which, the data are handled (§I.1-(1)).[15]

If data aren't identifiable they aren't "personal," and a variety of rights and obligations that apply to personal data may not be relevant. General rights of informational privacy, and professional obligations of medical confidentiality, for example, usually apply only to data that are "about" real people.

Almost all health data are initially collected as identified data, whether for healthcare, public health, or research purposes. Data can be de-identified in a variety of ways and to varying degrees, either irreversibly or reversibly. De-identification is a crucial strategy for research.

Identifiers

For any set of data about people, three sorts of identifying factors − commonly, although a bit too casually, referred to as "identifiers" − can be distinguished:

discuté. Or, ce soin apporté à la sélection des sujets et à l'obtention du consentement est impensable, de façon générale, dans les projets à venir qui mettront en jeu de nombreuses personnes. Sur le plan éthique, peut-on s'en remettre au consentement pour justifier la libre publication de données susceptibles d'être identifiées?

Il y a amplement de quoi s'inquiéter et du travail à faire.

2. Identifiabilité et identificateurs

« Identifiabilité »

Par « identifiabilité », on entend la possibilité d'associer les données aux personnes. Le degré d'identifiabilité varie, le sujet pouvant être soit identifié clairement, soit possiblement identifié par déduction ou encore absolument non identifiable.

Dans les régimes juridiques, l'identifiabilité indirecte est aussi importante que l'identifiabilité directe. Ainsi, le *Privacy Rule* de la HIPAA s'applique [traduction] « à l'information sur la santé individuellement identifiable », c. à-d. « l'information qui permet d'identifier une personne; ou au sujet de laquelle il y a tout lieu de croire qu'elle permet d'identifier une personne » (§160.103).

Dans la même veine, la *Data Protection Act* du Royaume-Uni s'applique à [traduction] toutes les « données personnelles », « données qui sont liées à une personne vivante qui peut être identifiée – a) à partir de ces données ou b) à partir de ces données ou d'autres informations qui sont en la possession ou sont susceptibles d'être en la possession du contrôleur des données », la personne légalement responsable de déterminer les fins pour lesquelles et la manière selon laquelle les données sont traitées (§I.1- (1))[15].

Si les données ne sont pas « identifiables », elles ne sont pas « personnelles » et il se peut que divers droits et obligations visant les données personnelles ne s'appliquent pas dans les circonstances. Les droits généraux à la confidentialité des renseignements et les obligations professionnelles de confidentialité médicale, par exemple, ne s'appliquent habituellement qu'aux données « concernant » des personnes réelles.

> **Administrative or demographic tags** (name, Social Security number, email address, hospital name, Zip code...)
>
> **Overt descriptors** (gender, eye color, height, blood type, scars, asthma...)
>
> **Indirect clues** (medication use, number of children, spouse's occupation, circumstances of emergency-room admission...).

Whether particular bits of data alone or in combination should be considered sufficient to identify a person is a matter of judgment. Much may depend on whether partial identifiers can be linked with identified or identifiable data in public or other databases.

The HIPAA Privacy Rule illustrates the practical challenges. The Rule provides that for data to be considered adequately de-identified and therefore not subject to its provisions, a number of listed descriptors must be absent. (See Figure 2, at the end of this chapter, known to aficionados as "the HIPAA List.").

The List comprises identifiers that are linked fairly directly somewhere to name-andaddress; it does not include all prime descriptors of persons. For example, there is no element on the list for health, illness, or disability characteristics, even those that may be evident to simple perception such as hearing impairment, palsy, albinism, limp, or wheelchair dependency. Presumably the assumption is that these will be caught by the "any other" element (R), even though this relegates a lot to judgment and the qualifier "unique" is subject to interpretation. One may wonder whether (R) covers, for example, an International Classification of Diseases code ("ICD-10 L40" = psoriasis vulgaris).

Knowing a few of the elements on the List may or may not allow identification, and even knowing a person-unique fact such as Social Security number allows identification only if it can be traced through some other source, such as a Social Security look-up database. It is obvious that some identifying elements are pretty weak in evidentiary power, at least if they aren't linked with other data. But some others, such as birthdate, are stronger.

A word about Limited Data Sets, since they are one of HIPAA's concessions to research and are being used in some genomic projects. For

La majorité des données sur la santé sont initialement recueillies sous la forme de données identifiées, que ce soit aux fins des soins de santé, de la santé publique ou de la recherche. On peut dépersonnaliser les données de diverses manières et à divers degrés, de manière irréversible ou réversible. La désidentification est une stratégie cruciale dans le domaine de la recherche.

Identificateurs

Pour tout ensemble de données concernant des personnes, il est possible de distinguer trois sortes de facteurs d'identification – couramment, quoiqu'un peu trop simplement, appelés des « identificateurs » :

> **Descripteurs administratifs ou démographiques** (nom, numéro de sécurité sociale, adresse électronique, nom de l'hôpital, code postal…).
>
> **Descripteurs explicites** (sexe, couleur des yeux, taille, groupe sanguin, cicatrices, asthme…).
>
> **Indices indirects** (utilisation de médicaments, nombre d'enfants, emploi du conjoint, circonstances de l'admission aux urgences…).

La question de savoir si des éléments de données particuliers, pris isolément ou associés, peuvent être jugés suffisants pour identifier une personne est affaire de jugement. Cela peut dépendre dans une large mesure de la capacité de coupler des identificateurs partiels avec des données identifiées ou identifiables dans des bases de données publiques ou d'autres bases de données.

Le *Privacy Rule* de la HIPAA illustre les défis pratiques que cela pose. Le règlement prévoit qu'un certain nombre de descripteurs expressément mentionnés doivent être absents pour que l'on puisse considérer que les données ont été dépersonnalisées de manière satisfaisante et que, par conséquent, elles ne sont pas visées par les dispositions du règlement. (Voir à la figure 2, à la fin du présent chapitre, la « liste de la HIPAA ».)

La liste comprend les identificateurs qui sont reliés assez directement quelque part au nom et à

research, the Privacy Rule allows data custodians to release Limited Data Sets, data from which many but not all core identifiers have been stripped (§164.514(b)(5)(e)(2)). Names, electronic communication addresses, and biometric identifiers must not be present, for example, but gender, birthdate, treatment dates, cities, states, Zip codes, and some other potentially identifying clues can remain, as well as substantive health information. When applying to a data custodian to use a Limited Data Set, researchers must specify which data they want and the intended uses, name who will be using the data, commit to enforcing safeguards, and promise that they will not attempt to identify the data-subjects or contact them.

Caution regarding the identifiability of Limited Data Sets is expressed in policies such as that of the Centers for Medicare & Medicaid Services:[16]

> Limited Data Sets (LDS) contain beneficiary level health information but exclude specified direct identifiers as outlined in the Privacy Rule. LDS are considered identifiable even without the specified direct identifiers. Because the information is considered identifiable, it remains subject to the Privacy Act of 1974 as well. These data are identifiable because of the potential for identifying a beneficiary due to technology, particularly in linking and re-identifying data files.

Because genomic data have exquisitely fine-grained informational structure, describe fundamental constituents of the person's body, don't change during the lifetime, and can be used for matching and possibly for profiling, surely they must be treated as strong identifiers. This has implications for the safeguarding of genomic data and the ethico-legal standards that govern that safeguarding. For present purposes it is important to contemplate whether now or in the future the HIPAA identifier elements (P), "Biometric identifiers," or (R), "Any other unique identifying number, characteristic, or code," should be interpreted as including genetic or genomic information.

This chapter has used the HIPAA Privacy Rule to illustrate the points, but similar issues arise with all privacy protection regimes.

l'adresse; elle ne comprend pas tous les descripteurs primaires d'une personne. Ainsi, aucun élément de la liste ne concerne les caractéristiques liées à la santé, à la maladie ou à l'incapacité, même celles qui pourraient être facilement observables, telles que le déficit auditif, la paralysie, l'albinisme, la boiterie ou l'utilisation d'un fauteuil roulant. On présume sans doute que ces caractéristiques entreront dans la catégorie « tout autre » élément même si cela a pour effet de laisser beaucoup au jugement et si le qualificatif « unique » est sujet à interprétation. On est en droit de se demander, par exemple, si le paragraphe (R) s'applique, par exemple, à un code de la *Classification internationale des maladies* (CIM‑10, L40 = psoriasis).

La connaissance de quelques éléments de la liste peut ou non permettre l'identification, et même la connaissance d'un renseignement propre à une seule personne, comme le numéro de sécurité sociale, ne rend possible l'identification que si l'on peut effectuer une recherche dans une autre source, par exemple la base de données de la Sécurité sociale. Manifestement, certains éléments d'identification sont assez faibles sur le plan probatoire, du moins s'ils ne sont pas couplés à d'autres données. Toutefois, d'autres éléments, comme la date de naissance, sont plus puissants.

Un mot concernant les « ensembles de données limités » (*Limited Data Sets*), puisqu'ils sont l'une des concessions de la HIPAA à la recherche et ne sont utilisés que dans certains projets en génomique. Aux fins de la recherche, le *Privacy Rule* autorise les gardiens de données à diffuser des ensembles de données limités, données desquelles un grand nombre d'identificateurs de base, mais pas tous, ont été supprimés (§164.514 (b)(5)(e)(2)). Ainsi, les noms, les adresses électroniques et les identificateurs biométriques ne doivent pas être présents, mais le sexe, la date de naissance, les dates de traitement, les villes, les États, les codes postaux et certains autres indices pouvant permettre l'identification peuvent demeurer, de même que des renseignements importants sur la santé. Lorsque les chercheurs demandent à un gardien de données l'autorisation d'utiliser un ensemble de données limité, ils doivent indiquer les données qu'ils veulent obtenir et les usages prévus, ainsi que le nom de la personne qui utilisera les données. Ils doivent aussi s'engager à appliquer des mesures de protection et promettre qu'ils ne tenteront pas

Terminology

A note regarding terms for various states of identifiability. This author prefers to speak of data as being, simply, either identified or identifiable, or key-coded, or non-identifiable.[17] Approximate synonyms used in various professional cultures are as follows. **(See Figure 1 in appendix 1.1)**

Use of "key-coded" avoids such awkward expressions as "pseudonymized" and helps the public understand the approach. "Encryption" is now taken in everyday speech to mean the scrambling of messages to keep them secret en route. "Coding" is universally used in the health sciences to refer to the classification of diseases, drugs, and procedures to standard categories. The central feature of a system that maintains the potential to reassociate substantive data with identifying data is the key: hence, key-code. **(See Figure 2 in appendix 2.2)**

3. Strategies for identifying non-identified genomic data

With detailed individual-level genomic data, such as SNPs and sequences, there are three basic and very different approaches to identifying the person to whom the data pertain, which will be discussed in turn:

- Matching genotype against reference genotype
- Linking genomic+associated data with other data
- Profiling from genomic characteristics.

These strategies can be used for socially desired purposes such as fighting crime and terrorism and identifying victims of accidents, disasters, epidemics, or war. They can also be used for nefarious purposes.

Matching genotype against reference genotype

Lin, Owen, and Altman have analyzed the chances of random matching of unidentified genotype data against reference-collection data and concluded that "specifying DNA sequence at only 30 to 80 statistically independent SNP positions will uniquely identify a single person." (They mean uniquely *match*, i.e. confirm

d'identifier les sujets sources de données ni de communiquer avec eux.

Des politiques comme celle des Centers for Medicare and Medicaid Services (CMS) renferment des mises en garde concernant l'identifiabilité des ensembles de données limités[16] :

> [traduction]
> Les ensembles de données (ED) limités renferment de l'information sur la santé à l'échelle du bénéficiaire, mais excluent les identificateurs directs mentionnés dans le *Privacy Rule*. Les ED limités sont considérés comme identifiables même en l'absence des identificateurs directs mentionnés. Vu que l'information est considérée comme identifiable, elle demeure également visée par la *Privacy Act* de 1974. Ces données sont identifiables en raison de la possibilité d'identifier un bénéficiaire grâce à la technologie, notamment en couplant et en réidentifiant les fichiers de données.

Étant donné que les données génomiques ont une structure informationnelle extrêmement fine, décrivent des constituants fondamentaux de l'organisme d'une personne, ne changent pas au cours de la vie et peuvent être utilisées à des fins d'appariement et, éventuellement, de profilage, il convient manifestement de les considérer comme des identificateurs puissants. Cela a des répercussions sur la protection des données génomiques et sur les normes éthico-juridiques qui régissent la protection. Aux fins du présent document, il importe de se demander s'il convient, aujourd'hui ou dans l'avenir, de considérer que les identificateurs de la HIPAA mentionnés aux paragraphes (P), « identificateurs biométriques » ou (R), « tout autre numéro d'identification, caractéristique ou code unique », incluent l'information génétique ou génomique.

Dans le présent chapitre, nous avons utilisé le *Privacy Rule* de la HIPAA pour illustrer notre propos, mais tous les autres régimes de protection de la confidentialité soulèvent des questions semblables.

Terminologie

Nous examinons ci-après les termes décrivant divers degrés d'identifiabilité. Dans le présent

that two samples come from the same identical person; whether it *identifies* anybody in the normal-talk sense depends on whether the reference data themselves are personally identified.) They go on and say that randomly changing 10% of SNPs, or binning using standard statistical techniques, do not change this conclusion.[18]

Identified forensic-purpose biospecimens from millions of people are held by the criminal justice system and the armed services. Most forensic matching focuses on a standard panel of short tandem repeat polymorphisms ("STRPs"). The identification efficacy is as high as it would be if SNPs were used.[19]

Literally countless biospecimens, and a growing number of genomic analyses, are held by medical, public health, and health research institutions. A practical source of reference for identifying victims and criminals is blood relatives' genotypes.[20]

Matching is possible to a high degree of certainty if very much of the genome is available. Its reliability is substantially higher than that with the matching of fingerprints or retinal scans.

Questions about matching include:

Q1: How "much" genome, i.e. how many megabases, SNPs, STRPs, traces, genes, or other amounts of information, is sufficient for identifying by matching? Surely a gene or two is not enough, but how much is?

(The answer seems to be, "It depends" − it depends on the density, or resolution, of mapping, the extent of genome covered, the rarity of variants, the degree of linkage disequilibrium, and other factors.)[21]

Q2: Is it possible to draft at least semi-quantitative criteria for setting thresholds of identifiability, for instance in technical guidance?

(The answer to the previous question implies that this will remain a matter of judgment. But maybe risk-analytical approaches can be devised.)

document, nous préférons considérer les données comme étant, simplement, identifiées ou identifiables, ou assorties d'un code d'identification ou non identifiables[17]. Voici divers synonymes utilisés selon la culture professionnelle : **(voir la Figure 1 dans l'appendice 1.1)**

L'utilisation du terme « assorti d'un code d'identification » permet d'éviter des expressions bizarres telles que « pseudonymisé » et aide le public à comprendre l'approche. On parle couramment de « cryptage » pour désigner l'embrouillage des données dans le but de préserver leur confidentialité durant leur transmission. Le terme « codage » est universellement utilisé en sciences de la santé pour désigner la classification des maladies, des médicaments et des interventions au moyen de catégories standard. L'élément central d'un système qui permet de conserver la possibilité de réassocier des données fondamentales avec des données d'identification est la clé, c. à-d. le code d'identification. **(voir la Figure 2 dans l'appendice 1.2)**

3. Stratégies d'identification des données génomiques non identifiées

Lorsque l'on dispose de données génomiques détaillées à l'échelle individuelle, telles que les polymorphismes d'un nucléotide simple (SNP) et les séquences, on peut avoir recours à trois approches fondamentales et très différentes pour identifier le sujet source des données. Nous examinerons l'une après l'autre ces trois approches ci-après :

- Appariement du génotype avec le génotype de référence
- Couplage des données génomiques + connexes avec d'autres données
- Profilage à partir des caractéristiques génomiques.

On peut utiliser ces stratégies aux fins socialement désirées (comme la lutte contre la criminalité ou le terrorisme ou l'identification des victimes d'accidents, de catastrophes, d'épidémies ou de guerre). On peut également les utiliser à mauvais escient.

Linking genotype+associated data with other data

A second route to identifying genotyped subjects is deduction by linking and matching genotype+phenotype (or +other data) with data in health, demographic, administrative, employment, criminal, military service, hazard exposure, disaster response, or other databases. Often the context as regards exposure, disease, or locale strongly suggests which external databases may yield useful information. If the data linked-to are overtly identified, the task is straightforward. If the data linked-to are not fully identified, inferential matching and narrowing-down may be possible. Statisticians possess an array of well tested techniques for identifying data-subjects from partial data, and an equally well tested array of techniques for obfuscating inferential identification.[22, 23]

There is no shortage of external sources of identified data. There are so many public and commercial databases about people's lives, especially in the U.S., that it requires databases of databases and enables a lucrative look-up/hunt-down industry.24 Just a few familiar examples of accessible data are those about birth, marriage, divorce, and death, home and business addresses and telephone numbers, voter registration, motor vehicle and boat registration, real estate ownership, police and court proceedings, organization membership, professional licensing, government employment, and in some states hospital discharge. Family pedigree databases can provide complementary information.

Beyond these databases, of course the health arena holds uncountably more confidential but conditionally accessible data ranging from medical care and payment records to perinatal genetic screening results, Guthrie cards, disease registries, and implant registries.

Obviously, barring the availability of a reference genotype collection, genotype+phenotype (or +other) data are much more vulnerable to being inferentially identified than genotype data alone are. Thus careful attention will have to be paid as health research moves, inevitably, toward linking genomic data with clinical and social data.

Appariement du génotype avec le génotype de référence

Lin, Owen et Altman ont analysé la probabilité de l'appariement aléatoire des données génotypiques non identifiées avec les données d'une collecte de référence et ont conclu que [traduction] « Le fait de préciser la séquence d'ADN d'aussi peu que 30 à 80 locus SNP indépendants peut permettre d'identifier une personne sans ambiguïté. » (Ils veulent dire ici « permettre un *appariement* unique, c. à-d. la confirmation que deux échantillons proviennent d'une même personne; la possibilité d'*identifier* une personne au sens habituel du terme dépend du fait que les données de référence elles-mêmes soient ou non personnellement identifiées.) Ils avancent en outre que le fait de modifier de manière aléatoire 10 % des SNP, ou la classification par fenêtre au moyen de méthodes statistiques standard, ne modifie pas cette conclusion[18].

Des échantillons biologiques identifiés à des fins judiciaires et prélevés chez des millions de personnes sont conservés par le système judiciaire et les forces armées. L'appariement à des fins judiciaires se concentre en général sur un panel standard de polymorphismes de microsatellites (STR). L'efficacité de l'identification est aussi élevée que si l'on utilisait les SNP[19].

Une quantité littéralement incommensurable d'échantillons biologiques et un nombre croissant d'analyses génomiques sont conservés par les établissements médicaux, les établissements de santé publique et les centres de recherche en santé. Le génotype sanguin des parents est une source commode d'information pour l'identification des victimes et des criminels[20].

L'appariement est possible avec un degré élevé de certitude si une quantité importante d'information génomique est disponible. Sa fiabilité sera beaucoup plus grande que celle obtenue au moyen des empreintes digitales ou des empreintes rétiniennes.

