Ethical and practical challenges of sharing data from genome-wide association studies: The eMERGE Consortium experience

Amy L. McGuire,[1,11] Melissa Basford,[2] Lynn G. Dressler,[3] Stephanie M. Fullerton,[4] Barbara A. Koenig,[5] Rongling Li,[6] Cathy A. McCarty,[7] Erin Ramos,[6] Maureen E. Smith,[8] Carol P. Somkin,[9] Carol Waudby,[7] Wendy A. Wolf,[10] and Ellen Wright Clayton[2]

[1]Center for Medical Ethics and Health Policy, Baylor College of Medicine, Houston, Texas 77030, USA; [2]Vanderbilt University Medical Center, Nashville, Tennessee 37232, USA; [3]University of North Carolina Eshelman School of Pharmacy, Chapel Hill, North Carolina 27599, USA; [4]University of Washington School of Medicine, Seattle, Washington 98195, USA; [5]Mayo Clinic, Rochester, Minnesota 55905, USA; [6]National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA; [7]Marshfield Clinic Research Foundation, Marshfield, Wisconsin 54449, USA; [8]Center for Genetic Medicine, Northwestern University, Chicago, Illinois 60611, USA; [9]Kaiser Permanente, Oakland, California 94611, USA; [10]Division of Genetics, Children's Hospital Boston, Boston, Massachusetts 02115, USA

In 2007, the National Human Genome Research Institute (NHGRI) established the Electronic MEdical Records and GEnomics (eMERGE) Consortium (www.gwas.net) to develop, disseminate, and apply approaches to research that combine DNA biorepositories with electronic medical record (EMR) systems for large-scale, high-throughput genetic research. One of the major ethical and administrative challenges for the eMERGE Consortium has been complying with existing data-sharing policies. This paper discusses the challenges of sharing genomic data linked to health information in the electronic medical record (EMR) and explores the issues as they relate to sharing both within a large consortium and in compliance with the National Institutes of Health (NIH) data-sharing policy. We use the eMERGE Consortium experience to explore data-sharing challenges from the perspective of multiple stakeholders (i.e., research participants, investigators, and research institutions), provide recommendations for researchers and institutions, and call for clearer guidance from the NIH regarding ethical implementation of its data-sharing policy.

Although data sharing among researchers has not always been embraced, its importance is increasing with the advent of new technological approaches requiring large data sets for analysis. To facilitate broad data sharing from genome-wide association studies (GWAS), the National Institutes of Health established the database of Genotypes and Phenotypes (dbGaP) (Mailman et al. 2007) and an accompanying GWAS Data-Sharing Policy (National Institutes of Health 2007), which strongly encourages submission of data from NIH-funded GWAS into dbGaP. As of November 2010, one hundred twenty studies have submitted data into dbGaP, but little is known about the experiences and perspectives of stakeholders involved in the process.

In late 2007, the National Human Genome Research Institute (NHGRI) established the Electronic MEdical Records and GEnomics (eMERGE) Consortium (www.gwas.net) to develop, disseminate, and apply approaches to research that combine DNA biorepositories with electronic medical record (EMR) systems for large-scale, high-throughput genetic research (McCarty et al. 2011a). The Consortium comprises five sites: Group Health Cooperative/University of Washington, Marshfield Clinic, Mayo Clinic, Northwestern University, and Vanderbilt University. The project includes a community consultation and ethics component to study the unique ethical, legal, and social implications (ELSI) of EMR-coupled biobanks (Clayton et al. 2010). A major focus of ELSI research within the eMERGE Consortium has been to assess the challenges of data sharing, as it is practiced both within the Consortium and through the database of Genotypes and Phenotypes (dbGaP), from the perspective of stakeholders involved in the process (research subjects, genome scientists, and institutions). This paper was written by a subcommittee of the Consent and Community Consultation (C&CC) Working Group, which includes investigators from each eMERGE site and NIH Project Leaders, as well as outside experts on the ethical and policy implications of broad data sharing. This paper describes the eMERGE experience with data sharing, presents several challenges to such data sharing from the perspective of study investigators, and summarizes participant perspectives on data sharing from site-specific studies. The purpose of this paper is to illustrate what data sharing actually entails and the ethical and practical challenges of implementing established data-sharing policies from within a multisite Consortium.