L'appariement soulève notamment deux questions :

Q1 : Quelle est la « quantité » d'information génomique (c. à-d. combien de millions de paires de bases, combien de SNP, de polymorphismes de STR, de tracés, de gènes ou d'autres informations)

Profiling from genomic characteristics

Increasingly now there is concern about whether a probabilistic profile of an individual can be inferred from genotype. This amounts to *describing* rather than actually *identifying*. Making such inferences depends on being able to interpret how particular genomic factors contribute to determining bodily characteristics or behavioral or disease likelihoods, and then developing a composite description of the person. Such a description can only be a probabilistic profile against which candidates can be screened, and any further narrowing-down depends on linking with other evidence.

As the population frequencies of SNPs and STRPs become better known, both kinds of markers are being used to construct profiles of ethnicity. Forensic approaches have tended to use STRPs, perhaps because STRP data are the principal sort held in police collections (for matching), but profiling based on them suffers from such limitations as the fact that STR loci mutate relatively rapidly.[25] As millions of SNPs become analyzed and the ancestry of the DNA sources is characterized, the patterns tend to suggest profiles.[26]

It is sometimes rumored that the FBI, CIA, and other agencies are developing systems for profiling suspects based on genomic data, and not just with reference to ethnicity. Apparently from time to time they have explored this avenue, but so far have not found it very productive.[27] They should be expected to pursue genomic profiling when the science has advanced sufficiently.

Again: Any inference derived by comparing an individual's SNP or STRP markers against the prevalence of those markers in culturally or geographically defined populations can at best yield only a likelihood ratio.[28, 29, 30]

Questions about profiling include:

Q3: How accurate can profiling be regarding possible ethnic/racial origins or appearance (acknowledging the definitional ambiguities)?

Q4: Which corporal features can be inferred now, apart from gender, the odds on blood type, skin pigmentation, and overt manifestations of Mendelian disorders?

suffisante pour procéder à une identification par appariement? Un gène ou deux ne suffit sans doute pas, mais combien en faut-il?

(La réponse semble être « cela dépend ». Cela dépend de la densité ou de la résolution de la cartographie, de l'étendue de l'information génomique examinée, de la rareté des variantes, du degré de déséquilibre de liaison et d'autres facteurs[21].)

Q2 : Est-il possible de déterminer au moins des critères semi-quantitatifs pour l'établissement de seuils d'identifiabilité, par exemple dans des directives techniques?

(La réponse à la question précédente sous-entend que cela demeurera une affaire de jugement. Mais il existe la possibilité de concevoir des approches analytiques du risque.)

Croisement des données génotypiques + connexes avec d'autres données

Une autre méthode d'identification des sujets génotypés est la déduction par le couplage et l'appariement des données génotypiques+phénotypiques (ou + d'autres données) avec des données qui sont contenues dans des bases de données démographiques, administratives, des bases de données sur la santé, l'emploi, la criminalité, le service militaire, l'exposition au risque, les interventions en cas de catastrophe ou d'autres bases de données. Souvent, le contexte entourant l'exposition, la maladie ou le lieu aide grandement à déterminer quelles bases de données externes il convient d'utiliser. Si les données avec lesquelles s'effectue le couplage sont clairement identifiées, la tâche est alors relativement simple. Cependant, si ces données ne sont pas complètement identifiées, il est parfois possible de procéder à un appariement inférentiel puis de circonscrire la recherche.

Les statisticiens disposent d'un éventail de techniques bien éprouvées pour l'identification des sujets sources des données à partir de données partielles, ainsi que d'une gamme de techniques tout aussi bien éprouvées pour rendre

Which might be expected to be deducible before long − height, shoulder width, or other aspects of skeletal build? Hair color or texture? Eye color? Eye shape or other facial features? Cranial, dental? Others?

Q5: Which behavioral or disease attributes − likelihood of being depressive, schizophrenic, alcoholic, violent? Diabetic, hypertensive? Others?

Q6: Shouldn't we assume that in 5−10 years many attributes will be profilable?

Q7: With profiling, "how much" genome has to be known before the data in themselves are usefully descriptive? Again, is it possible to set semiquantitative thresholds of identifiability?

(The answer will be very different from that to the same question regarding matching.)

4. Strategies for de-identifying genomic data

For reducing identifiability of genomic (perhaps +other) data before releasing them for research, there are three sorts of technical options, which will be discussed in turn:

- Limiting the proportion of genome released
- Statistically degrading the data before releasing
- Sequestering identifiers via key-coding.

Limiting the proportion of genome released

The first option is to publish only limited segments of genomes, such as sequence traces or only a few variants, along with minimum necessary phenotypic or other data. This requires judgment as to how much to release, which will depend on what genomic region is involved and the circumstances of data or biospecimen collection (for instance, whether the data are openly known to be about people having a particular disease, or who live in a certain region), and be difficult to generalize. It may deny important data to

difficile l'identification inférentielle[22,23].

Les sources externes de données identifiées ne manquent pas. Il existe un si grand nombre de bases de données publiques et privées sur la vie des gens, en particulier aux États-Unis, qu'il faut avoir recours à des « bases de données de bases de données », ce qui ouvre la voie à une lucrative industrie de recherche de renseignements[24]. Voici quelques exemples bien connus de données faciles d'accès parmi bien d'autres : données sur les naissances, mariages, divorces et décès, adresses et numéros de téléphone à la maison et au travail, inscriptions électorales, immatriculation des véhicules automobiles et des bateaux, possession de biens immobiliers, interventions policières et poursuites en justice, appartenance à des organisations, inscriptions au tableau d'un organisme de réglementation professionnelle, emploi au sein de la fonction publique et, dans certains États, congés des hôpitaux. Les bases de données généalogiques peuvent fournir des renseignements complémentaires.

Outre ces bases de données, le secteur de la santé dispose, bien sûr, d'une quantité phénoménale de données plus confidentielles, mais accessibles sous condition : données sur les soins médicaux, registres des paiements, résultats de tests de dépistage génétique périnatal, fiches de Guthrie, registres des maladies et registres des greffes.

Manifestement, si l'on excepte la disponibilité d'une collecte de génotypes de référence, les données génotypiques+phénotypiques (ou +d'autres données) se prêtent bien davantage à une identification référentielle que les données génotypiques seules. Au fil de l'évolution inévitable des recherches en santé, il faudra donc être très attentif au risque que des données génomiques soient couplées avec des données cliniques et sociales.

Profilage à partir des caractéristiques génomiques

On craint de plus en plus, aujourd'hui, qu'il soit possible de déduire le profil probabiliste d'une personne à partir du génotype. Cela constitue une *description* plutôt qu'une *identification* véritable. La possibilité de faire ce genre de déduction dépend de la capacité d'interpréter comment des facteurs génomiques particuliers contribuent à

researchers, including data about regions of the genome that they can't know whether they need to know.

In practice many disease-specific projects do limit the portion of genome they release, but it is not clear that they use any criteria other than "no more than necessary." How such a limitation might apply with whole genome association studies, however, is unclear [at least to this author]. Certainly, precautions can be taken, such as releasing sequence traces in such a separated manner that it isn't possible for an external analyst to reconstruct which traces pertain to a single individual's DNA.

The pivotal issue here is the same as was raised in the previous chapter: how to know "how much" genome is too much to release.

Statistically degrading the data before releasing

The second option is to degrade data before posting, such as reducing precision by lumping G and A as purines, and C and T as pyrimidines; or "fuzzing" data by adding statistical noise, i.e. spurious but not dissimilar data, to data-sets; or randomly altering or exchanging a small percentage of SNPs; or micro-aggregating or "binning" small subsets of data in various ways to reduce granularity. These are all standard statistical disclosure-reduction techniques, although they have to be carefully adapted for genomic data.[31, 32]

Data transformed in such ways may meet the needs of some query systems in which researchers pose questions to external databases. And they are fine for some higher level population surveys. But at the sequence level the human genome data-tape comprises some 3,000,000,000 data-cells − arrayed in linear order, though segmented into chromosomes − and the sensitivity of the human organism to variance is such that the occurrence of a T instead of a C in one data-cell can mean the difference between disease and health. So for many lines of genomic research, degrading of data simply degrades usefulness.

A pragmatic question on which it is very important to have calibration from genomicists is:

déterminer la probabilité qu'une personne présente des caractéristiques physiques, des comportements ou des maladies, puis à développer une description composite de cette personne. Une telle description équivaut tout au plus à établir un profil probabiliste en fonction duquel les candidats peuvent être triés, et la possibilité de circonscrire davantage le tri dépend de la possibilité d'effectuer un couplage avec d'autres données.

À mesure que l'on connaît davantage la fréquence des SNP et des polymorphismes de STR au sein d'une population, il devient possible d'utiliser d'autres sortes de marqueurs pour établir des profils d'appartenance ethnique. Les approches judiciaires ont plus souvent utilisé les polymorphismes de STR, probablement parce que les données à ce sujet sont celles le plus souvent conservées dans les bases de données des services de police (à des fins d'appariement). Toutefois, l'appariement fondé sur ce type de données présente des limites, notamment le fait que les locus des STR connaissent des mutations relativement rapides[25]. Au fur et à mesure que sont analysées des millions de SNP et que l'origine des sources de l'ADN est caractérisée, les modèles qui se dégagent finissent en général par évoquer des profils[26].

Selon certaines rumeurs, le FBI, la CIA et d'autres agences mettent en place des systèmes de profilage des suspects fondés sur les données génomiques, systèmes qui ne visent pas seulement la détermination de l'ethnicité. Apparemment, cette avenue a déjà été explorée à quelques reprises, mais on ne l'a pas jugée suffisamment prometteuse jusqu'ici[27]. On s'attend à ce que ces organismes effectuent un profilage génomique dès que la science aura fait des progrès suffisants dans ce domaine.

Encore une fois, il convient de rappeler que toute déduction fondée sur une comparaison entre les marqueurs SNP ou STR d'une personne et la prévalence de ces marqueurs au sein de populations définies sur les plans culturel ou géographique peut, dans le meilleur des cas, permettre d'établir uniquement un rapport de vraisemblance[28, 29, 30].

Le profilage soulève notamment les questions suivantes :

Q3 : Quel est le degré d'exactitude du

Q8: How serious an impediment to research is masking, binning, perturbing, or otherwise degrading genomic data? How does the answer vary with different techniques, and with different lines of research?

Sequestering identifiers via key-coding

The method most widely used in health research for de-identifying data is key-coding (reversibly de-identifying), in which potentially identifying data are separated from the substantive data, such as health data, but a link is maintained by assigning an arbitrary code number to each part of the data–identifier pair before they are separated. Held securely and separately, the key allows reassociating the substantive data with the identifiers if that is ever necessary. The key and the responsibility for its use can be delegated to a trusted party, and use of the key can be guided by agreed criteria and subjected to oversight.

The scientific and ethical advantages of reversible de-identification are widely appreciated and need not be reviewed here. Key-coding can be used among multiple databases and with biospecimens, and it can keep elements, such as clinical data and biospecimens, rossreferenced with each other even if the links to the data-subjects are irreversibly severed. For high-sensitivity data, the codes can be further encoded. Elaborate key-coding systems and identifiability vocabulary are being developed for pharmacogenetic data submitted in regulation.[33, 34]

Surely what is important in any instance is not whether a link of some sort exists somewhere, but *whether the identifiers can be known to the researchers* who study the substantive data. A carefully constructed key system can provide reliable safeguards. Furthermore, data-use agreements almost always require that data recipients promise not to attempt to re-identify the data-subjects, which obviously forbids abusing the key system.

The HHS Office for Human Research Protection (OHRP) has issued helpful guidance on these issues. The document should be consulted for details, but a main point is this:[35]

> OHRP considers private information or specimens not to be individually identifiable when they cannot be linked to specific

profilage en ce qui concerne la détermination des origines ethniques/ raciales possibles ou de l'apparence (compte tenu des ambiguïtés définitionnelles)?

Q4 : Quelles caractéristiques corporelles peut-on aujourd'hui déduire, mis à part le sexe, la probabilité du groupe sanguin, la pigmentation de la peau et les manifestations patentes de troubles mendéliens? Quelles sont celles qui pourront vraisemblablement être déduites avant longtemps – la taille, la largeur des épaules ou d'autres aspects de la charpente du squelette? La couleur ou la texture des cheveux? La couleur des yeux? La forme des yeux ou d'autres traits du visage? Des caractéristiques du crâne, de la dentition? D'autres encore?

Q5 : Quels sont les traits liés à des comportements ou à des maladies – probabilité de souffrir de dépression, de schizophrénie, d'alcoolisme, d'avoir un comportement violent? Risque de souffrir de diabète, d'hypertension? Autres?

Q6 : Ne pourrait-on pas présumer que, d'ici 5 à 10 ans, il sera possible de procéder au profilage de ces caractéristiques?

Q7 : En ce qui concerne le profilage, quelle « quantité » d'information génomique faut-il connaître avant que les données en elles-mêmes soient utiles sur le plan descriptif? Encore une fois, est-il possible d'établir des seuils d'identifiabilité?

(La réponse à cette question sera très différente de celle donnée à la même question concernant l'appariement.)

4. Stratégies de désidentification de l'information génomique

Pour restreindre l'identifiabilité de l'information génomique (et peut-être d'autres données) avant qu'elles ne soient transmises à des chercheurs, nous disposons de trois options techniques, que nous aborderons à tour de rôle :

- la limitation de la proportion de segments de

individuals by the investigator(s) either directly or indirectly through coding systems. For example, OHRP does not consider research involving *only* coded private information or specimens to involve human subjects as defined under 45 CFR 46.102(f) [of the Common Rule] if the following conditions are both met:

> (1) the private information or specimens were not collected specifically for the currently proposed research project through an interaction or intervention with living individuals; and
>
> (2) the investigator(s) cannot readily ascertain the identity of the individual(s) to whom the coded private information or specimens pertain because, for example:
>
> > (a) the key to decipher the code is destroyed before the research begins;
> >
> > (b) the investigators and the holder of the key enter into an agreement prohibiting the release of the key to the investigators under any circumstances, until the individuals are deceased...;
> >
> > (c) there are IRB-approved written policies and operating procedures for a repository or data management center that prohibit the release of the key to the investigators under any circumstances, until the individuals are deceased; or
> >
> > (d) there are other legal requirements prohibiting the release of the key to the investigators, until the individuals are deceased.

NIH and other investigators have had considerable experience with such key-coding, but the craft continues to deserve refinement. The problem is not one of technology development − many identifiability-protecting encryption programs exist, and ever more sophisticated ones are being developed (for protecting confidentiality while mining data in large sets electronic medical records, for example) − but one of discipline in day-to-day practice.

Key-coding can effectively sequester obvious identifiers. But before data are released, various indirect identifying bits (such as free-text narrative or mentions of family members) may have to be

génome diffusés;

- l'altération statistique des données avant leur communication;
- le verrouillage des identificateurs par l'attribution de codes d'identification.

Limitation de la proportion de segments de génome diffusés

La première option consiste à ne publier que des segments limités de génome, tels que des tracés de séquences, ou uniquement un petit nombre de variants accompagnés du minimum de données phénotypiques (ou d'autres données) nécessaires. Cette option exige un certain jugement quant à la quantité d'information génomique pouvant être communiquée, selon la région génomique à l'étude et les conditions de collecte de données ou d'échantillons biologiques (par exemple, si les données sont manifestement connues comme étant celles de personnes souffrant d'une maladie particulière ou vivant dans une certaine région). Cette méthode peut en outre s'avérer difficile à généraliser. Elle pourrait par ailleurs priver les chercheurs de données importantes, comme les données sur les régions génomiques auxquelles ils n'ont pas accès, s'ils en ont besoin.

Dans les faits, bon nombre de projets portant sur des maladies particulières limitent la part d'information génomique publiée, mais nous ne sommes pas certains qu'ils fassent appel à un critère quelconque, autre que celui du « minimum nécessaire ». La façon dont ce critère pourrait être appliqué aux études d'association à l'échelle du génome n'est toutefois pas claire [du moins pour l'auteur]. Certes, des précautions peuvent être prises, comme la publication de tracés de séquences segmentés de telle sorte qu'il soit impossible pour un analyste de l'extérieur de les reconstituer pour déterminer ceux qui appartiennent à l'ADN d'un individu donné.

La question fondamentale ici est la même que celle soulevée au chapitre précédent : comment faire pour déterminer quelle est la « quantité » maximale d'information génomique pouvant être diffusée?

stripped off, or if these data are absolutely necessary for the research, protected by a non-disclosure agreement. Artificial-intelligence programs can help, for example by screening for proper nouns or especially sensitive words or phrases, but the exercising of human judgment cannot be avoided.

Obviously the risk questions raised by this chapter are the same as those of the previous chapter, although coming from the opposite direction − essentially, "How much" genome is too much to hang out in public?

Related to de-identification are several questions of responsibility:

Q9: Who should be responsible for de-identifying data before providing them, via whatever route, for research? Are physicians and principal investigators adequately prepared to do this? What roles should IRBs play with respect to disclosure control? 36

Q10: For data or biospecimens provided to research resource platforms, such as GAIN or UK Biobank, should the platforms themselves conduct any identifiability or disclosure review before releasing the data onward, via whatever route, for research?

5. Controlled release

The alternative to open publication is release of data to researchers under agreements that, inter alia, protect privacy and confidentiality. Legally such agreements amount to contractual undertakings. In many instances they are also public promises, as when they pertain to data obtained from government institutions or sources supported by government funds. Often for legal clarity the definitions or conditions in agreements refer to definitions or requirements of the Common Rule, the HIPAA Privacy Rule, the Privacy Act, or other statutes, regulations, or guidance.

Many projects release data through two precautionary stages: first reducing identifiability to a reasonable extent by suppressing overt identifiers, broadening data (such as rounding birthdate to year of birth or age-range), and so on, and performing at least informal disclosure review;

Altération statistique des données avant leur communication

La deuxième option consiste à altérer les données par différentes techniques, avant qu'elles ne soient publiées : en réduisant leur précision (p. ex. : en regroupant la guanine et l'adénine dans la catégorie des purines, puis la cytosine et la thymine, dans celle des pyrimidines), en « brouillant » les données (p. ex. en ajoutant aux sous-ensembles de données un bruit de fond statistique, c'est-à-dire des données parasites, mais non dissemblables), en altérant ou en interchangeant de façon aléatoire un petit pourcentage des SNP, en créant des micro-agrégats ou en « stockant » de différentes façons de petits sous-ensembles de données afin de réduire la granularité. Ce sont toutes des techniques statistiques normalisées permettant de limiter la divulgation de données confidentielles, mais il est nécessaire qu'elles soient rigoureusement adaptées à l'information génomique[31, 32].

Les données ainsi transformées peuvent satisfaire aux besoins de certains systèmes d'interrogation utilisés par les chercheurs pour interroger des bases de données externes. Elles conviennent aussi pour certaines enquêtes sur la population à plus grande échelle. Cependant, au niveau séquentiel, la bande de données du génome humain comprend quelque trois milliards de cellules de données — ordonnées de façon linéaire, bien que segmentées en chromosomes —, et la sensibilité de l'organisme humain aux variations est telle, que la présence d'un T plutôt que d'un C dans une cellule de données peut faire toute la différence entre la maladie et la santé. Aussi, dans bien des secteurs de la recherche en génomique, l'altération des données est tout simplement synonyme d'appauvrissement de leur utilité.

Une question pragmatique à laquelle il est très important que les spécialistes en génomique répondent par une valeur quantitative est la suivante :

Q8 : Dans quelle mesure un obstacle à la recherche peut-il masquer, écarter, perturber ou sinon altérer l'information génomique? Dans quelle mesure la réponse variera-t-elle si des techniques différentes sont utilisées? Et si les secteurs de recherche ne sont pas les

and then providing access under a controlled-release agreement.

Terms of agreements

The terms commonly addressed by controlled-release agreements are shown in Figure 3 (at the end of this chapter), meant here to telegraph the considerations. Probably no agreement anywhere includes all of these terms, and some agreements may incorporate additional ones, but the table is a basic menu. Among the terms most relevant for protection of identifiability are the following.

Consent relating to identifiability. May address the honoring of commitments to protect identifiability in various ways. May set limitations on purpose, allowed users, or other aspects of data use, either for the whole data-set or for particular data-subjects, which may reflect perceptions of trustworthiness. May address publication, and either promise that identities will be thoroughly obscured or warn that some non-negligible identifiability risk may accompany publication.

Confidentiality protection. Must always include physical, organizational, and IT security. May make reference to compliance with International Standards Organization, HIPAA, or other security standards. Usually specifies who, if anyone, will be responsible for de-identifying data and the procedures and criteria for doing this. May cover processes of key-coding, safeguarding of the key, and use of the key. Almost always states that researchers will make no attempt to re-identify or recontact the data-subjects.

Limiting of onward transfer. Restricts transfer (or access, which amounts to the same thing) and extends the chain of confidentiality and the accompanying obligations.

Linking. May be discussed if, inter alia, linking is contemplated that might have the effect of increasing identifiability.

Publication, release, and returning of findings. Specifies whether data must be fed back to the original data resource, and whether publication of detailed findings, such as sequences, may be fully open, as on a publicly accessible database, or must be by further controlled release. Usually requires that the identifiability of any derived data must be protected to the same extent as that of the data being provided.

mêmes?

Verrouillage des identificateurs par l'attribution de codes d'identification

Dans le domaine de la recherche en santé, l'attribution de codes d'identification est la méthode de désidentification (réversible) des données la plus répandue, dans laquelle les données pouvant servir d'identificateurs potentiels sont séparées des données fondamentales, telles que celles sur la santé. Avant d'être séparées, un numéro de code arbitraire (clé) est assigné à chaque constituant de cette paire de données, de manière à maintenir le lien entre les données d'identification et les données fondamentales correspondantes. Cette clé, conservée en sécurité dans un lieu distinct, permet, au besoin, de réassocier les données fondamentales avec les données d'identification. La responsabilité de cette clé et de son utilisation peut être confiée à un tiers et soumise à certains critères convenus, dont le respect est assuré par un mécanisme de surveillance.

Les avantages scientifiques et éthiques de la désidentification réversible sont largement reconnus; il n'est donc pas nécessaire d'en reparler ici. L'attribution de codes d'identification peut être utilisée dans quantité de bases de données et avec une multitude d'échantillons biologiques. Elle permet de préserver la concordance entre des éléments tels des données et des échantillons biologiques cliniques, même si les liens vers les données confidentielles sur les sujets sont rigoureusement dépersonnalisés. Des données de la plus grande confidentialité peuvent être accompagnées de codes d'identification plus complexes. Des systèmes perfectionnés à clé de cryptage et une terminologie de l'identifiabilité pour les renseignements pharmacogénétiques qui y sont soumis en vertu de la réglementation sont en cours d'élaboration[33, 34].

Il est certain que l'important, en toutes circonstances, n'est pas de savoir si un lien quelconque existe quelque part, mais si les identificateurs peuvent être connus des chercheurs qui étudient les données fondamentales. Un système à clé de cryptage peut fournir un dispositif de protection fiable. De plus, le recours à des ententes en matière d'utilisation des données requiert presque toujours des destinataires qu'ils s'engagent à ne pas tenter

IRB or other ethics approval. Deferred to for oversight regarding conformance to the Common Rule and other obligations. An issue may be at what stage(s) IRB review should be carried out.

Arrangements for controlled release

Many details of practicality and governance have to be attended to, but data-providers and data-users are generally familiar with controlled-release arrangements. These three examples provide a few specifics.

Wellcome Trust Case Control Consortium. The Consortium, which involves a number of university units in the U.K., is genotyping de-identified DNA samples from thousands of people with known chronic diseases, and controls. Access to cleaned, raw, and summary data is decided by a Consortium Data Access Committee and is subject to a Data Access Agreement.[37]

Framingham Heart Study. Framingham releases DNA and data via evaluation by two committees and a Data and Materials Distribution Agreement.[38]

Genetic Association Information Network (GAIN). Because it will genotype DNA submitted by a network of contributing academic disease-specific studies having diverse auspices, consents, and IRB stances, and because it is concerned about identifiability of the extensive data it expects to generate, the GAIN project recently changed its policy from one envisioning fairly open publication to one using controlled release. De-identification will primarily be the responsibility of the data providers. Access will be controlled by a Data Access Committee and will be subject to Data Access Certification.[39]

The GAIN example illustrates the importance for the genomic research community of exploring criteria for deciding whether particular sorts and "amounts" of data can be posted publicly or must be managed by controlled access.

Alternatively to transferring full data-sets to external researchers, more restricted channels can be employed, such as:

- on-site data enclaves (or Research Data Centers, as several Federal ones are called) to which researchers come and perform studies on a database in a secure, dedicated,

de réidentifier les données confidentielles, ce qui signifie manifestement que toute exploitation abusive du système à clé de cryptage est interdite.

Le Office for Human Research Protection (OHRP) du Department of Health and Human Services (HHS) a publié des lignes directrices très utiles sur ces questions. Bien que le document mérite d'être consulté pour plus de détails, il convient d'en faire ressortir le point suivant[35] :

> [traduction]
> L'OHRP considère comme étant non individuellement identifiable tout renseignement ou échantillon confidentiel qu'un système à clé de cryptage interdit aux chercheurs d'associer directement ou indirectement à un individu en particulier. À titre d'exemple, l'OHRP ne considère pas comme étant fondées sur des sujets humains des recherches faisant uniquement appel à des renseignements ou à des échantillons confidentiels codés, conformément à la définition de l'alinéa 46.102 f) du titre 45 du CFR [de la « règle commune »], si les dispositions suivantes du Code of Federal Regulations (CFR) des États-Unis sont toutes deux satisfaites :
>
> (1) les renseignements ou les échantillons confidentiels n'ont pas été spécifiquement recueillis pour le projet de recherche actuellement proposé au cours d'interactions ou d'interventions auprès de personnes vivantes; et
>
> (2) les chercheurs ne peuvent pas déterminer facilement l'identité des individus auxquels se rapporte un renseignement ou un échantillon confidentiel codé, par exemple :
>
> (a) la clé pour décrypter le code est détruite avant que la recherche ne débute;
>
> (b) les chercheurs et le détenteur de la clé concluent une entente interdisant aux chercheurs de divulguer la clé tant que les personnes concernées ne sont pas décédées;
>
> (c) des politiques et des procédures consignées par écrit, portant sur l'exploitation d'une banque de données ou d'un centre de gestion de données et

monitored server

- remote-query systems, in which researchers interrogate databases and obtain responses, possibly veiled in some ways, via secure telecommunications

- service analyses that analyze data or biospecimens according to agreed methodology and provide the results to the commissioning researchers.

Ways must be devised to make controlled release practical, binding, and palatable − conditions that are not procedurally onerous but that at the same time secure genuine, formal, enforceable commitments. Perhaps for some group of projects a general data-use license can be worked out, for instance, through which researchers, with their institutions, agree to terms and gain entrée to a large suite of data and/or biospecimens.

Q11: Is there any appeal in exploring broad data-use licenses for access to centrally held genomic data? Has there been any relevant experience with such a scheme?

(See Figure 3 in appendix 2.3)

6. Identifiability risks, overall

All of this must be examined from risk perspectives − risks to data-subjects, risks to data stewards, risks to researchers and their institutions, even risks to the genomic research enterprise. Concern must be about whether data can be used to (a) deduce the identity of data-subjects or (b) deduce facts about data-subjects, and whether in either case this can lead to harm.

Sizing-up risks of any kind involves two activities. First, "risk assessment" estimates the probability of undesired events compounded by the severity of the likely consequences. Then, "risk appraisal" weighs the risks in perspective of personal or societal values. Appraisal can be cast as willingness to invest in reducing the risk by reducing the odds or the stakes or both. A broader appraisal can weigh the risks against benefits gained in the risktaking and consider whether the risks are acceptable. This is the way people think about most situations in life, whether they realize it or not.[41]

approuvées par un CEI, stipulent que les chercheurs ne doivent en aucune circonstance divulguer la clé tant que les personnes concernées ne sont pas décédées; ou

(d) il existe d'autres dispositions légales interdisant de divulguer la clé aux chercheurs tant que les personnes concernées ne sont pas décédées.

Les National Institutes of Health et d'autres chercheurs ont acquis une grande expérience dans le domaine de l'attribution de codes d'identification, mais ces techniques ont encore besoin d'être perfectionnées. Le problème n'est pas lié aux progrès technologiques, car il existe un grand nombre de programmes de cryptage permettant de protéger l'identifiabilité. En outre, des programmes encore plus perfectionnés sont en cours d'élaboration (par exemple, pour protéger la confidentialité tout en explorant de vastes ensembles de données de dossiers de santé électroniques). Le problème est plutôt lié à la discipline que cela requiert dans la pratique quotidienne.

L'utilisation de codes d'identification permet en effet de verrouiller tout identificateur évident. Cependant, cela peut nécessiter que les segments d'information servant d'identificateurs indirects (comme les exposés de faits en texte libre ou les remarques concernant des membres de la famille) soient éliminés avant que les données ne soient publiées, ou si ces données sont absolument nécessaires pour la recherche, qu'elles soient protégées par une entente de non-divulgation. Les programmes d'intelligence artificielle peuvent aider, par exemple en contrôlant les noms propres ou encore les mots ou les expressions particulièrement révélatrices, mais le jugement humain reste essentiel dans tous les cas.

De toute évidence, les questions de risque soulevées dans ce chapitre sont les mêmes que celles du chapitre précédent, même si la perspective est diamétralement opposée. Essentiellement, elles convergent toutes vers la question suivante : quelle est la « quantité » maximale d'information génomique pouvant être publiée?

La désidentification soulève plusieurs questions sur le plan de la responsabilité :

Genomic disclosure-risk-assessment must take account of such factors as the extent of genome covered; the density, or resolution, of mapping; the rarity of variants (because rarity increases identifiability); the degree of linkage disequilibrium; and the specificity with which gene effects are known. A special consideration with genomics is the disclosure risk for blood-relatives of people whose genome is studied, which has implications for consent and for safeguards.

Safeguards can't be discussed here except to state the sermonic point that an array of physical, procedural, cybersecurity, training, and legal safeguards must be in place against both accidental release and intrusive access. Among other reasons*, safeguards are what justify asking for broad consent.*

(A large topic that must be left to other forums is the need for genetic anti-discrimination laws, which tend to focus more on harmful consequences than on processes.)

What are the threats? We know that computerized systems can be broken into, data obtained by subterfuge, laptops stolen, and biospecimens transferred improperly. To the present there have been remarkably few proven abuses of medical data, much less health research data. But with the coming of electronic medical records, increased linking of databases, and so on − and given the vague foreboding that many people feel about anything "genomic" − public concerns are intensifying. As was mentioned at the outset of this document, the abuses that can be imagined range from embarrassment, blackmail, fraud, and group stigmatization, to negative discrimination for health or life insurance, employment, promotion, mortgages, or loans. Another possible abuse, depending on point of view, is unconsented parentage testing.

Should we expect accidental releases, hacking, and attempts at abuse? Certainly. Detailed speculation is fruitless, although risk-anticipation exercises can help identify vulnerabilities and suggest defenses. It must be assumed that some threats are real possibilities, and that some can have serious consequences.

A plea for risk aversion. Surely it will be important not to expose "too much" of people's genomes in the coming years, only to regret it in the future when the analytic technologies become more

Q9 : À qui incombe-t-il de dépersonnaliser l'information, et par quel moyen, avant qu'elle ne soit transmise aux chercheurs? Les médecins et les chercheurs principaux y sont-ils adéquatement préparés? Quels rôles les CEI devraient-ils jouer en matière de contrôle des données divulguées[36]?

Q10 : En ce qui concerne les données ou les échantillons biologiques versés dans des plateformes de ressources de la recherche comme le Genetic Association Information Network (GAIN) ou la UK Biobank, devrait-on obliger ces plateformes, et par quel moyen, à contrôler l'identifiabilité ou la diffusion de l'information avant son transfert en aval aux chercheurs?

5. Diffusion contrôlée

Au lieu d'être diffusées à grande échelle, les données peuvent être communiquées aux chercheurs en vertu d'accords visant entre autres à protéger la vie privée et la confidentialité. Sur le plan légal, de tels accords correspondent à des engagements contractuels. Souvent, il s'agit de promesses publiques, comme dans le cas où les données proviennent d'établissements gouvernementaux ou d'organisations financées par des fonds publics. Souvent, par souci de clarté juridique, les définitions ou les conditions stipulées dans les accords font référence à des définitions ou à des exigences du Code de réglementation fédérale des États-Unis (le Common Rule), le Privacy Rule de la HIPAA (*Health Insurance Portability and Accountability Act*) ou d'autres législations, règlements ou directives.

De nombreux projets génèrent des données ayant fait l'objet de deux mesures de précaution : la première visant à réduire l'identifiabilité à un degré raisonnable en éliminant les identificateurs directs, en arrondissant les données (par exemple la date de naissance à l'année de naissance ou à une fourchette d'âge) et ainsi de suite, et comprend également un examen, à tout le moins officieux, de la divulgation; la deuxième mesure autorisant l'accès aux données en vertu d'un accord de diffusion contrôlée.

robust, affordable, and routine, and genomic information becomes more easily abusable.

Q12: Overall, how much should we fret over genomic disclosure risks? (Details?)

Q13: Should the research community worry very much about access to protected genomic research data or biospecimens by the police, FBI, or other forces of public order, as compared with accidental release or malicious intrusion? Why? What, if anything, should it be doing differently?

Q14: Is there any reason to re-examine the Fort Lauderdale Principles, given that they are flexible and voluntary? Would any other forms of guidance be useful?

7. Flanking issues

Here, in no particular order, are some aspects of the larger puzzle that this project couldn't address but that very much need to be pursued.

Construal of genomic "human subject" under the Common Rule and other regulations. The fact that this is a perennial issue doesn't mean it shouldn't be worked on. Genomic research faces difficult questions regarding such matters as the status of people as "subjects" (or not) whose data or biospecimens have been assiduously de-identified, and the status as subjects of uninvolved relatives of people whose specimens are genotyped or being considered for genotyping. The answers have implications regarding, inter alia, identification and consent.

Consent. As an ethical matter, should consent be relied upon to justify deposition, in a publicly-accessible database, of data that have some realistic chance of being identifiable?

Controlled-release arrangements. As was suggested at the end of chapter 5, arrangements need to be explored that meet the ethical, legal, IT, managerial, and public perception challenges, and at the same time don't erect impractical barriers to research. Needing to be addressed with this are the special issues that arise when access to data is provided across national jurisdictions.

Termes des accords

Le tableau 3 (à la fin de ce chapitre) présente une liste non exhaustive des termes et expressions couramment utilisés dans ce type d'accord. Naturellement, il est peu probable qu'un accord contienne tous les termes figurant dans ce tableau, et certains accords pourraient contenir des termes ou expressions qui n'y figurent pas; il s'agit d'une liste de base. Les termes qui suivent sont parmi les plus pertinents en matière de protection de l'identifiabilité.

Consentement relatif à l'identifiabilité. Ce consentement peut porter sur le respect des engagements visant à protéger l'identifiabilité de diverses façons. Il peut par exemple comprendre des restrictions sur les utilisateurs autorisés des données en question ou sur l'objet ou d'autres aspects de l'utilisation des données, et ce, pour l'ensemble complet des données ou encore pour certains des sujets sources; ces restrictions peuvent être indicatrices de la confiance inspirée. Le consentement peut en outre porter sur la publication et certifier que l'identité des sujets sera rigoureusement dissimulée ou, au contraire, prévenir de certains risques non négligeables que pourraient comporter la publication quant à l'identifiabilité.

Protection de la confidentialité. Cette disposition doit toujours comprendre les mesures de sécurité prévues sur les plans physique, organisationnel et informatique. Peut faire référence à la conformité à l'égard des normes de l'Organisation internationale de normalisation, de la HIPAA ou d'autres normes de sécurité. Précise habituellement qui, s'il y a lieu, sera responsable de dépersonnaliser les données ainsi que la procédure et les critères pour ce faire. Peut mentionner les procédés de codage, les méthodes d'utilisation et les mesures de protection de la clé employée pour le codage. Stipule presque toujours que les chercheurs ne doivent pas tenter de réidentifier les sujets sources ni d'entrer en contact avec eux.

Limitation des transferts subséquents. Cette disposition vise à restreindre le transfert subséquent de données (ou l'accès aux données, ce qui revient au même) et à prolonger la chaîne de confidentialité ainsi que les obligations qui y sont associées.

Couplage. Le couplage des données peut être

Certificates of Confidentiality, the legal assurances that NIH can issue under the Public Health Service Act that "allow the investigator and others who have access to research records to refuse to disclose identifying information on research participants in any civil, criminal, administrative, legislative, or other proceeding, whether at the federal, state, or local level."42 They offer protection, but have limitations. How useful can the Certificates, or for that matter any conceivable legal ring-fence against forced disclosure, be for genomic projects?

Genetic anti-discrimination laws. The legislative saga rumbles on....

Protection of information on deceased persons. Continuing protection after death is required under the HIPAA Privacy Rule, but [bizarrely, to this observer], not under the Common Rule. The directness of the implications of people's genomic data for surviving relatives makes this issue more important for genomics than for most other health sciences.

Appendix. Sketches of a few projects

Framingham Heart Study. A study, begun in 1948, of the causes of cardiovascular and related diseases that has followed a cohort of some 5,200 people originally living around Framingham, Massachusetts, and many of their children, and is now recruiting grandchildren. In its latest phase Framingham has begun examining genetic factors. (www.framingham.com/heart, and www.nhlbi.nih.gov/about/framingham)

Genes and Environment Initiative (GEI). An ambitious proposed NIH-wide program to analyze genomic factors, develop improved technologies for monitoring exposures, and study how genes and exposures interact as risk factors of disease. While Congressional budget approval is pending, elaborate concept exploration is being conducted. (http://grants1.nih.gov/grants/guide/rfa-files/RFA-HG-06-033.html)

Genetic Association Information Network (GAIN). A public-private cooperative project of the Foundation for the NIH, NIH, Pfizer Inc, Affymetrix Inc., the Broad Institute, and Abbott Laboratories. Will perform whole genome association studies on samples provided from existing case-control studies of patients having common diseases. Full

discuté, notamment s'il risque d'accroître l'identifiabilité.