eMERGE data sharing

One of the goals of eMERGE is to develop methods for extracting phenotype data from EMR systems for use in genome-wide association studies (GWAS) that could, with minimal effort, be implemented in various institutions and data systems. Work toward this goal necessitated considerable collaboration between sites as each implemented and tested algorithms developed by other sites. Initially, data sharing in eMERGE was only anticipated between individual sites on an ad hoc basis, as early collaborative

[11]Corresponding author.
E-mail amcguire@bcm.edu.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.120329.111.

efforts developed pairwise. One site would test the implementation of an electronic phenotype algorithm from another site. Such efforts proved to be successful, and sites realized that cohort size could be rapidly expanded by using subjects from other Consortium sites. For example, Vanderbilt actively worked with Northwestern to implement and validate Northwestern's Type 2 diabetes phenotype algorithm and were then able to contribute over 2000 primarily African-American research subjects from Vanderbilt to the Northwestern study (Table 1).

To maximize these collaborations between sites, participating institutions had to develop Data Use Agreements (DUA) in order to share de-identified research data, including the Health Insurance Portability and Accountability Act (HIPAA)-defined limited data sets, with other sites within the Consortium. A DUA is a legal agreement between institutions, in this case, signed by the investigators and the designated institutional officials from each of the data-sharing organizations, that establishes what data may be shared, the ways in which the information in the data set may be used, and how the data will be protected. The DUA generally describes the research project, types of data elements to be shared, and who (or what classes of people, e.g., researchers and project staff) have access to the specific data.

This early stage of eMERGE utilized site-specific and project-specific DUAs, enabling data to be shared between pairs of institutions. Electronic phenotyping efforts were remarkably successful, and all site-led studies included data from more than one institution, thus increasing sample size and study power. The Consortium quickly realized the strength and uniqueness of its combined data set of over 17,000 GWAS subjects and the benefits of sharing data across all sites to further increase the power of each study and to provide viable replication data for promising associations (Fig. 1). Just as quickly, it became apparent that pairwise DUAs constrained the ability of the Consortium to function effectively, as sharing between more than two institutions became a necessity for each new phenotyping project, requiring the development of multiple DUAs for each study.

Further, six phenotypes were chosen for network-wide studies that would leverage EMR data and genotype data from samples originally analyzed for primary phenotypes. To enable these analyses, the Administrative Coordinating Center (ACC), funded within the grant and hosted by Vanderbilt University Medical Center, facilitated the sharing of genotype and phenotype data across all sites (see Fig. 1). For the network phenotypes, it was agreed that the ACC would centralize phenotypic and genomic data for network studies and facilitate the efficient flow of data among all sites. To enable this activity, the ACC worked with representatives from each site and developed a single DUA and Memorandum of Understanding with language and content that was agreeable to all Consortium members. This agreement allowed data to be shared between each site and with the ACC and also, critically, allowed the ACC to share all network data, including the full genomic data set, with all sites, thus becoming a data hub.

All eMERGE sites also share data with the broader research community by submitting phenotype and genotype data to

Table 1. Characteristics of eMERGE biobank populations and phenotypes for GWAS

Institution	Biobank population	Biobank size and demographics	Ongoing participant interactions	Primary GWAS phenotypes	Other sites contributing samples for genotyping
Group Health Cooperative (Seattle, WA)	Disease specific: Adult Changes in Thought (ACT) Study cohort (Kukull et al. 2002); source of cases and controls randomly sampled from HMO and not demented at time of enrollment	~4000 ACT participants Age 65+ 96% European ancestry	Yes, through bi-annual in-person visits, quarterly newsletters, birthday cards	Alzheimer's disease (n = 3390)	Marshfield Clinic; Vanderbilt University
Marshfield Clinic (Marshfield, WI)	Broad population: Personalized Medicine Research Project (McCarty et al. 2005); population-based ascertainment from Marshfield Clinic catchment area	20,000 participants Age 18+ 98% European ancestry	Yes, through three newsletters per year and as needed for specific studies	HDL, cataract (n = 3968)	None
Mayo Clinic (Rochester, MN)	Disease specific: Cases identified from noninvasive vascular lab database; controls identified from the Cardiovascular Health Clinic (Kullo et al. 2010)	1641 cases and 1604 controls Age: mean 66 +/− 11 yr, cases; 61 +/− 8 yr, controls 96% European ancestry	No	Peripheral Arterial Disease (n = 3335)	None
Northwestern University (Chicago, IL)	Broad population: NUgene Project (Wolf et al. 2003); ascertained from clinic- and hospital-based population	~10,000 participants Age 18+ 70% European ancestry 12% AA 8% Hispanic	No	Type 2 diabetes (n = 3498)	Vanderbilt University
Vanderbilt University (Nashville, TN)	Broad population: BioVU (Roden et al. 2008); use of discarded blood/nonhuman subjects linked to EMRs	>100,000 samples All ages 70% European ancestry 10% AA	N/A	QRS duration (n = 3192)	Northwestern University