Publication, diffusion et communication des résultats. On précise dans cette rubrique si certaines données doivent être renvoyées à la source originale des données et si les résultats détaillés, comme des séquences, peuvent être rendus publics, c'est-à-dire publiés dans des bases de données accessibles au grand public, ou si les données doivent faire l'objet d'un contrôle plus étroit de la diffusion. La publication, la diffusion et la communication des données requièrent habituellement que l'identifiabilité des renseignements dérivant des données soit protégée de la même façon que le sont les données fournies.

Approbation d'un comité d'examen institutionnel ou d'un autre comité d'éthique. Cette disposition permet d'obtenir l'avis d'un comité d'examen ou d'éthique sur la conformité à l'égard du Common Rule ou d'autres obligations. Reste à savoir à quelle étape ou à quel moment on doit demander à un comité d'examen institutionnel d'examiner la situation.

Accords de diffusion contrôlée

De nombreux aspects pratiques et détails de gestion doivent être pris en compte, mais les fournisseurs et les utilisateurs de données connaissent habituellement bien les accords de diffusion contrôlée. Les trois exemples ci-dessous illustrent différents types d'accords de diffusion contrôlée.

- Wellcome Trust Case Control Consortium. Le consortium, qui comprend un certain nombre d'unités universitaires du R.-U., procède au génotypage d'échantillons d'ADN dépersonnalisés provenant de milliers de personnes atteintes de maladies chroniques ainsi que de sujets témoins. L'accès aux données épurées, brutes et sommaires est régi par un comité d'accès aux données (Consortium Data Access Committee) et doit faire l'objet d'une entente (Data Access Agreement)[37].

- Framingham Heart Study. Les données génétiques et autres de l'étude de Framingham sont diffusées en vertu d'une entente (Data and Materials Distribution

planning is underway, with initial funding decisions to be announced soon. (www.fnih.org/GAIN/GAIN_home.shtml)

Genome wide association study (GWAS). A generic term, which NIH defines as including "any study of genetic variation across the entire human genome that is designed to identify genetic associations with observable traits (such as blood pressure or weight), or the presence or absence of a disease or condition."[43]

National Health and Nutrition Examination Survey (NHANES). A long-running series of examination surveys conducted by the Centers for Disease Control and Prevention (CDC). NHANES makes DNA from many later-phase cohort members available for analysis under very tightly controlled conditions, but it does not allow release of any genomic data. (www.cdc.gov/nchs/about/major/nhanes/research_proposal_guidelines.htm)

NHGRI Medical Sequencing Program (MSP). A program in which contributing investigators will submit samples and phenotypic data, and NHGRI will perform sequencing, maintain all the data in a database, and manage release of the data. Initial intentions are to sequence intervals associated with Mendelian disorders, and to sequence large numbers of samples from studies of complex disorders in order to gauge the distribution and frequency of medically relevant genes. Detailed planning and pilot analyses are underway. (http://www.genome.gov/15014882)

The Cancer Genome Atlas (TCGA). A proposed project to chart the inherited and acquired mutations that relate to the onset, diagnosis, progression, and treatment of cancers, by genotyping biospecimens and examining the genomic data in light of clinical data on the patients. Piloting is starting. (http://cancergenome.nih.gov)

UK Biobank. A project that at the end of 2006 will start recruiting 500,000 people around the U.K. in the age range 40−69, conduct physical examinations, collect biospecimens and lifestyle data, link to NHS medical records, and follow the health trajectories of the participants for several decades. Consent will be very broad and will cover possible genotyping. (www.ukbiobank.org)

Agreement)[38] et seulement après que deux comités en aient approuvé la publication.

- Genetic Association Information Network (GAIN). Comme le projet GAIN porte sur le génotypage d'ADN soumis par divers groupes d'études universitaires sur des maladies précises, des groupes relevant de différents comités d'examen institutionnels, financés par des sources différentes et ayant conclu des ententes de consentement différentes, et parce que les membres du projet se préoccupent de l'identifiabilité des données qu'ils s'attendent à générer en abondance, ils ont récemment modifié leur politique de diffusion à échelle relativement grande pour une diffusion contrôlée. Les fournisseurs de données seront les principaux responsables de la desidentification des données. L'accès aux données sera régi par un comité d'accès aux données et devra faire l'objet d'une certification (Data Access Certification)[39].

L'exemple du réseau GAIN illustre l'importance, pour le milieu de la recherche en génomique, d'établir des critères pour déterminer les types et la quantité de données pouvant être publiés à grande échelle et ceux devant faire l'objet d'une diffusion contrôlée.

Au lieu de transférer des ensembles complets de données à des chercheurs de l'extérieur, on peut utiliser des dossiers plus restreints, notamment :

- des unités internes sur place (ou des centres de données de recherche, comme plusieurs centres fédéraux sont appelés) où viennent les chercheurs pour effectuer des études sur une base de données stockée sur un serveur surveillé, distinct et sécurisé;

- des systèmes d'interrogation à distance, au moyen desquels les chercheurs interrogent, par des systèmes de télécommunications sécurisés, les bases de données et obtiennent des réponses, possiblement occultées d'une manière ou d'une autre;

- des recommandations de services qui analysent les données ou des échantillons biologiques selon des méthodes convenues et fournissent les résultats aux chercheurs demandeurs.

End notes

[1] Federal Policy for the Protection of Human Subjects, HHS version (revised June 23, 2005); http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.htm.

[2] HHS, Standards for Privacy of Individually Identifiable Health Information; www.hhs.gov/ocr/hipaa. For interpretation see "Protecting personal health information in research: understanding the HIPAA Privacy Rule," http://privacyruleandresearch.nih.gov/pr_02.asp.

[3] Regarding overlap between the two Rules, see Congressional Research Service, report LC (X):RL32909, "Federal protection for human research subjects: an analysis of the Common Rule and its interactions with FDA regulations and the HIPAA privacy rule" (updated June 2, 2005); among other websites, posted at www.fas.org/sgp/crs/misc/RL32909.pdf.

[4] Portal to the EU Data Protection Directive, the national laws, and the authorities who administer them: http://ec.europa.eu/justice_home/fsj/privacy.

[5] David R. Bentley, "Genomic sequence information should be released immediately and freely into the public domain," *Science* 274, 533-534 (1996).

[6] NHGRI, "Reaffirmation and extension of NHGRI rapid data release policies" (February 2003); www.genome.gov/10506537.

[7] Wellcome Trust (writing as convenor), "Sharing data from large-scale biological research projects: A system of tripartite responsibility" (2003); www.wellcome.ac.uk/assets/wtd003207.pdf.

[8] "The responsible use and publication of HapMap data"; www.hapmap.org/guidelines_hapmap_data.html.en.

[9] NIH, "Final NIH statement on sharing research data" (2003) and related documents; http://grants.nih.gov/grants/policy/data_sharing.

[10] For brief sketches of a few such projects, see the Appendix.

[11] Bradley Malin and Latanya Sweeney, "Determining the identifiability of DNA database entries," *Proceedings of the American Medical Informatics Association Symposium 2000*, 537-541 (2000); available at http://privacy.cs.cmu.edu/dataprivacy/projects/genetic/dna1.html.

[12] Bradley A. Malin, "An evaluation of the current state of genomic data privacy protection technology and a roadmap for the future," *Journal of the American Medical Informatics Association* 12, 28-34 (2005).

Il faut concevoir des façons de rendre la diffusion contrôlée pratique, obligatoire et acceptable – conditions qui ne sont pas trop onéreuses du point de vue des procédures, mais qui permettent en même temps d'obtenir des engagements véritables, officiels et applicables. Il se peut qu'on établisse, par exemple pour certains groupes de projets, une licence générale d'utilisation des données aux termes de laquelle les chercheurs et leur établissement acceptent certaines conditions et puissent accéder à un large éventail de données ou d'échantillons biologiques.

Q11 : Serait-il intéressant d'étudier la possibilité d'établir des licences générales d'utilisation des données qui permettent d'avoir accès à des données génomiques centralisées? A-t-on déjà fait l'expérience d'un tel mécanisme?

(voir la Figure 3 dans l'appendice 1.3)

6. Risques généraux liés à l'identifiabilité

Tout cela doit être examiné du point de vue des risques – risques pour les sujets sources de données, risques pour les gardiens des données, risques pour les chercheurs et leurs établissements, et même risque pour l'entreprise de recherche en génomique. On doit se demander si les données peuvent être utilisées pour a) déduire l'identité des sujets sources de données ou b) déduire des faits relativement à des sujets sources de données, et il faut déterminer dans l'un ou l'autre cas si cela peut être préjudiciable.

Pour mesurer les risques de tout type, il faut mener à bien deux activités. Tout d'abord, l'« évaluation du risque » estime la probabilité que des événements indésirables accroissent la gravité des conséquences possibles. Puis, l'« appréciation du risque » soupèse les risques du point de vue des valeurs personnelles ou sociales. Cette appréciation peut tenir compte de la volonté d'investir dans la réduction du risque en diminuant la probabilité ou les enjeux ou les deux. Une appréciation plus large peut consister à soupeser les risques et les avantages et à déterminer si les risques sont acceptables. C'est la façon dont les gens examinent la plupart des situations dans leur vie, qu'ils en soient conscients ou non[41].

[13] Amy L. McGuire and Richard A. Gibbs, "No longer de-identified," *Science* 312, 370-371 (2006).
[14] Morris W. Foster and Richard R. Sharp, "Ethical issues in medical-sequencing research: implications of genotype−phenotype studies for individuals and populations," *Human Molecular Genetics* 15, R45-R49 (2006).
[15] Although the Act applies only to data about living individuals, professional guidance in the U.K. advises that medical data should be held in confidence after death as well. Because of the implications for relatives, the issue of protection of DNA and genomic data after death warrants re-evaluation everywhere now.
[16] "Procedures for Limited Data Sets," www.cms.hhs.gov/PrivProtectedData/10_LimitedDataSets.asp. Provision of CMS data is governed by strict Data User Agreements.
[17] For some considerations in key-coding see William W. Lowrance, *Learning from Experience: Privacy and the Secondary Use of Data in Health Research* (The Nuffield Trust, London, November 2002), pp. 32-33; www.nuffieldtrust.org.uk/ecomm/files/161202learning.pdf
[18] Zhen Lin, Art B. Owen, and Russ B. Altman, "Genomic research and human subject privacy," *Science* 303, 183 (2004), with supporting calculations at www.sciencemag.org/cgi/content/full/305/5681/183/DC1.
[19] A standard text is John M. Butler, *Forensic DNA Typing*, 2nd edition (Elsevier, Amsterdam and Boston, 2005).
[20] Frederick K. Bieber, Charles H. Brenner, and David Lazar, "Finding criminals through DNA of their relatives," *Science* 312, 1315-1316 (2006).
[21] Zhen Lin, Russ B. Altman, and Art B. Owen, Letter, "Confidentiality in genome research," *Science* 313, 441-442 (2006).
[22] A rich source is the American Statistical Association's website on Privacy, Confidentiality, and Data Security; www.amstat.org/comm/CmtePC. Another is the Federal Committee on Statistical Methodology's *Working Paper 22*; "Report on statistical disclosure limitation methodology," www.fcsm.gov/working-papers/spwp22.html.
[23] A standard text is Josep Domingo-Ferrer, editor, *Inference Control in Statistical Databases* (Springer-Verlag, Berlin, 2002).
[24] Thousands of databases, many of them holding voluminous personally identified information, can be consulted via services such as www.searchsystems.net and www.choicepoint.com.

L'évaluation du risque lié à la divulgation des renseignements génomiques doit tenir compte de certains facteurs comme l'étendue du génome en question, la densité ou la résolution de la cartographie, la rareté des variantes (car la rareté augmente l'identifiabilité), le degré de déséquilibre du croisement, et la spécificité avec laquelle les effets génétiques sont connus. En génomique, il faut porter une attention spéciale au risque lié à la divulgation dans le cas des parents par le sang de personnes dont le génome est étudié, risque qui a des répercussions sur le consentement et les mesures de protection.

Nous ne pouvons pas aborder ici les mesures de protection sauf pour déclarer de façon impérative qu'un éventail de mesures de protection physiques, procédurales, juridiques et liées à la cybersécurité et à la formation doivent être mises en place pour prévenir la diffusion accidentelle de même que l'accès importun à une ressource. Entre autres, *les mesures de protection sont ce qui justifie la sollicitation d'un consentement général*.

(Nous laisserons à d'autres le soin d'aborder la question générale des lois contre la discrimination génétique, qui tendent à mettre davantage l'accent sur les conséquences néfastes que sur les processus.)

Quels sont les dangers? Nous savons qu'on peut pénétrer dans les systèmes informatiques, obtenir des données par subterfuge, voler des ordinateurs portatifs et transférer inadéquatement des échantillons biologiques. Jusqu'à maintenant, dans un nombre étonnamment faible de cas, on a démontré que des données médicales, et plus rarement des données de recherche en santé, ont été utilisées à mauvais escient. Mais par suite de l'introduction des dossiers médicaux électroniques, du croisement accru des bases de données, etc.. – et compte tenu de l'inquiétude vague qu'éprouvent de nombreuses personnes à l'égard de tout ce qui est « génomique » — les craintes de la population ne cessent de croître. Comme nous l'avons mentionné au début du présent document, les usages abusifs qu'on peut imaginer vont de l'embarras, du chantage, de la fraude et de la stigmatisation de groupes à la discrimination négative pour l'assurance maladie ou l'assurance vie, l'emploi, les promotions, les hypothèques ou les prêts. Les tests de paternité obtenus sans consentement constituent, selon le point de vue, un autre usage abusif possible.

[25] For a review see John M. Butler, "Genetics and genomics of core short tandem repeat loci used in human identity testing," *Journal of Forensic Sciences* 51, 253-265 (2006); ethnicity profiling is discussed on p. 260.

[26] David A. Hinds, Laura L. Stuve, Geoffrey B. Nilsen, Eran Halperin, Eleazar Eskin, Dennis G. Ballinger, Kelly A. Frazer, and David R. Cox, "Whole genome patterns of common DNA variation in three human populations," *Science* 307, 1072-1079 (2005).

[27] Anecdotal communications to the author

[28] For perspective see Noah A. Rosenberg, Jonathan K. Pritchard, James L. Weber, Howard M. Cann, Kenneth K. Kidd, Lev A. Zhivotovsky, and Marcus W. Feldman, "Genetic structure of human populations," *Science* 298, 2381-2385 (2002).

[29] NHGRI Race, Ethnicity, and Genetics Working Group, "The use of racial, ethnic, and ancestral categories in human genetics research," *American Journal of Human Genetics 77*, 519-532 (2005).

[30] Susanne B. Haga, "Policy implications of defining race and more by genome profiling," *Genomics, Societyand Policy* [online] 2 (1), 57-71 (2006); www.gspjournal.com.

[31] Regarding such techniques generally see footnote 22 above.

[32] Statistical approaches to estimating and reducing the disclosure risks of SNP databases are explored in Zhen Lin, "Balancing utility and anonymity in public biomedical databases," doctoral dissertation, Stanford University (April 2005); http://helix-web.stanford.edu/people/zlin/pubs/zlin_thesis.pdf.

[33] European Medicines Evaluation Agency, "Position paper on terminology in pharmacogenetics" (2002); www.emea.eu.int/pdfs/human/press/pp/307001en.pdf.

[34] Standardization of vocabulary is now being considered by the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH).

[35] HHS, Office for Human Research Protection, "Guidance on research involving coded private information or biological specimens" (2004); www.hhs.gov/ohrp/humansubjects/guidance/cdebiol.pdf

[36] Regarding IRB issues see Virginia de Wolf, Joan E. Silber, Philip M. Steel, and Alvan O. Zarate, "Meeting the challenge when data sharing is required," *IRB: Ethics & Human Research* 28 (2), 10-15 (2006).

Devrait-on s'attendre à des diffusions accidentelles, à du piratage informatique et à des tentatives d'usage abusif? Assurément! Il est inutile de spéculer sur les détails, mais des exercices de prévision du risque peuvent aider à identifier les vulnérabilités et à proposer des mécanismes de défense. Il faut présumer que certains dangers sont des possibilités réelles et que certains peuvent avoir des conséquences graves.

Il est essentiel de prévenir les dangers. Il serait certes important de ne pas dévoiler « une trop grande partie » du génome des personnes dans les années à venir, pour ensuite le regretter lorsque les techniques d'analyse deviendront plus fermes, moins coûteuses et plus courantes et qu'il sera plus facile d'utiliser à mauvais escient l'information génomique.

Q12 : En général, à quel point devrait-on s'inquiéter des risques de divulgation des renseignements génomiques? (Précisions?)

Q13 : Les chercheurs devraient-ils s'inquiéter grandement de l'accès par les policiers, le FBI ou d'autres forces de l'ordre à des données ou à des échantillons biologiques protégés issus de la recherche en génomique, comparativement à la communication accidentelle ou à l'intrusion malicieuse? Pourquoi? Qu'est-ce que les chercheurs devraient faire différemment, le cas échéant?

Q14 : Y a-t-il lieu de réexaminer les Fort Lauderdale Principles, étant donné qu'ils sont souples et volontaires? D'autres formes de lignes directrices seraient-elles utiles?

7. Questions secondaires

Nous présentons ici dans le désordre certains des aspects du tableau d'ensemble auxquels le présent projet ne pouvait s'attaquer mais qu'il faut vraiment examiner.

Interprétation de ce qu'on entend par « sujet humain » en génomique en vertu du *Common Rule* et d'autres règlements. Le fait qu'il s'agisse d'une question perpétuelle ne signifie pas qu'on ne devrait pas s'y attaquer. La recherche en

[37] www.wtccc.org.uk.
[38] www.nhlbi.nih.gov/about/framingham/policies/index.htm.
[39] www.fnih.org/GAIN/Updated_Data_Access_Policy.shtml.
[40] Adapted from William W. Lowrance, "Access to Collections of Data and Materials for Health Research" (March 2006); http://www.wellcome.ac.uk/doc_WTX030843.html.
[41] Such odds−stakes thinking comes naturally when deciding where to cross a street, which ski run to attempt, whether to carry an umbrella (to the gym vs to a wedding), what intimate concerns to confide (to a close friend vs to a casual acquaintance), how candid to be in answering a questionnaire, whether to volunteer as a subject in alcoholism research....
[42] NIH Certificates of Confidentiality Kiosk: http://grants.nih.gov/grants/policy/coc.
[43] NIH, "Proposed policy for sharing of data obtained in NIH supported or conducted genome-wide association studies (GWAS)," 71 *Federal Register*, 51629-51631 (August 30, 2006).

génomique est confrontée à des questions difficiles concernant, notamment, le statut des personnes utilisées comme « sujets » (ou non) dont les données ou les échantillons biologiques ont été assidûment dépersonnalisés, et le statut comme sujets des non-participants parents de personnes dont les échantillons sont génotypés ou pris en considération pour un génotypage. Les réponses à ces questions ont des répercussions, entre autres, sur l'identification et le consentement.

Consentement. Du point de vue éthique, devrait-on s'appuyer sur le consentement pour justifier le dépôt, dans une base de données accessible au public, de données qui courent un risque réel d'être identifiables?

Accords de diffusion contrôlés. Comme nous l'avons proposé à la fin du chapitre 5, il faut examiner la possibilité d'accords qui répondent aux problèmes éthiques, juridiques, informatiques, aux problèmes de gestion et de perception publique et qui, en même temps, n'érigent pas de barrières insurmontables pour la recherche. Il faut se pencher parallèlement sur les problèmes spéciaux qui surgissent lorsque l'accès aux données est accordé outre frontière.

Certificats de confidentialité. Les garanties juridiques que le NIH peut délivrer en vertu de la Public Health Service Act qui [traduction] « permettent aux chercheurs et à d'autres qui ont accès aux dossiers de recherche de refuser de divulguer des renseignements permettant d'identifier des participants à la recherche dans des poursuites civiles, criminelles, des travaux administratifs, législatifs ou autres, que ce soit à l'échelle de l'administration fédérale, de l'État ou à l'échelle locale »[42]. Ces certificats offrent une protection, mais comportent des limites. Quelle utilité peuvent avoir les certificats, ou toute barrière juridique concevable, contre la divulgation forcée dans le cas des projets en génomique?

Lois contre la discrimination génétique. La saga législative se poursuit…

Protection des renseignements sur les personnes décédées. Le maintien de la protection après le décès est obligatoire en vertu du *Privacy Rule* de la HIPAA, mais [bizarrement selon moi], pas en vertu du *Common Rule*. Les répercussions directes des données génomiques personnelles sur les parents survivants font en sorte que cette

question est plus importante en génomique que dans la plupart des autres sciences de la santé.

Annexe. Aperçus de quelques projets

Framingham Heart Study. Cette étude, qui a débuté en 1948, porte sur les causes des maladies cardio-vasculaires et apparentées et a suivi une population de quelque 5 200 personnes qui vivaient à l'origine près de Framingham, Massachusetts, ainsi que bon nombre de leurs enfants; on recrute maintenant les petits-enfants. Dans sa dernière phase, l'étude a commencé à examiner les facteurs génétiques (www.framingham.com/heart et www.nhlbi.nih.gov/about/framingham).

Genes and Environment Initiative (GEI). Ce programme ambitieux proposé à l'échelle du NIH vise à analyser les facteurs génomiques, à améliorer les techniques pour surveiller les expositions et à étudier la façon dont les gènes et les expositions interagissent comme facteurs de risque de maladies. Bien que le budget n'ait pas encore été approuvé par le Congrès, on explore actuellement les concepts (http://grants1.nih.gov/grants/guide/rfa-files/RFA-HG-06-033.html).

Genetic Association Information Network (GAIN). Projet de coopération public-privé de la Foundation for the NIH, du NIH, de Pfizer Inc., d'Affymetrix Inc., du Broad Institute et des Laboratoires Abbott. On effectuera des études d'association génomique à partir d'échantillons fournis par les études cas/témoins existantes portant sur des patients qui ont des maladies communes. On a commencé la planification, et les décisions initiales relatives au financement seront bientôt annoncées (www.fnih.org/GAIN/GAIN_home.shtml).

Étude d'association à l'échelle du génome (GWAS). Terme générique qui, selon le NIH, inclut [traduction] « toute étude de la variation génétique à l'échelle de tout le génome humain qui vise à identifier les associations génétiques avec des traits observables (comme la pression sanguine ou le poids) ou la présence ou l'absence d'une maladie ou d'un trouble »[43].

National Health and Nutrition Examination Survey (NHANES). Série d'enquêtes effectuées depuis longtemps par les Centers for Disease Control and Prevention (CDC). La NHANES rend

l'ADN de nombreux membres de la population de la dernière phase de l'enquête accessible pour l'analyse si l'on respecte un certain nombre de conditions très strictes, mais elle ne permet pas la diffusion de données génomiques (www.cdc.gov/nchs/about/major/nhanes/research_proposal_guidelines.htm).

NHGRI Medical Sequencing Program (MSP). Programme qui sera utilisé par les chercheurs-collaborateurs pour soumettre des échantillons et des données phénotypiques et qu'emploiera le NHGRI pour effectuer le séquençage, tenir à jour toutes les données dans une base de données et gérer la diffusion des données. Les objectifs initiaux sont de séquencer des intervalles associés à des affections mendéliennes et de séquencer un grand nombre d'échantillons provenant d'études de troubles complexes afin de mesurer la distribution et la fréquence des gènes présentant un intérêt médical. Une planification détaillée et des analyses pilotes sont en cours (http://www.genome.gov/15014882).

The Cancer Genome Atlas (TCGA). Projet proposé de cartographie des mutations héréditaires et acquises ayant un lien avec l'apparition, le diagnostic, la progression et le traitement des cancers, au moyen du génotypage des échantillons biologiques et de l'examen des données génomiques à la lumière des données cliniques sur les patients. Un essai pilote vient de démarrer (http://cancergenome.nih.gov).

UK Biobank. Projet dans le cadre duquel on commencera à la fin de 2006 à recruter 500 000 personnes au R.-U. âgées de 40 à 69 ans, sur lesquelles des examens physiques seront effectués, des échantillons biologiques et des données sur le mode de vie seront recueillis, les données seront comparées avec les dossiers médicaux du NHS et l'évolution de la santé des participants sera suivie pendant plusieurs décennies. Le consentement sera très large et englobera un génotypage possible (www.ukbiobank.org).

Notes en bas de page

[1] Federal Policy for the Protection of Human Subjects, version du département de la Santé et des Services humanitaires (HHS) (document révisé le 23 juin 2005). Sur Internet : <http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.htm>.

[2] HHS, Standards for Privacy of Individually Identifiable Health Information. Sur Internet : <www.hhs.gov/ocr/hipaa>. Pour l'interprétation à donner, voir « Protecting personal health information in research: understanding the HIPAA Privacy Rule ». Sur Internet : http://privacyruleandresearch.nih.gov/pr_02.asp.

[3] Concernant le chevauchement entre les deux règlements, voir le rapport du service de recherche du Congrès LC(X) :RL32909, « Federal protection for human research subjects: an analysis of the Common Rule and its interactions with FDA regulations and the HIPAA privacy rule » (document mis à jour le 2 juin 2005). Document affiché, entre autres, sur le site <www.fas.org/sgp/crs/misc/RL32909.pdf>.

[4] Portail d'accès à la Directive relative à la protection des données de l'UE, aux lois nationales et aux autorités qui les appliquent :< http://ec.europa.eu/justice_home/fsj/privacy/index_fr.htm>.

[5] David R. Bentley, « Genomic sequence information should be released immediately and freely into the public domain », Science 274, 533-534 (1996).

[6] NHGRI, « Reaffirmation and extension of NHGRI rapid data release policies » (février 2003). Sur Internet : www.genome.gov/10506537.

[7] Wellcome Trust (à titre de responsable), « Sharing data from large-scale biological research projects: A system of tripartite responsibility » (2003). Sur Internet : <www.wellcome.ac.uk/assets/wtd003207.pdf>.

[8] « L'utilisation et la publication des données de HapMap ». Sur Internet : <www.hapmap.org/guidelines_hapmap_data.html.fr>.

[9] NIH, « Final NIH statement on sharing research data » (2003) et documents connexes. Sur Internet : <http://grants.nih.gov/grants/policy / data_sharing>.

[10] Un bref aperçu de quelques projets est donné en annexe.

[11] Bradley Malin et Latanya Sweeney, « Determining the identifiability of DNA database entries », Proceedings of the American Medical Informatics Association Symposium 2000, 537-541 (2000). Consultable à l'adresse <http://privacy.cs.cmu.edu/dataprivacy/projects/genetic/dna1.html>.

[12] Bradley A. Malin, « An evaluation of the current state of genomic data privacy protection technology and a roadmap for the future », Journal of the American Medical Informatics Association 12, 28-34 (2005).

[13] Amy L. McGuire et Richard A. Gibbs, « No

longer de-identified », Science 312, 370-371 (2006).

[14] Morris W. Foster et Richard R. Sharp, « Ethical issues in medical-sequencing research: implications of genotype-phenotype studies for individuals and populations », Human Molecular Genetics 15, R45-R49 (2006).

[15] La Loi s'applique uniquement aux données sur des personnes vivantes mais, au Royaume-Uni, des directives à l'intention des professionnels recommandent que les données médicales soient gardées confidentielles même après le décès. En raison des incidences sur les proches, la question de la protection des données sur l'ADN et des données génomiques après le décès devrait aujourd'hui être réexaminée partout dans le monde.

[16] « Procedures for Limited Data Sets ». Sur Internet : www.cms.hhs.gov/PrivProtectedData/10_LimitedDataSets.asp. La communication des données des CMS est régie par des ententes strictes avec les utilisateurs de données.

[17] Pour de plus amples renseignements concernant l'utilisation de codes d'identification, veuillez consulter William W. Lowrance, Learning from Experience : Privacy and the Secondary Use of Data in Health Research (The Nuffield Trust, London, novembre 2002), p. 32-33. Sur Internet : www.nuffieldtrust.org.uk/ecomm/files/161202learning.pdf.

[18] Zhen Lin, Art B. Owen, and Russ B. Altman, « Genomic research and human subject privacy », Science 303, 183 (2004), avec calculs à l'appui. Sur Internet : www.sciencemag.org/cgi/content/full/305/5681/183/DCI.

[19] Voir à ce sujet John M. Butler, Forensic DNA Typing, 2nd edition (Elsevier, Amsterdam and Boston, 2005).

[20] Frecerick K. Bieber, Charles H. Brenner, and David Lazar, « Finding criminals through DNA of their relatives », Science 312, 1315-1316 (2006).

[21] Zhen Lin, Russ B. Altman, and Art B. Owen, Lettre, « Confidentiality in genome research », Science 313, 441-442 (2006).

[22] À ce sujet, le site Web sur la confidentialité et la sécurité des données de l'American Statistical Association est particulièrement riche d'informations. Sur Internet : www.amstat.org/comm/CmtePC. Un autre document intéressant est le Working Paper 22, du Federal Committee on Statistical Methodology; « Report on statistical disclosure limitation methodology ». Sur Internet : www.fcsm.gov/working/papers/spwp22.html.

[23] Document de fond : Josep Domingo-Ferrer,

editor, Inference Control in Statistical Databases (Springer-Verlag, Berlin, 2002).

[24] Il est possible de consulter des milliers de bases de données, dont beaucoup contiennent une quantité considérable de renseignements personnellement identifiés, en ayant recours à des services comme www.searchsytstems.net et www.choicepoint.com.

[25] Pour un survol de la question, consulter John M. Butler, « Genetics and genomics of core short tandem repeat loci used in human identity testing », Journal of Forensic Sciences 51, 253-265 (2006). La question du recours au profilage pour déterminer l'ethnicité est abordée à la page 260.

[26] David A Hinds, Laura L. Stuve, Geoffrey B. Nielsen, Eran Halperin, Eleazar Eskin, Dennis G. Ballinger, Kelly A. Frazer et David R. Cox, « Whole genome patterns of common DNA variation in three human populations », Science 307, 1072-1079 (2005).

[27] Communications anecdotiques de l'auteur.

[28] Pour un survol de la question, se reporter à Noah A. Rosenberg, Jonathan K. Pritchard, James L. Weber, Howard M. Cann, Kenneth K. Kidd, Lev A. Zhivotovsky et Marcus W. Feldman, « Genetic structure of human populations », Science 298, 2381-2385 (2002).

[29] NHGRI Race, Ethnicity, and Genetics Working Group, « The use of racial, ethnic, and ancestral categories in human genetics research », American Journal of Human Genetics 77, 519-532 (2005).

[30] Susanne B. Haga, « Policy implications of defining race and more by genome profiling », Genomics, Society and Policy [en ligne] 2 (1), 57-71 (2006); www.gspjournal.com.

[31] En ce qui concerne ces techniques en général, consulter la note de bas de page no 22, plus haut.

[32] Les méthodes statistiques d'estimation et de réduction du risque de divulgation des bases de données SNP sont explorées dans un mémoire rédigé par Zhen Lin, « Balancing utility and anonymity in public biomedical databases », Université de Stanford (avril 2005). Sur Internet : http://helix-web.stanford.edu/people/zlin/pubs/zlin_thesis.pdf.

[33] Agence européenne du médicament (EMEA), « Position paper on terminology in pharmacogenetics » (2002). Sur Internet : www.emea.eu.int/pdfs/human/press/pp/307001en.pdf.

[34] La Conférence internationale sur l'harmonisation des exigences techniques relatives à l'homologation des produits pharmaceutiques à usage humain (CIH) se

penche actuellement sur la normalisation de la terminologie.

[35] HHS, Office for Human Research Protection, « Guidance on research involving coded private information or biological specimens » (2004). Sur Internet : www.hhs.gov/ohrp/humansubjects/guidance/cdebiol.pdf.

[36] Pour de plus amples renseignements sur les enjeux du CEI, consulter : Virginia de Wolf, Joan E. Silber, Philip M. Steel, and Alvan O. Zarate, « Meeting the challenge when data sharing is required », IRB: Ethics & Human Research 28 (2), 10-15 (2006).

[37] Sur Internet : www.wtccc.org.uk.

[38] Sur Internet : www.nhlbi.nih.gov/about/framingham/policies/index.htm.

[39] Sur Internet : www.fnih.org/GAIN/Updated_Data_Access_Policy.shtml.

[40] Adaptation de William W. Lowrance, "Access to Collections of Data and Materials for Health Research" (March 2006); http://www.wellcome.ac.uk/doc_WTX030843.html.

[41] Ce calcul de probabilité est naturel lorsqu'on choisit une rue à traverser, une pente de ski à descendre, décide s'il est utile d'apporter un parapluie (au gymnase p. opp. à un mariage), quels soucis personnels confier (à un ami intime p. opp. à une simple connaissance), dans quelle mesure on répondra franchement à un questionnaire, si l'on se portera volontaire comme sujet dans une recherche sur l'alcoolisme…

[42] NIH Certificates of Confidentiality Kiosk: http://grants.nih.gov/grants/policy/coc.

[43] NIH, "Proposed policy for sharing of data obtained in NIH supported or conducted genome-wide association studies (GWAS)," 71 Federal Register, 51629-51631 (30 août 2006).

Appendice 1.1

Figure 1. Concordance des termes relatifs à l'identifiabilité		
identifié ou identifiable	**assorti d'un code d'identification**	**non identifiable**
personnel nominatif	désidentifié de manière réversible anonymisé lié pseudonymisé pseudo-anonymisé crypté codé	désidentifié de manière irréversible anonymisé non lié non identifiable anonyme

Appendice 1.2

Figure 2. Liste d'identificateurs du *Privacy Rule* de l'HIPAA (§164.514(B)(2))

[traduction]
(i) L'entité couverte peut juger que l'information sur la santé n'est pas individuellement identifiable seulement si (…) les identificateurs suivants du sujet ou des parents, employeurs ou membres du ménage du sujet sont supprimés :

(A) Noms;

(B) Toutes les subdivisions géographiques plus petites qu'un État, y compris l'adresse municipale, la ville, le pays, la circonscription, le code postal et les géocodes équivalents, à l'exception des trois premiers chiffres d'un code postal si, selon les données publiques du Bureau du recensement :

(1) L'unité géographique formée en combinant tous les codes postaux ayant les trois mêmes premiers chiffres contient plus de 20 000 personnes; et (2) les trois premiers chiffres d'un code postal pour toutes les unités géographiques contenant 20 000 personnes ou moins sont remplacés par 000;

(C) Tous les éléments des dates (sauf l'année), dans le cas des dates directement liées à un sujet, y compris la date de naissance, la date de l'admission, la date du congé, la date du décès; et tous les âges supérieurs à 89 ans, ainsi que tous les éléments des dates (y compris l'année) indiquant un tel âge, à l'exception des âges et des éléments qui peuvent être regroupés au sein d'un seul groupe d'âge de 90 ans et plus;

(D) Les numéros de téléphone;

(E) Les numéros de télécopieur;

(F) Les adresses électroniques;

(G) Les numéros de sécurité sociale;

(H) Les numéros de dossier médical;

(I) Les numéros de bénéficiaire d'un régime de santé

(J) Les numéros de compte;

(K) Les numéros de certificat/licence;

(L) Les identificateurs et numéros de série des véhicules, y compris les numéros de plaque d'immatriculation;

(M) Les identificateurs et numéros de série de dispositifs;

(N) Les localisateurs de ressources uniformes (URL);

(O) Les numéros de protocole Internet (IP);

(P) Les identificateurs biométriques, y compris les empreintes digitales et vocales;

(Q) Les images photographiques de face et toute image comparable;

(R) Tout autre numéro d'identification, caractéristique ou code unique (…)

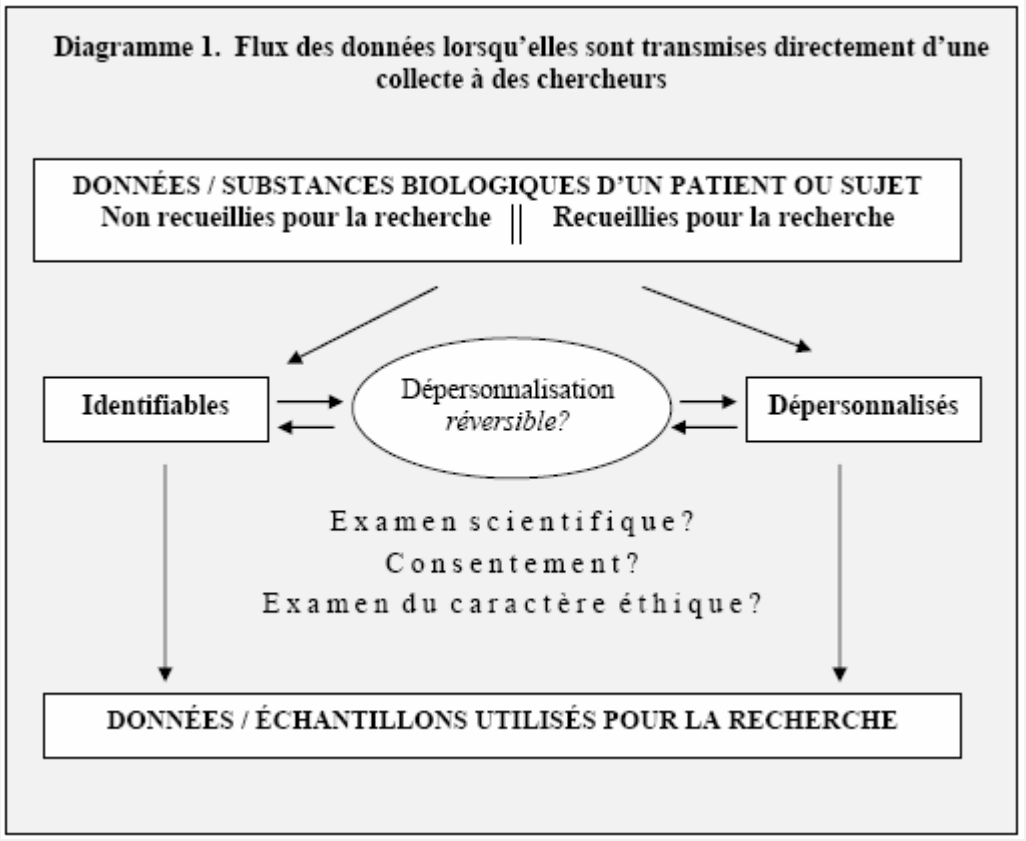
[et]

L'entité couverte n'a pas été informée du fait que l'information pourrait être utilisée seule ou en association avec d'autres informations pour identifier une personne qui est l'objet de l'information.