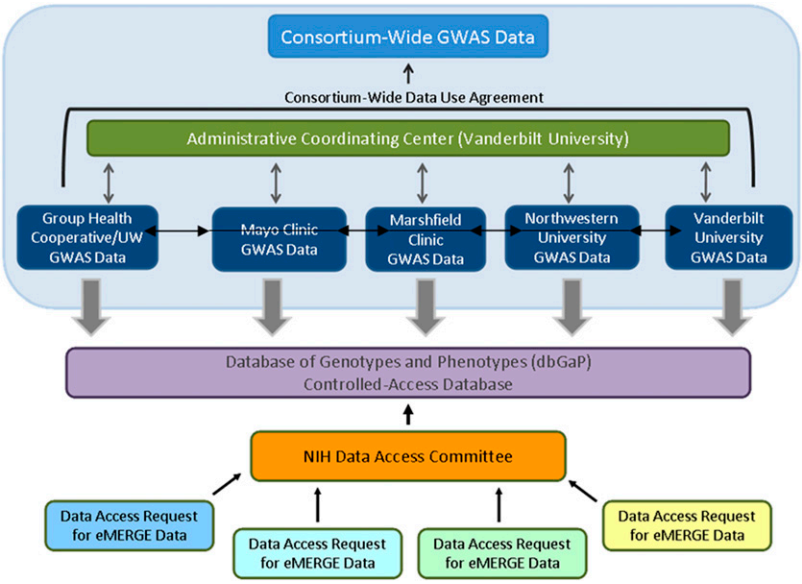

Figure 1. Data sharing for eMERGE is composed of three central tenets: (1) sharing of research data, including genomic data, between sites; (2) sharing of research data among all sites and the Administrative Coordination Center (ACC) with the facilitation of data flow enabled by the ACC or any particular site; and (3) sharing of data with secondary investigators through external NIH databases, namely dbGaP.

dbGaP for each GWAS conducted. Data submission is largely the responsibility of each site, although the ACC aggregated and submitted all data for the larger Consortium projects. Regardless of whether the data are submitted to dbGaP by the ACC for network-wide studies or by individual sites for site-specific GWAS, all subjects are identified by their site or study of origin, and the data maintains any restrictions dictated by the original informed-consent agreement.

All eMERGE sites have shared their primary study data with dbGaP. The Consortium continues to evaluate systematically various risks of re-identification of participants based on the types of data submitted to external databases like dbGaP and the policies governing those databases. To date, eMERGE researchers have developed a framework to understand events that might precipitate re-identification, including intentional attack (Benitez and Malin 2010; Malin et al. 2010) and processes to mitigate this risk (Malin et al. 2011). Researchers are also working to improve the processes for submitting data, including collaborative submissions, and to expand the various file types for submitting different classes of data, such as medications, reimbursement codes, and vital signs. In addition, the eMERGE experience suggests the need for shorter, more simplified DUAs that are easy to implement, provide clear safeguards for data, and can be expanded to include new partners joining the Consortium. Such a document is currently being pilot tested within eMERGE.