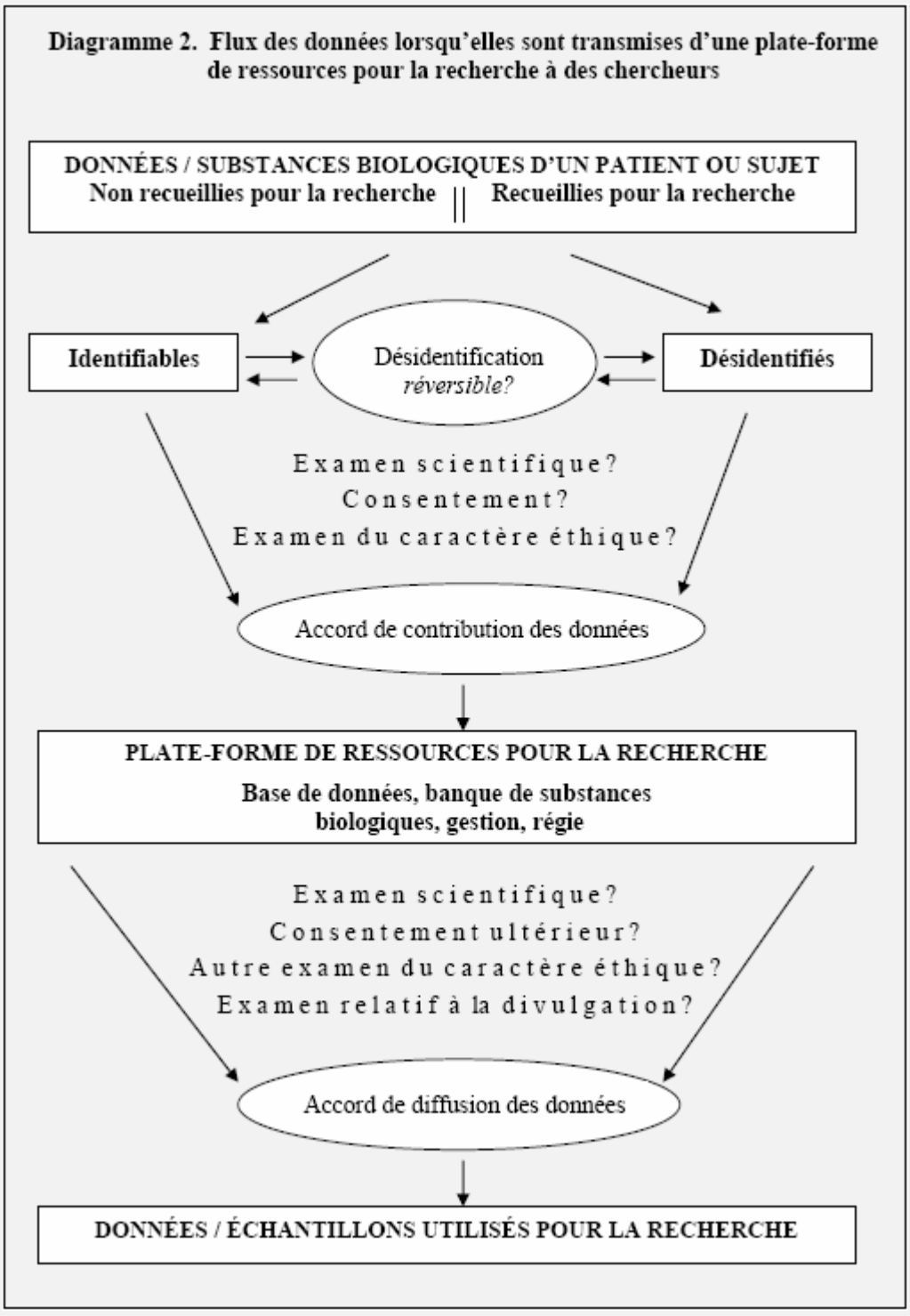
Appendice 1.3

Figure 3. Certaines des conditions des accords de diffusion des données et des échantillons biologiques	
Évaluation préliminaire de la compétence et du bien-fondé scientifiques	(Pourquoi une telle évaluation – pour protéger la ressource? Pour épargner des efforts?) Objets, rentabilité scientifique potentielle…
Spécification de ce qui est fourni	Données, substances biologiques, analyse, informatique, croisement, assistance, formation, archivage?
Consentement	Ce qu'il englobe, où est-il consigné, comment est-il retracé…
Limitation des usages	Pour une maladie particulière? Création de lignées cellulaires? Usage commercial?
Confidentialité	Désidentifier (comment)? Promettre de ne pas essayer de ré-identifier, sécurité, formation, …
CEI ou une autre approbation du caractère éthique	…au point de collecte des données? Au stade de la plate-forme des ressources pour la recherche avant la diffusion?
Limitation du transfert subséquent	Restrictions
Couplage	Attentes, restrictions
Communication à nouveau avec les sujets sources des données	Justifications, nouveau contact par qui et comment
Maintien de la qualité de la ressource	Promesse de s'occuper des erreurs ou de la contamination
Publication ou retour des résultats	Requis? Mode de publication? Moment?
Remerciements	« Redevable »
Coauteurs	Mention exigée pour le contrôle ou le partage du mérite?
Enrichissement de la ressource	Intégrer les résultats dans la ressource? Qui est responsable de la qualité?
Information des sujets sources des données	… au sujet des progrès? Des résultats relatifs à la personne?
Archivage	Comment? Qui paie? Conditions d'accès?
Droits de propriété intellectuelle	Attributions ou renonciation de la PI
Réponse si un sujet se retire	Détruire les données, les échantillons biologiques ou les liens?
Retour ou destruction du matériel	…à la fin? Si les engagements ne sont pas tenus?
Ordre de priorité pour l'accès aux données ou aux échantillons biologiques	…si la quantité de substances biologiques ou les ressources analytiques ou de TI est limitée
Frais	Pourquoi? Les frais dépendent-ils des possibilités relatives à la PI?
Application transfrontalière	Contraintes juridiques, approbation du caractère éthique, droit des sujets
Surveillance, supervision ou vérification	Plans, attentes
Éventualités si la ressource ou les éléments du projet sont abandonnés	Détruire les ressources? Transférer dans un autre établissement qui respectera les conditions?
Avis juridiques de non-responsabilité	[Non responsable pour la qualité ou les conséquences]

Appendice 1.4



Appendice 1.5



Appendix 2.1

Figure 1. Concordance of identifiability terms		
identified or identifiable	**key-coded**	**non-identifiable**
personal nominative	reversibly de-identified linked anonymized pseudonymized pseudoanonymized encrypted coded	irreversibly de-identified unlinked anonymized unidentifiable anonymous

Appendix 2.2

Figure 2. The HIPAA Privacy Rule's identifier list (§164.514(b)(2))

(i) A covered entity may determine that health information is not individually identifiable health information only if... the following identifiers of the individual or of relatives, employers, or household members of the individual, are removed:

(A) Names;

(B) All geographic subdivisions smaller than a State, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial three digits of a zip code if, according to the current publicly available data from the Bureau of the Census:

(1) The geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people; and (2) the initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000;

(C) All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older;

(D) Telephone numbers;

(E) Fax numbers;

(F) Electronic mail addresses;

(G) Social security numbers;

(H) Medical record numbers;

(I) Health plan beneficiary numbers;

(J) Account numbers;

(K) Certificate/license numbers;

(L) Vehicle identifiers and serial numbers, including license plate numbers;

(M) Device identifiers and serial numbers;

(N) Web Universal Resource Locators (URLs);

(O) Internet Protocol (IP) address numbers;

(P) Biometric identifiers, including finger and voice prints;

(Q) Full face photographic images and any comparable images; and

(R) Any other unique identifying number, characteristic, or code...

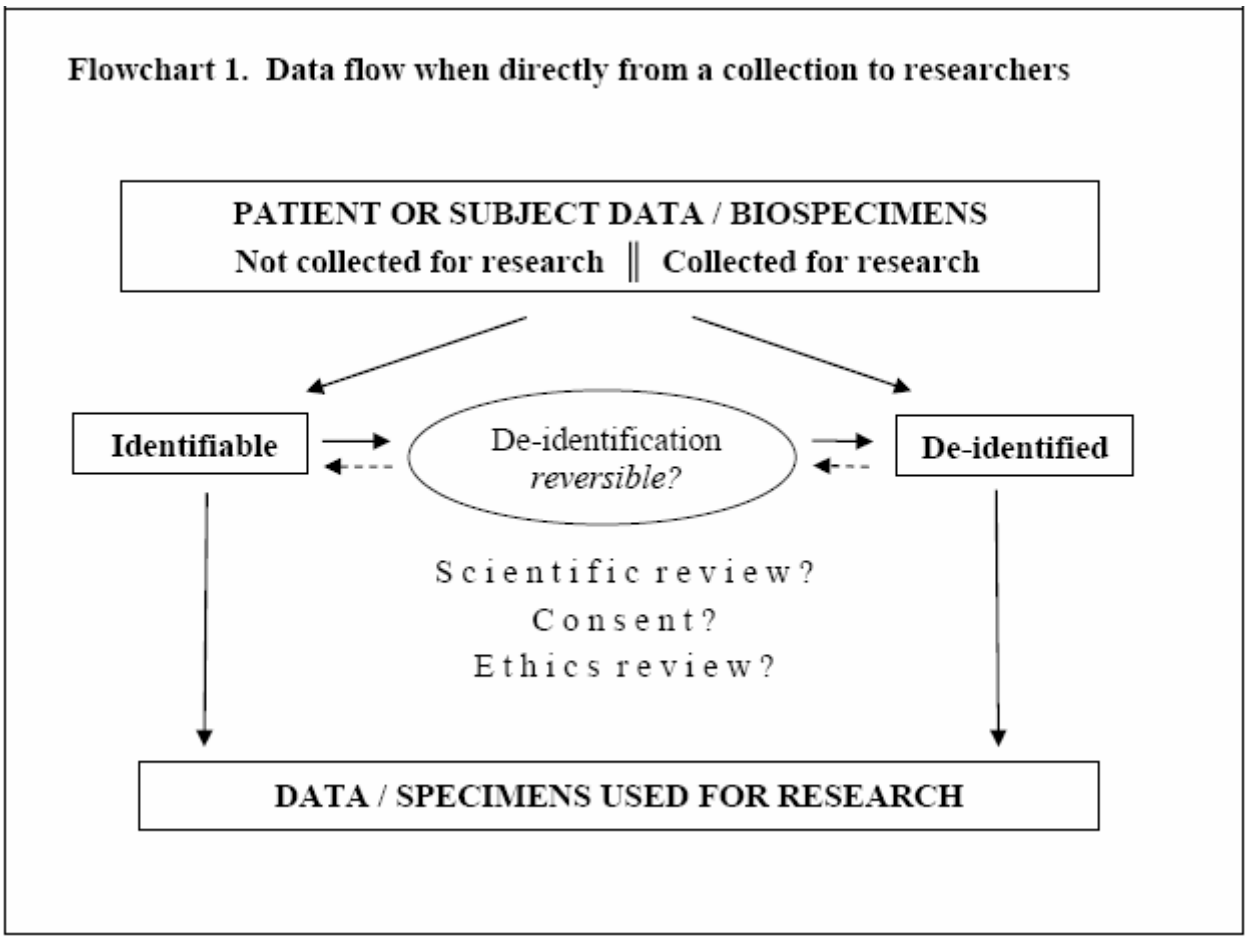
[and]

(ii) The covered entity does not have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a subject of the information.

Appendix 2.3

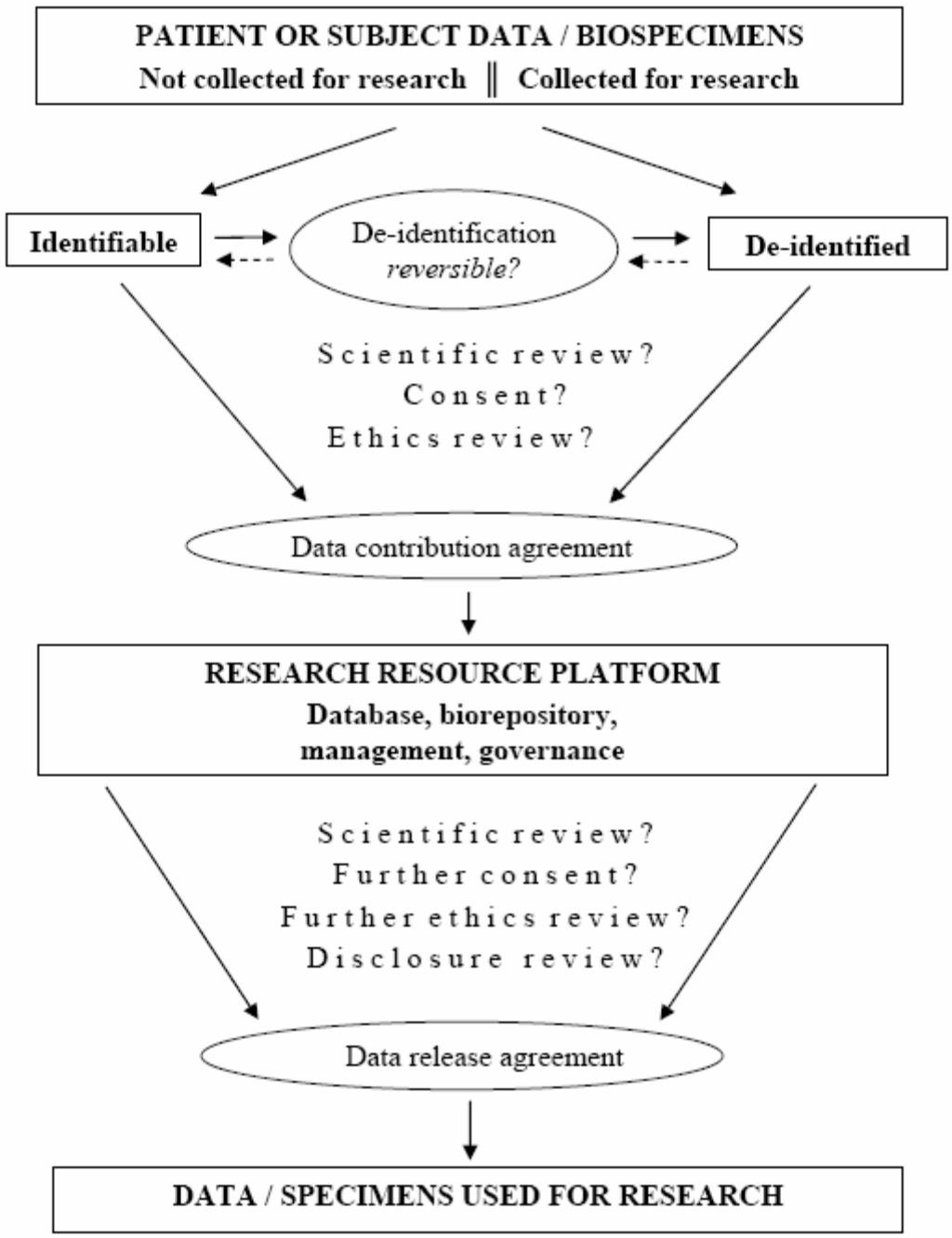
Figure 3. Some terms of data and biospecimen release agreements [40]	
Screening of scientific competence and merit	(why screen – to protect the resource? to conserve effort?) purposes, potential scientific payoff...
Specification of what is provided	data, biospecimens, analysis, informatics, linking, assistance, training, archiving?
Consent	coverage, documented where, tracked how...
Purpose limitation	disease-specific use? start cell lines? commercial use?
Confidentiality	de-identify (how)?, promise not to try to re-identify, security, training, ...
IRB or other ethics approval	...at point of data collection? at research resource platform stage before release?
Limiting of onward transfer	restrictions
Linking	expectations, restrictions
Recontacting data-subjects	justifications, recontact by whom and how
Maintaining quality of resource	promise to deal with errors or contamination
Publication or returning of findings	required? publish how? timing?
Acknowledgments	"much obliged"
Co-authoring	required for control or credit-sharing?
Enriching the resource	integrate findings into the resource? who is responsible for quality?
Informing data-subjects	...of progress? of person-specific findings?
Archiving	how? who pays? conditions of access?
Intellectual property rights	IP assignments or waiving
Responding if a subject withdraws	destroy data, biospecimens, or links?
Returning or destroying materials	...when finished? if commitments are broken?
Prioritization of access to data or biospecimens	...if biospecimen quantity or analytic or IT resources are limited
Fees	for what? does fee depend on IP prospects?
Transborder enforcement	legal constraints, ethics approval, subjects' rights
Monitoring, oversight, or audit	plans, expectations
Contingencies if resource or project elements are terminated	destroy the resources? pass on to another institution that will preserve the conditions?
Legal disclaimers	[not responsible for quality or consequences]

Appendix 2.4


Flowchart 1. Data flow when directly from a collection to researchers
PATIENT OR SUBJECT DATA / BIOSPECIMENS
Not collected for research ‖ Collected for research
Identifiable
De-identification
reversible?
De-identified
Scientific review?
Consent?
Ethics review?
DATA / SPECIMENS USED FOR RESEARCH

Appendix 2.5

Flowchart 2. Data flow when via a research resource platform to researchers



29E CONFÉRENCE INTERNATIONALE DES COMMISSAIRES
À LA PROTECTION DES DONNÉES ET DE LA VIE PRIVÉE

LES HORIZONS DE LA PROTECTION DE LA VIE PRIVÉE

TERRA INCOGNITA

PRIVACY HORIZONS

29TH INTERNATIONAL CONFERENCE OF
DATA PROTECTION AND PRIVACY COMMISSIONERS

Solutions de rechange à l'obtention du consentement pour chaque projet lorsqu'on utilise des renseignements personnels aux fins de la recherche en santé. Qu'en pensent les Canadiennes et les Canadiens ?

Alternatives to project-specific consent for access for personal information for health research. What do Canadians think?

Par/by:

Donald J. Willison, Lisa Schwartz, Julia Abelson, Cathy Charles, Marilyn Swinton, David Northrup et/and Lehana Thabane

Office of the Privacy Commissioner of Canada

Commissariat à la protection de la vie privée du Canada

Objectif

Notre objectif était de connaître l'opinion du public sur les solutions de rechange à l'obtention du consentement pour chaque projet lorsqu'on utilise des renseignements personnels aux fins de la recherche en santé.