Data-sharing challenges

The Request for Proposals for the eMERGE Consortium was published in March 2007, and funding began in September 2007, one year after dbGaP was established and one month after the notice for the NIH Data Sharing Policy for Genome-Wide Association Studies was first published in the federal register. Although the NIH policy became effective only for competing applications and proposals submitted after January 25, 2008, the Request for Applications (RFA) describing the eMERGE program indicated that eMERGE investigators and institutions would be expected to be in compliance with the policy prior to funding. eMERGE investigators faced several challenges as they sought to meet both new expectations for broad data sharing through dbGaP and the demands of membership in the NIH Consortium. A survey of all site PIs, plus the experience of the authors who participated in decisions in real time, reveal how data-sharing obligations were balanced with institutional commitments to research participants recruited at each site. Researchers at each site encountered different challenges specific to the unique nature of the institutional and community context. Most of the challenges researchers encountered can be loosely categorized as administrative (or bureaucratic) challenges and challenges in honoring ethical obligations to study participants and their communities. However, these categories intertwine as administrative requirements are often put in place to try to enforce or promote ethically responsible conduct (Table 2).

As will be the case with any multisite consortium, harmonization across sites was a major challenge within eMERGE. Each site had their own study protocol, and the Institutional Review Board (IRB) at each site requested conditions or modifications based on institutional policy and local concerns. Consent forms also varied greatly. Since eMERGE was utilizing existing samples and clinical data from each site, some sites had multiple consent forms that were updated and changed over time. None of the consent documents specifically anticipated this type of research use of the samples/data or the broad data sharing required by the NIH data-sharing policy. In some cases, inconsistent institutional policies resulted in disparate treatment of subjects. For example, despite the fact that four of the five eMERGE sites originally obtained broad consent for the collection and subsequent use of samples,

Table 2. Data-sharing challenges and policy considerations

Use of archived specimens creates special issues
Specimens collected over time, using multiple consents, with evolving revisions
Lack of consistency across consortium sites in IRB determination for need for re-contact and new consent for GWAS, federal data sharing
Balancing compliance with NIH GWAS data-sharing policy with site-specific responsibility for ethical review and data submission to dbGaP
Varying levels of institutional preparedness for addressing GWAS data sharing
Different existing policies for sharing research subject level data outside the institution
Lack of harmonization of policies across sites for reviewing data and certifying compliance with GWAS data-sharing policy
Inconsistent policies within an institution
Missed opportunity for cross-site learning

the institutional review process at one of these sites resulted in a requirement that investigators re-consent participants specifically for GWAS data sharing (Ludman et al. 2010). The IRB at the three other sites determined that the broad consent originally obtained from participants allowed data sharing. The last site did not require informed consent because the design of their biobank was found to be consistent with nonhuman subjects research. Their model uses blood remaining from routine clinical care linked to de-identified clinical data from the EMR (Roden et al. 2008). In this model, patients are informed about the program when they seek care and are informed that they may opt out of the program at any time.

Inconsistency in IRB review and demands has long been a source of frustration for investigators (and IRBs alike) and is seen as a barrier to conducting multisite clinical trials (Menikoff 2010). The problem is exacerbated by a national policy that expects broad data sharing across studies, while placing responsibility for ethical review and coordination of the administrative aspects of data submission with institutions and their human subjects protection programs (e.g., IRBs or Privacy Boards). This intentionally allows for community input but also invariably leads to variation in policies and practices and tensions between federal and local requirements that can be confusing for investigators (and IRBs) collaborating within a large consortium.

Institutions conducting GWAS must also develop new policies and processes in order to comply with the NIH data-sharing policy. For example, the policy expects that each institution provide certification that it approves submission of the data to dbGaP. In order to sign such an agreement, the submitting institution's IRB and/or Privacy Board must review and certify that the investigator's plan for de-identifying data sets prior to submission is consistent with NIH policy, ensure that the uses and exclusions of the data (i.e., data use limitations) are in accordance with what participants agreed to during the informed consent process, and certify that the genotype and phenotypes were collected in a manner consistent with federal regulations (U. S. Dept. of Health and Human Services 2009).

Little cross-site learning or standardization took place within the eMERGE Consortium when completing the institutional certification due to variability in how prepared each submitting institution was to comply with federal policy and how they ultimately chose to review and certify compliance with the NIH GWAS data-sharing policy. At one site, the IRB had already developed a review process and a template for institutional certification. At another, no process was in place, but the IRB and institutional officials worked together through the experience to develop a process that was eventually incorporated into all electronic project submissions for IRB approval. At several other institutions, there were no formal processes in place, but investigators worked with their IRBs and/or institutional officials to obtain institutional certification.

The lack of institutional preparedness to comply with changing federal policy may reflect conflict with the ethical values underlying how different sites treat data collected from the community. For example, some sites historically have established a close and trusting relationship with the community, and there is a strong tradition of never allowing subject-level information outside of institutional walls. This must change for the institution to conduct large NIH-funded GWAS or to participate in national projects like eMERGE as long as the NIH continues to require broad data sharing. However, some institutional officials and individual investigators are reluctant to change their practices and share data more broadly because they believe it would be inconsistent with community expectations and desires and would, therefore, pose a potential threat to public trust.

Complying with data-sharing policies can also make it difficult to respect the wishes of participants who prefer more limited data sharing or would prefer that oversight of their samples and data be more tightly controlled by the institution collecting samples and data. Effective implementation of data-sharing policies should promote scientific progress while protecting participant interests and preserving public trust. As part of the implementation process, each eMERGE site solicited input from participants and community members in order to anticipate and respond effectively to ethical challenges raised by this sort of data sharing.

Stakeholder perspectives on data sharing

Each eMERGE site conducted independent community consultation activities and studies of stakeholders' (largely research participants' and potential participants') views. As summarized in Table 3, activities ranged from notification of planned data sharing via a study newsletter (Marshfield), consultation with study Community Advisory Boards (Marshfield, Mayo, Northwestern, Vanderbilt), solicitation of views as part of a community-based deliberative democracy exercise (Mayo), interviews with study participants and refusers (Mayo), focus groups with participants and potential participants (Group Health/UW, Marshfield, Northwestern), and surveys of potential participants (Vanderbilt) and participants (Group Health/UW). While there was no formal attempt to coordinate either the content of the information provided to stakeholders or the specific questions posed, all sites did discuss the NIH GWAS data-sharing policy and/or the likely deposition of data into the national data repository dbGaP, as well as security precautions taken to protect the confidentiality of shared data. Four of five sites also asked about what informed-consent requirements should be implemented for GWAS-related data sharing, as well as stakeholders' attitudes with respect to with whom (investigators, organizations) data might be shared (Table 3). Participant re-consent for dbGaP deposition, mechanisms to ensure third-party compliance with local oversight standards, and withdrawing from research once data were shared were each only addressed by one or two sites, so we have a less comprehensive picture of stakeholder perspectives on these topics.

Limited data collected as part of our community consultation activities suggests that participants and potential participants hold generally favorable views on sharing of genetic and linked health data in order to support and enable clinical research. A majority [61% (n = 27/44)] of research participants interviewed at the Mayo Clinic expressed neutral reactions to the sharing of research data (including sharing of data with federal entities), while 30% (n = 13/44) expressed negative views, and 9% (n = 4/44) were positive. Participants in focus groups conducted at Group Health, Northwestern, and Marshfield (n = 161 individuals participating in 20 focus groups across all three studies) also expressed support for broad data sharing, particularly where such sharing might promote more efficient use of study resources (Lemke et al. 2010; Trinidad et al. 2010; McCarty et al. 2011b). Furthermore, most respondents [69.5% (n = 2800/4040)] to a Vanderbilt survey of its faculty and staff (who overwhelmingly use Vanderbilt University Medical Center and so constitute at least some potential participants in the BioVU repository, though their views on data sharing may well be different from those of other potential participants) reported that sending data to a national database would make no difference to their willingness to participate, and a proportion

Table 3. Methods used and topics addressed with participants and community advisory boards

Institution	GHC	Marshfield	Mayo	Northwestern	Vanderbilt
Consultation method(s)	Focus groups (n = 79), telephone survey (n = 365), consensus panel (n = 13)	Focus groups (n = 33), newsletter (n ~ 12,500), Community Advisory Board (CAB, n = 20)	Interviews (n = 50), deliberative engagement (n = 21), CAB (n = 20)	Focus groups (n = 49), web survey of IRB members (n = 208), CAB (n = 25)	Web survey (n = 4037), CAB (n = 10)
Topics addressed					
NIH GWAS data-sharing policy/deposition of data in national database dbGaP	X	X	X	X	X
Benefits of sharing de-identified genotypic and phenotypic data	X	X	X	X	
Informed consent requirements	X	X	X	X	
Participant re-consent for dbGaP deposition	X				
With whom (investigators, organizations) data might be shared	X	X	X	X	
Sharing with for-profit organizations	X	X	X	X	
Security to prevent unauthorized access and protect confidentiality	X	X	X	X	X
Mechanisms to ensure third-party compliance with local standards, especially with regard to secondary use and/or privacy			X	X	
Withdrawing from research once data are shared	X		X		

[18.5% (n = 745/4040)] said it would make them *more* likely to participate (KB Brothers, DR Morrison, and EW Clayton, in prep.). Finally, at Marshfield Clinic, after dbGaP-related data sharing was outlined in a newsletter mailed to all 20,000 biobank participants, only one request for withdrawal was received.