Nature de l'enquête

Nous avons mené une enquête téléphonique à composition aléatoire et à réponses fixes auprès de 1 230 adultes dans tout le Canada.

Mesures

Nous avons :

- mesuré les attitudes à l'égard de la protection de la vie privée et de la recherche médicale;
- mesuré la confiance en la capacité des différentes institutions à protéger la confidentialité des renseignements;
- déterminé le degré de consentement approprié pour l'utilisation en recherche des renseignements médicaux personnels nécessitant : l'examen du dossier médical, l'extraction automatique de renseignements du dossier médical électronique et la corrélation entre le niveau d'études ou le revenu et les données sur la santé. Les choix de réponse étaient les suivants :
    - Cette information ne devrait pas (du tout) être utilisée.
    - On devrait obtenir votre permission avant chaque utilisation.
    - On devrait obtenir votre permission générale et reprendre contact avec vous périodiquement.
    - On devrait obtenir votre permission générale une fois.
    - On devrait vous aviser quand les renseignements sont utilisés.
    - Il n'est pas nécessaire de savoir. Qu'on les utilise, tout simplement.

Pour chacun des scénarios, nous avons signalé que les identificateurs directs – nom adresse, numéro d'assurance maladie – ne seraient pas recueillis et qu'il serait par conséquent difficile, mais pas impossible, d'identifier les participants de nouveau.

Principales constatations

En général, nous avons constaté que la

Objective

Our objective was to determine public opinion on alternatives to project-specific consent for use of their personal information for health research.

Design

We conducted a fixed-response random-digit dialled telephone survey of 1230 adults across Canada.

Measurements

We measured:

- attitudes toward privacy and health research;
- trust in different institutions to keep information confidential; and
- preferred consent model for research use of one's own health information involving: medical record review; automated abstraction of information from the electronic medical record (EMR); and linking education or income with health data. Choices included:
    - This information should not be used (at all);
    - They should get your permission beforehand for each use;
    - They should get your general permission, with periodic re-contacting;
    - They should get your general permission once;
    - They should notify you of the use of the information;
    - There is no need to know. Just use it.

For each scenario we indicated that directly identifying information–name, address, health insurance number–would not be collected and that this would make it difficult, though not impossible, to re-identify individuals.

Major Findings:

In general, we found the Canadian public supports both research and privacy. Ninety per cent were either "somewhat" (53 per cent) or "very" concerned (37 per cent) if allowing health research made it difficult to control how their information was being used. At the same time, 88 per cent were either "somewhat" (56 per cent) or "very" concerned (32 per cent) if protecting people's rights to control access to their information made it difficult or impossible to conduct health research.

Support for using their health information was strongest for research that improves public health

population canadienne est pour la recherche, mais également pour la protection de la vie privée. Quatre-vingt-dix pour cent des répondants disent qu'ils seraient « assez » (53 %) ou « très » (37 %) préoccupés si le fait d'autoriser la recherche en santé implique qu'il soit difficile de contrôler la façon dont les renseignements fournis sont utilisés. Par ailleurs, 88 % des répondants disent qu'ils seraient « assez » (56 %) ou « très » (32 %) préoccupés si le fait de protéger le droit des personnes à contrôler l'accès à leurs renseignements personnels rend difficile ou impossible la recherche en santé.

Le taux d'acceptation de l'utilisation des renseignements médicaux personnels est plus élevé lorsque la recherche sert à améliorer la santé publique et la qualité des soins (85-89 %). Le soutien est considérablement plus faible si la recherche sert à des fins commerciales. Même dans le meilleur scénario – recherche permettant de surveiller les maladies transmissibles et d'améliorer la qualité des soins – une partie importante de la population (40-45 %) n'est que « plutôt » favorable à ces utilisations de leurs renseignements médicaux personnels. Nos observations et nos travaux antérieurs permettent de penser que cette acceptation dépend des fins visées et des utilisateurs des données, ainsi que des mesures de protection employées. (1;2)

Les répondants ont indiqué qu'ils accordent une confiance relativement élevée aux fondations de maladies, aux hôpitaux, aux chercheurs universitaires et aux organisations de collecte de données, comme Statistique Canada, pour protéger la confidentialité de leurs renseignements personnels. Entre 26 % et 35 % des répondants font grandement confiance à ces organismes, et entre 76 % et 81 % leur font « plutôt » ou « grandement » confiance. Il y a un degré de *méfiance* relativement élevé à l'égard de l'industrie des assurances (42 %), des entreprises pharmaceutiques (28 %) et du gouvernement (27 %).

Bien qu'ils soient favorables à l'utilisation de leurs renseignements personnels dans le cadre de la recherche, la majorité des répondants souhaitent toujours maintenir un certain degré de contrôle sur l'utilisation de cette information. Dans le cas de la recherche nécessitant l'extraction de données du dossier médical, la plupart des répondants (64 %) sont disposés à accepter des solutions de rechange moins strictes que l'habituel

and quality of care (85-89 per cent). Support was substantially lower if research was used for commercial purposes. Even in the "best case" scenarios–research to track communicable diseases and to improve quality of care–a substantial portion of the population (40-45 per cent) was only "somewhat" supportive of these uses of their personal health information. Our findings and previous work suggest that this support is dependent on the intended uses and users of the data, and on the safeguards applied. (1;2)

Respondents exhibited relatively high trust in disease-based foundations, hospitals, university researchers and data collection organizations such as Statistics Canada to keep their information confidential. From 26 to 35 per cent trusted these organizations a great deal, and 76 per cent to 81 per cent trusted them either somewhat or a great deal. There was a relatively high *distrust* of the insurance industry (42 per cent), drug companies (28 per cent) and government (27 per cent).

While supportive of research uses of their information, the majority of respondents still wished to maintain some level of control over use of their information. For research involving abstraction of data from the medical record, most (64 per cent) were willing to consider less stringent alternatives to conventional study-by-study consent, including a broad authorization for research uses (28 per cent), and notification with opt-out (24 per cent). Few (12 per cent) felt it was acceptable to use the information without prior permission or notification. On the other hand, very few (4 per cent) were completely opposed to use of their health information for research.

These findings are consistent with most other studies examining research use of personal information. (1;3-7) However, they are in stark contrast with a recent study by Barrett and colleagues who found a high acceptance (72 per cent) among U.K. residents of the practice of using personal information–including directly identifying information–without consent for a national cancer registry. (8) In part, this may be due to the framing of Barrett's question which gave only one option–use of this information without consent. Respondents were asked if they thought this was a breach of privacy. As well, respondents may have considered the cancer registry more like a public health *service* activity than research, which could affect perceptions of the acceptability of using this information. Finally, cancer itself may hold a special status in the public's mind, distinct from other health research.

consentement accordé pour une étude précise, notamment l'autorisation générale aux fins de recherche (28 %) et la notification avec option de non-participation (24 %). Peu de répondants (12 %) estiment qu'il soit acceptable d'utiliser les renseignements sans autorisation préalable ou sans notification. Par contre, très peu de répondants (4 %) sont totalement opposés à l'utilisation de leurs renseignements médicaux personnels aux fins de la recherche.

Ces constatations concordent avec celles de la plupart des autres études portant sur l'utilisation de renseignements personnels dans le cadre de la recherche. (1;3-7) Cependant, elles s'opposent d'une façon frappante à celles de l'étude récente menée par Barrett et ses collègues, qui ont constaté que les résidants du Royaume-Uni acceptaient dans une grande proportion (72 %) l'utilisation sans consentement des renseignements personnels – y compris des identificateurs directs – dans le cadre d'un registre national du cancer. (8) Cela est peut-être attribuable en partie à la formulation de la question de Barrett, qui n'offrait qu'une seule option : l'utilisation de cette information sans consentement. Les répondants devaient dire s'ils pensaient qu'il s'agissait là d'une atteinte à la vie privée. De plus, il est possible que les répondants aient jugé que le registre du cancer s'apparentait davantage à un service de santé publique qu'à une recherche, ce qui a pu modifier leurs perceptions quant à l'acceptabilité de l'utilisation de ces renseignements. Enfin, dans l'esprit du public, le cancer, en soi, pourrait avoir un statut spécial qui le distinguerait des autres domaines de recherche en santé.

Une question se pose toujours lorsque la recherche exige l'examen des dossiers médicaux : qui peut extraire les données du dossier médical? Selon nos constatations, le public est presque autant rassuré si cette tâche est effectuée par un assistant de recherche universitaire ou par une infirmière, et moins rassuré si cela est fait par un employé de bureau. Étant donné que le personnel médical manque de temps, un protocole commun devrait être établi pour qu'un assistant de recherche formé adéquatement ait accès, dans certaines limites, aux dossiers médicaux pour cette fin.

Limites de l'étude

Nous présentons une vue transversale de

An ongoing question in research requiring review of medical records is who may abstract information from the health record. Our findings suggest that the public is almost as comfortable with an academically-based research assistant collecting the information as they are with a nurse who works in the practice, and more comfortable than with a clerical person. Given clinical employees' limited time, a common protocol should be established that would allow a properly trained research assistant limited access to health records for this purpose.

Study Limitations

We present a cross-sectional view of public opinion at one point in time. It provides no information on how firmly the opinions were held–whether, if challenged, people would change their views. Survey respondents were better educated than the general public. Although our data did not find differences in response by education, other studies have noted that higher education is associated with increased privacy consciousness. (9;10) Therefore, our findings may present a slightly more restrictive attitude towards consent than exists in the general public.

Policy implications

Individuals differ substantially in the amount of control they would like to exercise over research uses of their personal health information. There is no one approach that satisfies even a simple majority of the population. However, the findings do suggest insufficient public support for across-the-board *assumed* or *deemed* consent for research uses of one's health information.

A logical conclusion would be to document individuals' consent choices for research and other secondary use of their information. However, there are several legal and technical challenges, including:

- the need for legal recognition of a broad authorization for future uses of one's personal information for research purposes;
- an appropriate repository to track consent choices throughout the health care system;
- safeguards and governance structures that would ensure that the consent choices of individuals are honoured; and
- an appropriate method of eliciting those consent choices and keeping them up-to-date.

In Canada, the architecture for a common electronic medical record for recording all treatments

l'opinion publique à un moment précis. Elle ne nous dit pas à quel point les opinions exprimées étaient fermes; si les répondants auraient changé d'avis s'ils avaient été contredits. Les répondants étaient plus instruits que la population en général. Même si nos données n'indiquent pas que les réponses diffèrent en fonction du niveau d'instruction, d'autres études ont mis en évidence qu'un niveau de scolarité plus élevé va de pair avec une sensibilisation accrue par rapport à la vie privée. (9;10) Par conséquent, nos constatations peuvent indiquer une attitude légèrement plus réservée par rapport au consentement que ce qu'on noterait dans la population en général.

Incidences sur les politiques

Les gens ont des opinions assez divergentes en ce qui concerne le degré de contrôle qu'ils voudraient exercer sur l'utilisation de leurs renseignements médicaux personnels à des fins de la recherche. Il n'y a aucune méthode qui satisfait même une majorité relative de la population. Cependant, les données recueillies indiquent un manque d'appui de la part du public pour un consentement *présumé* d'application générale aux fins de l'utilisation des renseignements médicaux personnels dans le cadre de la recherche.

La conclusion logique serait de réunir des données sur les choix des personnes en matière de consentement pour la recherche et les autres usages secondaires de leurs renseignements personnels. Il y a toutefois plusieurs difficultés juridiques et techniques, notamment :

- nécessité de la reconnaissance juridique d'une autorisation de portée générale pour les futures utilisations des renseignements personnels à des fins de recherche;
- organe d'archivage approprié pour pouvoir connaître les choix en matière de consentement dans tout le système de soins de santé;
- mesures de protection et structures de gouvernance qui garantiraient que les choix d'une personne en matière de consentement soient respectés;
- méthode appropriée pour connaître les choix en matière de consentement et tenir l'information à jour.

Au Canada, l'architecture d'un dossier médical électronique commun, dans lequel seraient

and diagnostics is being developed. (11) Provision could be made for recording consent choices for different uses of one's personal health information and for linkage of that information with other health-related information like income and education. Although, technically, it is already possible to restrict access to and uses of the data to reflect consent choices, the challenge comes with ensuring organizational compliance with those protocols. (12;13) In addition, there is no good mechanism at present for eliciting individuals' consent choices nor for ensuring those choices are up-to-date. Physicians and other health care providers are too busy to take it on. Nor would it be appropriate, given concerns over the potential for undue influence.

That leaves the ethics review of each research project to determine on a case-by-case basis whether the project requires individual consent. Concerns have been raised over ethics boards' inconsistent requirements, as well as institutional hurdles that go beyond the requirements of the law. (14;15) Recent guidance from the Canadian Institutes of Health Research should help harmonize these policies. (16)

Conclusion

The Canadian public supports health research and is open to alternatives to a conventional project-by-project consent. However, they do not wish to completely relinquish control over use of their personal health information. Any long term solution must take into account the variety of consent choices in order to maintain public confidence in the confidentiality of the information they share with their physicians. While the electronic medical record may play a role here, the outstanding challenge is how best to elicit individuals' consent preferences and keep them up-to-date. There are no easy solutions. Any solutions put forward should be vetted with the public.

Reference List

(1) Robling MR, Hood K, Houston H, Pill R, Fay J, Evans HM. Public attitudes towards the use of primary care patient record data in medical research without consent: a qualitative study. J Med Ethics 2004 Feb;30(1):104-9.

(2) Nair K, Willison D, Holbrook A, Keshavjee

consignés tous les traitements et diagnostics, est en voie d'élaboration. (11) Des dispositions pourraient être prises pour consigner les choix de consentement en vue des différents usages des renseignements médicaux personnels et du recoupement de cette information avec d'autres renseignements liés à la santé, comme le revenu et la scolarité. Bien que, techniquement parlant, il soit déjà possible de restreindre l'accès aux données et leur utilisation pour tenir compte des choix de consentement, ce qui est difficile, c'est de s'assurer que les organisations se conforment à ces protocoles. (12;13) En outre, il n'y a pas, à l'heure actuelle, de mécanisme adéquat pour connaître les choix des personnes en matière de consentement ni pour tenir cette information à jour. Les médecins et les autres fournisseurs de soins de santé sont trop occupés pour s'en charger. Il ne serait pas non plus approprié qu'ils assument cette responsabilité compte tenu de l'influence indue qu'ils pourraient exercer.

Il reste donc l'examen déontologique de chaque projet de recherche pour déterminer au cas par cas si le projet requiert le consentement individuel. Mais les exigences incohérentes des comités d'éthique et les obstacles institutionnels qui vont au-delà des dispositions législatives ont suscité des préoccupations. (14;15) Les directives récentes des Instituts de recherche en santé du Canada devraient faciliter l'harmonisation de ces politiques. (16)

Conclusion

La population canadienne est favorable à la recherche et est réceptive à des solutions de rechange à l'obtention du consentement pour chaque projet. Elle n'est cependant pas disposée à renoncer complètement à l'encadrement de l'utilisation des renseignements médicaux personnels. Toute solution à long terme doit tenir compte de la diversité des choix de consentement si on veut maintenir la confiance des gens à l'égard de la protection de la confidentialité des renseignements qu'ils communiquent à leurs médecins. Même si le dossier médical électronique peut jouer un rôle à cet effet, la difficulté qui persiste, c'est de trouver le meilleur moyen de connaître les préférences des personnes en matière de consentement et de tenir cette information à jour. Il n'y a pas de solutions faciles. Toutes les solutions proposées devraient être examinées soigneusement avec la population.

K. Patients' consent preferences regarding the use of their health information for research purposes: a qualitative study. J Health Serv Res Policy 2004 Jan;9(1):22-7.

(3) Willison DJ, Keshavjee K, Nair K, Goldsmith C, Holbrook AM, for the COMPETE investigators. Patients' consent preferences for research uses of information in electronic medical records: Interview and survey data. BMJ 2003;326:373-6.

(4) Kass NE, Natowicz MR, Hull SC, Faden RR, Plantinga L, Gostin LO, et al. The use of medical records in research: what do patients want?
14. J Law Med Ethics 2003;31(3):429-33.

(5) The Gallup Organization for Institute for Health Freedom. Public attitudes toward medical privacy. Available at: http://www.forhealthfreedom.org/Gallupsurvey/IHF-Gallup.pdf. Princeton, New Jersey: The Gallup Organization; 2000.

(6) Americans support online personal health records; patient privacy and control over their own information are crucial to acceptance. Available at: http://www.rwjf.org/newsroom/newsreleasesdetail.jsp?id=10369&gsa=1. The Robert Wood Johnson Foundation . 2005.

(7) Angus Reid Group. Canadians and the Confidentiality of Their Personal Health Information. 1999 Feb.

(8) Barrett G, Cassell JA, Peacock JL, Coleman MP. National survey of British public's views on use of identifiable medical data by the National Cancer Registry. BMJ 2006 May 6;332(7549):1068-72.

(9) Pan-Canadian Health Information Privacy and Confidentiality Framework. Ottawa: EKOS Research Associates Inc; 2004 Nov.

(10) Berger E. Attitudes to privacy, health records and interconnection: Implications for healthcare organizations. Hospital Quarterly 2002;5(4):40-5.

(11) Canada Health Infoway and Health Coun-

Liste de référence

(1) Robling MR, Hood K, Houston H, Pill R, Fay J, Evans HM. Public attitudes towards the use of primary care patient record data in medical research without consent: a qualitative study. J Med Ethics, février 2004;30(1):104-9.

(2) Nair K, Willison D, Holbrook A, Keshavjee K. Patients' consent preferences regarding the use of their health information for research purposes: a qualitative study. J Health Serv Res Policy, janvier 2004;9 (1):22-7.

(3) Willison DJ, Keshavjee K, Nair K, Goldsmith C, Holbrook AM, for the COMPETE investigators. Patients' consent preferences for research uses of information in electronic medical records: Interview and survey data. BMJ 2003;326:373-6.

(4) Kass NE, Natowicz MR, Hull SC, Faden RR, Plantinga L, Gostin LO, et al. The use of medical records in research: what do patients want? 14. J Law Med Ethics 2003;31(3):429-33.

(5) The Gallup Organization for Institute for Health Freedom. Public attitudes toward medical privacy. Lien Internet : http://www.forhealthfreedom.org/Gallupsurvey/IHF-Gallup.pdf. Princeton, New Jersey: The Gallup Organization; 2000.

(6) Americans support online personal health records; patient privacy and control over their own information are crucial to acceptance. Lien Internet : http://www.rwjf.org/newsroom/newsreleasesdetail.jsp?id=10369&gsa=1. The Robert Wood Johnson Foundation. 2005.

(7) Angus Reid Group. Canadians and the Confidentiality of Their Personal Health Information. Février 1999.

(8) Barrett G, Cassell JA, Peacock JL, Coleman MP. National survey of British public's views on use of identifiable medical data by the National Cancer Registry. BMJ, 6 mai 2006;332(7549):1068-72.

(9) Cadre pancanadien de protection de la confidentialité des renseignements personnels sur la santé. Ottawa: Santé Canada, janvier 2005 (sondage d'opinion public réalisé par Les Associés de recherche Ekos inc., automne 2004)

(10) Berger E. Attitudes to privacy, health records and interconnection: Implications for healthcare organizations. Hospital

cil of Canada. Beyond good intentions: Accelerating the electronic health record in Canada. Available at: http://www.infoway-inforoute.ca/Admin/Upload/Dev/Document/Conference%20Executive%20Summary_EN.pdf. 2006.