Where stakeholders' views on dbGaP-related data sharing were explored in greater detail, some participants expressed concern about the possibility of the federal government's storing and controlling access to their genetic and health-related data. However, it was not clear to what degree, if any, such concerns influenced potential participants' willingness to participate in research. In a four-day deliberative community engagement exercise held with 21 Olmsted County, Minnesota, residents to guide planning of the Mayo Clinic Biobank, participants endorsed data sharing as long as the appropriate safeguards were put in place and approved by the institution participants had entrusted with their data and samples (in this case, the Mayo Clinic). Focus group discussions at Group Health, Northwestern, and Marshfield also suggested that where concerns about federal control were expressed, they were often balanced by trust in the home institution and confidence in the ability of local investigators to ensure that any data shared with federal entities would be suitably protected (Trinidad et al. 2011). Participants at Northwestern strongly recommended increased efforts to educate the public about the uses of data and general outcomes as a means to increase public trust (Lemke et al. 2010). At Group Health, where the IRB determined that re-consent for dbGaP data sharing was required, 152 (11%) of 1340 cognitively intact study participants declined to give permission for such sharing, and 90% of 365 who had re-consented said that it was very or somewhat important that they had been asked for their permission (Ludman et al. 2010).

Another major theme identified by several of the study sites was stakeholders' discomfort with the possibility that their data might be shared with commercial entities. At Marshfield, the Community Advisory Board expressed the desire that, where possible, Marshfield would work collaboratively with industry to advance the community's interests. As with federal data sharing, it was unclear whether or not participants who expressed concerns about commercial data sharing would withdraw from the research if they were informed that it is allowable following deposition into dbGaP unless explicitly precluded in the informed consent. In the Group Health survey of participants who had re-consented for dbGaP submission, slightly more than 40% (n = 151/365) expressed concerns about the for-profit use of their shared data (Ludman et al. 2010). Yet, as mentioned above, nearly 90% of these participants re-consented to broad data sharing, suggesting that, although these concerns exist, at least in some cases they are outweighed by participants' desire to contribute to research.

Discussion

Participant perspectives on data sharing deserve careful consideration by policymakers, funders, and researchers. Although the eMERGE community engagement activities described above were generally conducted long after participants were enrolled into the different studies, in many cases the sort of limitations desired by participants was incorporated into the original informed-consent document. For example, some eMERGE institutions made assurances in the original consent that data will not be used by for-profit companies. Per the NIH GWAS data-sharing policy, institutional certification for submission to dbGaP must ensure that data-use limitations reflect what was communicated to the research participants in the informed-consent process about how their data may be used and who will have access to it (National Institutes of Health 2007). Investigators who submit data access requests for study data sets through dbGaP sign a Data Use Certification and

agree to abide by these data-use limitations (https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?view_pdf&stacc=phs000007.v2.p1). However, for many NIH grant-making programs with finite funds available, breadth of data sharing permitted is a factor considered in funding decisions (http://grants.nih.gov/grants/gwas/GWAS_faw.htm). Studies at sites that are not able to share data broadly might be considered a lower priority for funding, a result which some might object to as inappropriate policy. This raises important ethical challenges and questions about how much control individual sites should be able to retain over their shared data sets, given their responsibilities as the sole steward of data and biological materials supplied by the original participants.

The policy also anticipates that, in some cases, circumstances beyond the control of the investigators may preclude submission of GWAS data to dbGaP. For example, certain groups (e.g., tribal groups or other potentially identifiable population-based cohorts) may wish to maintain control over the data and prohibit submission to dbGaP altogether. In these rare cases, the investigator is expected to provide in the grant application an explanation for why submission to dbGaP is not possible. The appropriate NIH Institute or Center will then consider this explanation when making the funding decision (National Institutes of Health 2007). Information about how many requests for such limitations to or exclusions from data sharing have been submitted to NIH, how many have been accommodated, and what influence, if any, this has had on funding decisions is not available but would be helpful to determine whether and to what extent this creates a problem for investigators.

It also remains unclear whether institutions submitting data to dbGaP can require review by a secondary user's (requestor's) IRB based on assurances made in the original consent or for certain types of data (e.g., those associated with a stigmatized disorder, such as substance abuse). Typically, requestors who are obtaining de-identified data from dbGaP are not considered to be conducting human subjects research (Office of Human Research Protections 2008) and so no IRB review is required for their research. However, because DNA is uniquely identifying (Lin et al. 2004; Homer et al. 2008), in order to protect the privacy of certain groups and to respect the wishes of participants, some investigators have requested an additional layer of ethical review prior to the downstream use of certain data sets. There are a few examples of this restriction being implemented in dbGaP (see e.g., Framingham SNP Health Association Resource, http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000007.v13.p5 and CIDR: Collaborative Study on the Genetics of Alcoholism Case Control Study, http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000125.v1.p1), but decisions about when such requests will be accommodated are made by each Institute or Center (IC) and may depend on the uniqueness of the data resource and whether alternative solutions are possible.

Consistent policies within and between institutions are needed so that investigators, IRBs, and other institutional officials know what level of protection they can promise to participants, and participants can make decisions based on accurate and truthful information. Clearer guidance is also needed from NIH as a whole and its individual Institutes and Centers so that investigators and institutions can know what to expect when submitting grant applications and requests to limit data access. Creating data-use limitations that are responsive to participants' concerns shows respect for those who contribute to research and may increase trust, resulting in greater research participation. However, investigators may believe that they cannot place such limits on data sharing if it precludes them from funding opportunities. There are many benefits to broad data sharing that the NIH GWAS data-sharing policy promotes. Our preliminary research with eMERGE participants suggests that, although most are willing to share their data, there is a strong desire for some data-use limitations. Requiring all participants to agree to unrestricted data release may result in certain groups choosing not to participate in research, which could create subject bias and influence the ability of investigators to identify disease variants relevant to the population at large (Kohane and Altman 2005; McGuire and Gibbs 2006).

Conclusion

Data sharing creates novel challenges for researchers and institutions, both ethically and administratively. Solutions can be developed ad hoc, but this approach runs the risk of data sharing being rushed and implemented without full consideration and may lead to poor decisions and outcomes. The eMERGE Consortium experience illuminates some of the challenges associated with implementing the NIH GWAS data-sharing policy. Many of these challenges can be addressed by clearer guidance from NIH to ensure that the ethical safeguards built into the policy are upheld without significant burden to investigators or institutions. However, in some cases, the NIH may need to change its policies and/or practices.

The eMERGE Consortium experience also illustrates the need for institutions and researchers to adjust to a new status quo in research where data-sharing requirements are the norm. Processes to facilitate data sharing among institutions and IRBs should be developed and ethical challenges should be addressed preemptively. Researchers and institutions must also be ready to share data consistent with the NIH GWAS data-sharing policy. A process should be in place to ensure compliance with the policy, and relevant expertise should be available to assist researchers and review protocols ahead of time. To the extent possible, harmonization across sites, especially for large consortia, should be a priority. Finally, in order to address community concerns, researchers should integrate community engagement and ethics evaluation into their study design and budget. Ultimately, institutions and researchers must meet their stewardship obligations by developing policies and practices that nurture trustworthiness in relationships with participants and communities.

Acknowledgments

The eMERGE Network was initiated and funded by NHGRI, in conjunction with additional funding from the National Institute of General Medical Sciences (NIGMS) through the following grants: U01-HG-004610 (Group Health Cooperative); U01-HG-004608 (Marshfield Clinic); U01-HG-04599 (Mayo Clinic); U01-HG-004609 (Northwestern University); and U01-HG-04603 (Vanderbilt University, also serving as the Administrative Coordinating Center). Additional funding provided by the Clinic Center for Translational Science Activities (UL 1 RR024150) from the National Center for Research Resources (B.K.); NHGRI R01HG004333 (A.L.M.); University of North Carolina Center for Genomics and Society NHGRI P50HG004488 (L.D.); the Institute of Translational Health Sciences UL1RR025014 (NCRR) and the Center for Genomics and Health Care Equality P50HG003374 (NHGRI) (S.M.F.); the Wayne and Gladys Valley Foundation 03-071, the Ellison Medical Foundation AG-IA-0046-04, the RWJ Foundation 64362, and a generous grant from KP Community