(12) Cavoukian A. Information and Privacy Commissioner Order H0-002 under the Ontario Personal Health Information Protection Act. Available at: http://www.ipc.on.ca/docs/ho-002.pdf. Ottawa: Information and Privacy Commissioner of Ontario; 2006.

(13) Willison DJ. Trends in collection, use and disclosure of personal information in contemporary health research: challenges for research governance. Health Law Review 2005;13(2-3):107-13.

(14) Willison DJ, Kapral MK, Peladeau P, Richards JA, Fang J, Silver FL. Variation in recruitment across sites in a consent-based clinical data registry: lessons from the Canadian Stroke Network. BMC Medical Ethics 2006;7(1):E6.

(15) Shalowitz D, Wendler D. Informed consent for research and authorization under the Health Insurance Portability and Accountability Act Privacy Rule: an integrated approach. Ann Intern Med 2006 May 2;144(9):685-8.

(16) Canadian Institutes of Health Research Privacy Advisory Committee. CIHR Best Practices for Protecting Privacy in Health Research - September 2005. Available at: http://www.cihr-irsc.gc.ca/e/documents/pbp_sept2005_e.pdf. Ottawa: Public Works and Government Services Canada; 2005.

Quarterly 2002;5(4):40-5.
(11) Inforoute Santé du Canada et Conseil canadien de la santé. Au-delà des bonnes intentions : accélérer le dossier de santé électronique au Canada. Lien Internet : http://www.infoway-inforoute.ca/Admin/Upload/Dev/Document/Conference_Executive_Summary_FR.pdf. 2006.

(12) Cavoukian A. Décret H0-002 de la commissaire à l'information et à la protection de la vie privée en vertu de la *Loi sur la protection des renseignements personnels sur la santé.* Lien Internet : http://www.ipc.on.ca/docs/ho-002.pdf. Ottawa: Commissaire à l'information et à la protection de la vie privée de l'Ontario; 2006.
(13) Willison DJ. Trends in collection, use and disclosure of personal information in contemporary health research: challenges for research governance. Health Law Review 2005;13(2-3):107-13.
(14) Willison DJ, Kapral MK, Peladeau P, Richards JA, Fang J, Silver FL. Variation in recruitment across sites in a consent-based clinical data registry: lessons from the Canadian Stroke Network. BMC Medical Ethics 2006;7(1):E6.
(15) Shalowitz D, Wendler D. Informed consent for research and authorization under the Health Insurance Portability and Accountability Act Privacy Rule: an integrated approach. Ann Intern Med, 2 mai 2006;144(9):685-8.
(16) Instituts de recherche en santé du Canada - Comité consultatif sur la protection de la vie privée. Pratiques exemplaires des IRSC en matière de protection de la vie privée dans la recherche en santé, septembre 2005. Lien Internet : http://www.cihr-irsc.gc.ca/e/documents/pbp_sept2005_f.pdf. Ottawa: Travaux publics et Services gouvernementaux Canada, 2005.

29E CONFÉRENCE INTERNATIONALE DES COMMISSAIRES
À LA PROTECTION DES DONNÉES ET DE LA VIE PRIVÉE

LES HORIZONS DE LA PROTECTION DE LA VIE PRIVÉE

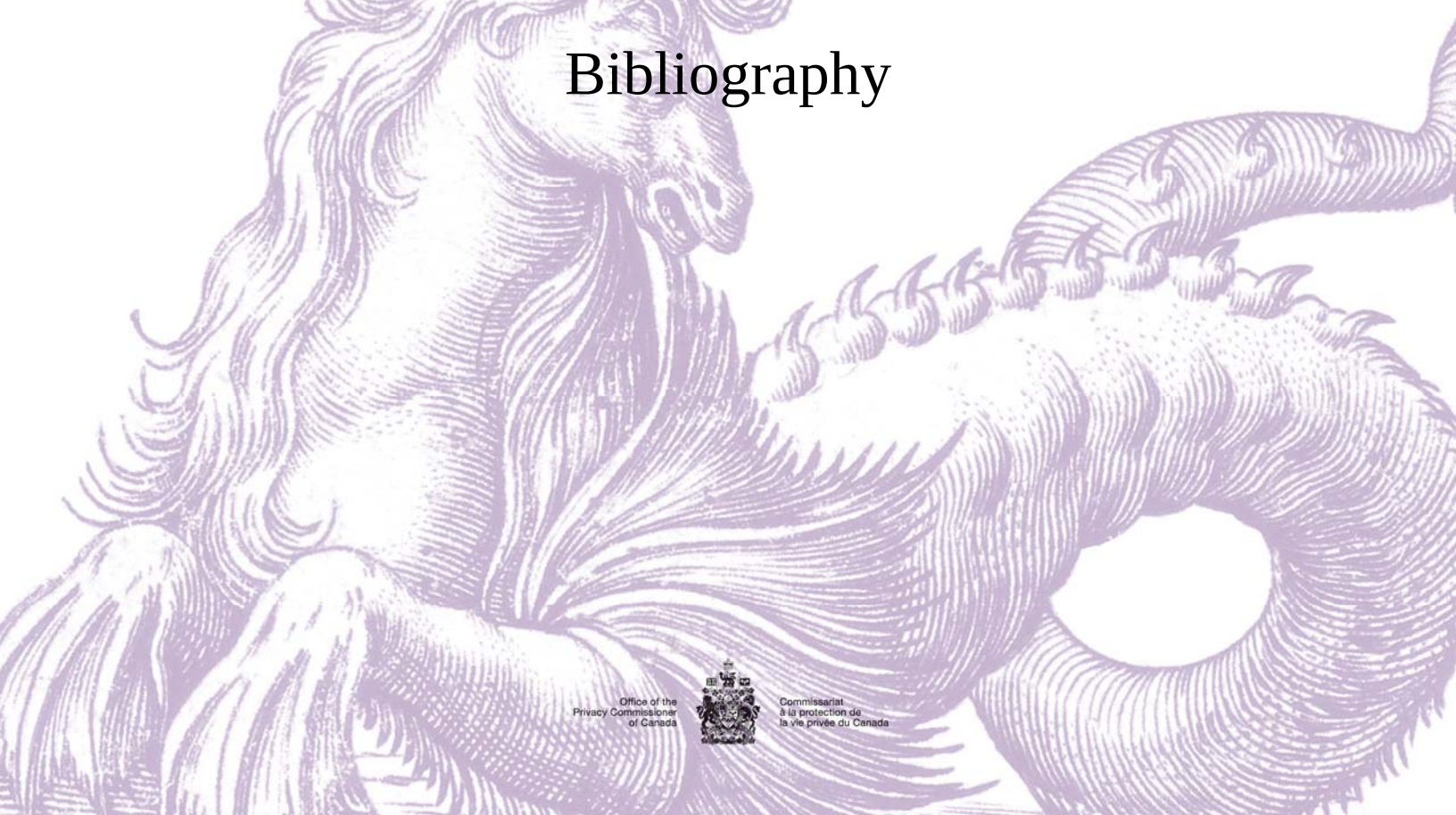
TERRA INCOGNITA

PRIVACY HORIZONS

29TH INTERNATIONAL CONFERENCE OF
DATA PROTECTION AND PRIVACY COMMISSIONERS

Bibliographie

Bibliography



Office of the Privacy Commissioner of Canada
Commissariat à la protection de la vie privée du Canada

Articles importants

Evaluating Common De-Identification Heuristics for Personal Health Information

Par Khaled El Emam

RÉSUMÉ

Contexte
L'adoption croissante des dossiers médicaux électroniques entraîne une demande accrue pour l'utilisation de données cliniques électroniques dans le cadre de la recherche par observation. L'une des exigences répandues des comités sur l'éthique pour autoriser l'usage secondaire des renseignements personnels sur la santé dans le cadre de recherches par observation est la dépersonnalisation des données. Des méthodes heuristiques de dépersonnalisation sont fournies dans la règle de protection de la vie privée de la *Health Insurance Portability and Accountability Act*, dans les lignes directrices sur la protection de la vie privée à l'intention des organismes de financement et des associations professionnelles et dans la pratique courante.

Objectif
L'étude visait à évaluer si les méthodes heuristiques de dépersonnalisation les plus répandues présentent des risques suffisamment minimes relativement à la repersonnalisation par couplage de dossiers et à déterminer s'il s'agit d'un risque stable pour toutes les tailles d'échantillon et tous les ensembles de données.

El Emam, K., Jabbouri, S., Sams, S., Drouet, Y. et Power, M. « Evaluating common de-identification heuristics for personal health information », *Journal of Medical Internet Research*, volume 8, p. 28, 2006.
https://tspace.library.utoronto.ca/html/1807/9799/jmir.html

* * *

Article 29 : Avis 4/2007 sur le concept de données à caractère personnel
Par le Groupe de travail sur la protection des données de l'UE

Introduction
Le Groupe de travail est conscient du besoin de mener une analyse poussée du concept de données à caractère personnel. L'information sur

Important Papers

Evaluating Common De-Identification Heuristics for Personal Health Information
By: Khaled El Emam

ABSTRACT

Background:
With the growing adoption of electronic medical records, there are increasing demands for the use of this electronic clinical data in observational research. A frequent ethics board requirement for such secondary use of personal health information in observational research is that the data be de-identified. De-identification heuristics are provided in the Health Insurance Portability and Accountability Act Privacy Rule, funding agency and professional association privacy guidelines, and common practice.

Objective:
The aim of the study was to evaluate whether the re-identification risks due to record linkage are sufficiently low when following common de-identification heuristics and whether the risk is stable across sample sizes and data sets.

K. El Emam, S. Jabbouri, S. Sams, Y. Drouet, and M. Power, "Evaluating common de-identification heuristics for personal health information," Journal of Medical Internet Research, vol. 8, p. e28, 2006.
https://tspace.library.utoronto.ca/html/1807/9799/jmir.html

* * *

Article 29: Opinion on the Concept of Personal Data
By: EU Data Protection Working Party

Introduction:
The Working Party is aware of the need to conduct a deep analysis of the concept of personal data. Information about current practice in EU Member States suggests that there is some uncertainty and some diversity in practice among Member States as to important aspects of this concept which may affect the proper functioning of the existing data protection framework in different contexts. The outcome of this analysis of a central element for the application and interpretation of data protection rules is bound to have a profound impact on a number of important issues, and will be particularly relevant for topics

les pratiques actuelles des États membres de l'UE laisse croire qu'il existe une certaine incertitude et une certaine diversité quant aux pratiques en vigueur parmi les États membres en ce qui a trait à des aspects importants de ce concept; cela peut avoir des répercussions sur le fonctionnement adéquat du cadre actuel de protection des données dans différents contextes. Les conclusions de cette analyse, liées à un élément central de la mise en application et de l'interprétation de règles de protection des données, auront forcément des répercussions marquées sur bon nombre d'enjeux importants et seront tout particulièrement pertinentes pour des thèmes comme la gestion de l'identité dans le contexte du gouvernement et de la santé en ligne, ainsi que dans le contexte du recours à des dispositifs d'identification par radiofréquence (IRF).

Le présent avis du Groupe de travail a pour objectif d'en arriver à une interprétation commune du concept des données à caractère personnel, des situations dans lesquelles une loi nationale de protection des données devrait être mise en application, et de la manière de le faire. L'élaboration d'une définition commune du concept de données à caractère personnel équivaut à définir les limites de la portée des règles de la protection des données. Un des corollaires de ce travail consiste à fournir des conseils sur la manière d'appliquer les règles nationales de protection des données dans certaines catégories de situations qui ont cours dans l'ensemble de l'Europe, contribuant ainsi à l'application uniforme de telles normes, ce qui constitue une fonction essentielle du Groupe de travail « Article 29 ».

Groupe de travail « Article 29 » sur la protection des données de l'UE. « Avis 4/2007 sur le concept de données à caractère personnel », adopté le 20 juin 2007; http://ec.europa.eu/justice_home/fsj/privacy/docs/wpdocs/ 2007/wp136_fr.pdf.

*** * ***

Identifiability in Genomic Research
Par William Lowrance et Francis Collins

La recherche en génétique peut désormais générer facilement des données qui couvrent des portions considérables du génome humain, à un degré de détails unique à une personne en

such as Identity management in the context of e-Government and e-Health, as well as in the RFID context.

The objective of the present opinion of the Working Party is to come to a common understanding of the concept of personal data, the situations in which national data protection legislation should be applied, and the way it should be applied. Working on a common definition of the notion of personal data is tantamount to defining what falls inside or outside the scope of data protection rules. A corollary of this work is to provide guidance on the way national data protection rules should be applied to certain categories of situations occurring Europe-wide, thus contributing to the uniform application of such norms, which is a core function of the Article 29 Working Party.

EU Data Protection Working Party, "Re: Article 29, "Opinion on the concept of personal data," adopted 20 June 2007; http://ec.europa.eu/justice_home/fsj/privacy/docs/wpdocs/ 2007/wp136_en.pdf.

*** * ***

Identifiability in Genomic Research
By William Lowrance and Francis Collins

"Genomic research can now readily generate data that cover significant portions of the human genome at levels of detail unique to individuals. Data can now be categorized with respect to disease-related genes and linked to clinical, family, and social data. Identifiability, the potential for such data to be associated with specific individuals, is therefore a pivotal concern."

William W. Lowrance and Francis S. Collins, "Identifiability in genomic research," *Science* 317, 600−602 (August 3, 2007)

*** * ***

How (Not) to Protect Genomic Data Privacy in a Distributed Network: Using Trail Re-identification to Evaluate and Design Privacy Protection Systems
By: Bradley Malin and Latanya Sweeney

The increasing integration of patient-specific genomic data into clinical practice and research

particulier. Il est maintenant possible de classifier les données en fonction de gènes propres à une maladie et de les lier à des données cliniques, familiales et sociales. L'identifiabilité, c'est-à-dire le potentiel d'associer de telles données à une personne particulière, devient donc une préoccupation essentielle.

Lowrance, William W., et Collins, Francis S. « Identifiability in genomic research », *Science* 317, pages 600 à 602 (3 août 2007).

* * *

How (Not) to Protect Genomic Data Privacy in a Distributed Network: Using Trail Re-identification to Evaluate and Design Privacy Protection Systems
Par Bradley Malin et Latanya Sweeney

L'intégration croissante de données génétiques propres à un patient dans la pratique et la recherche cliniques soulève de graves préoccupations en matière de protection de la vie privée. Divers systèmes ont été proposés pour protéger la vie privée, puisqu'ils suppriment ou chiffrent (à l'aide de pseudonymes) les renseignements permettant d'identifier les patients de manière explicite, comme le nom ou le numéro d'assurance sociale. Même si les responsables de ces systèmes prétendent qu'ils protègent la communication de l'identité, il y a peu de preuves formelles. Dans cet article, nous analysons l'érosion de la vie privée lorsque des données génétiques, que ce soit sous forme de pseudonymes ou de données prétendument anonymes, sont communiquées dans un environnement de soins de santé réparti. Nous examinons plusieurs algorithmes, portant le nom collectif de *RE-Identification of Data In Trails (REIDIT)*, qui permettent d'établir des liens entre les données génétiques et des personnes nommées dans des dossiers disponibles au public, en tirant profit des caractéristiques uniques des schèmes de visite patient-endroit. Les preuves algorithmiques de nouvelle personnalisation ne sont pas négligeables ni ne résultent d'occurrences isolées singulières. Nous proposons qu'il soit possible d'appliquer de telles techniques en guise de test des capacités de protection de la vie privée des systèmes.

* * *

MALIN, Bradley et SWEENEY, Latanya. « How raises serious privacy concerns. Various systems have been proposed that protect privacy by removing or encrypting explicitly identifying information, such as name or social security number, into pseudonyms. Though these systems claim to protect identity from being disclosed, they lack formal proofs. In this paper, we study the erosion of privacy when genomic data, either pseudonyms or data believed to be anonymous, is released into a distributed healthcare environment. Several algorithms are introduced, collectively called RE-Identification of Data In Trails (REIDIT), which link genomic data to named individuals in publicly available records by leveraging unique features in patient-location visit patterns. Algorithmic proofs of re-identification is neither trivial nor the result of bizarre isolated occurrences. We propose that such techniques can be applied as system tests of privacy protection capabilities.

Bradley Malin, Latanya Sweeney, How (Not) to Protect Genomic Data Privacy in a Distributed Network: Using Trail Re-identification to Evaluate and Design Privacy Protection Systems, April 2004, http://reports-archive.adm.cs.cmu.edu/anon/isri2004/CMU-ISRI-04-115.pdf

* * *

Key Web site:

American Statistical Association's Privacy, Confidentiality, and Data Security web site
http://www.amstat.org/comm/cmtepc/index.cfm?fuseaction=main

This site consists of a list of links to privacy and confidentiality-related web sites. The list is divided into the following 9 major topic areas:

I. Privacy, Confidentiality, and Data Dissemination Guidelines for Government Agencies and International Organizations
II. Statistical Methods for Privacy, Confidentiality, and Disclosure Protection
III. Human Subjects' Protection and Institutional Review Board (IRB)
IV. Health Care, Bioethics, and Personal Health Information
V. Topics in Education
VI. Topics in Finance
VII. Ethics, Principles, and Standards
VIII. Legal and Regulatory Sites
IX. Training Opportunities (e.g. fellowships and academic programs)

(Not) to Protect Genomic Data Privacy in a Distributed Network: Using Trail Re-identification to Evaluate and Design Privacy Protection Systems », avril 2004, http://reports-archive.adm.cs.cmu.edu/anon/isri2004/CMU-ISRI-04-115.pdf

* * *

Site Web important

Site Web *American Statistical Association's Privacy, Confidentiality, and Data Security*
http://www.amstat.org/comm/cmtepc/index.cfm?fuseaction=main

Ce site renferme une liste de liens vers des sites Web consacrés à la confidentialité et à la protection de la vie privée. La liste est séparée en neuf principaux thèmes :

I. *Privacy, Confidentiality, and Data Dissemination Guidelines for Government Agencies and International Organizations* (Lignes directrices sur la protection de la vie privée, la confidentialité et la communication de données, à l'intention des organismes gouvernementaux et internationaux)
II. *Statistical Methods for Privacy, Confidentiality, and Disclosure Protection* (Méthodes statistiques relatives à la protection de la vie privée, à la confidentialité et à la communication de données)
III. *Human Subjects' Protection and Institutional Review Board (IRB)* (Protection des sujets humains et comité d'examen en établissement)
IV. *Health Care, Bioethics, and Personal Health Information* (Soins de santé, bioéthique et renseignements personnels sur la santé)
V. *Topics in Education* (Éducation)
VI. *Topics in Finance* (Finances)
VII. *Ethics, Principles, and Standards* (Éthique, principes et normes)
VIII. *Legal and Regulatory Sites* (Sites juridiques et réglementaires)
IX. *Training Opportunities* (Formation – p. ex., bourses de recherche et programmes d'études)