Benefit (C.P.S.). The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of the National Human Genome Research Institute or the National Institutes of Health. We thank Veida Elliott and Kyle Brothers (Vanderbilt University), Jill Oliver (Baylor College of Medicine), and Joel Wu (Mayo Clinic) for research assistance and editorial support.

References

Benitez K, Malin B. 2010. Evaluating re-identification risks with respect to the HIPAA privacy rule. *J Am Med Inform Assoc* **17:** 169–177.

Clayton EW, Smith M, Fullerton SM, Burke W, McCarty CA, Koenig BA, McGuire AL, Beskow LM, Dressler L, Lemke AA, et al. 2010. Confronting real time ethical, legal, and social issues in the Electronic Medical Records and Genomics (eMERGE) Consortium. *Genet Med* **12:** 616–620.

Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, Pearson JV, Stephan DA, Nelson SF, Craig DW. 2008. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet* **4:** e1000167. doi: 10.1371/journal.pgen.1000167.

Kohane IS, Altman RB. 2005. Health-information altruists: A potentially critical resource. *N Engl J Med* **353:** 2074–2077.

Kukull WA, Higdon R, Bowen JD, McCormick WC, Teri L, Schellenberg GD, van Belle G, Jolley L, Larson EB. 2002. Dementia and Alzheimer disease incidence: A prospective cohort study. *Arch Neurol* **59:** 1737–1746.

Kullo IJ, Fan J, Pathak J, Savova GK, Ali Z, Chute CG. 2010. Leveraging informatics for genetic studies: Use of the electronic medical record to enable a genome-wide association study of peripheral arterial disease. *J Am Med Inform Assoc* **17:** 568–574.

Lemke AA, Wolf WA, Hebert-Beirne J, Smith ME. 2010. Public and biobank participant attitudes toward genetic research participation and data sharing. *Public Health Genomics* **13:** 368–377.

Lin Z, Owen AB, Altman RB. 2004. Genomic research and human subject privacy. *Science* **205:** 183.

Ludman EJ, Fullerton SM, Spangler L, Trinidad SB, Fujii MM, Jarvik GP, Larson EB, Burke W. 2010. Glad you asked: Participants' opinions of re-consent for dbGaP data submission. *J Empir Res Hum Res Ethics* **5:** 9–16.

Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L, Kiang A, Paschall J, Phan L, et al. 2007. The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet* **39:** 1181–1186.

Malin B, Karp D, Scheuermann RH. 2010. Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research. *J Investig Med* **58:** 11–18.

Malin B, Benitez K, Masys D. 2011. Never too old for anonymity: A statistical standard for demographic data sharing via the HIPAA privacy rule. *J Am Med Inform Assoc* **18:** 3–10.

McCarty CA, Wilke RA, Giampietro PF, Wesbrook SD, Caldwell MD. 2005. Marshfield Clinic Personalized Medicine Research Project (PMRP): Design, methods, and recruitment for a large population-based biobank. *Per Med* **2:** 49–79.

McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, Li R, Masys DR, Ritchie MD, Roden DM, et al. 2011a. The eMERGE Network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med Genomics* **4:**13. doi: 10.1186/1755-8794-4-13.

McCarty CA, Garber A, Reeser JC, Fost NC. 2011b. Study newsletters, community and advisory board input, and focus groups discussions provide ongoing consultation for a large biobank. *Am J Med Genet A* (in press).

McGuire AL, Gibbs RA. 2006. No longer de-identified. *Science* **312:** 370–371.

Menikoff J. 2010. The paradoxical problem with multiple-IRB review. *N Engl J Med* **363:** 1591–1593.

National Institutes of Health. 2007. Implementation guidance and instructions for applicants: Policy for sharing of data obtained in NIH-supported or conducted genome-wide association studies (GWAS). http://grants.nih.gov/grants/guide/notice-files/NOT-OD-08-013.html

Office of Human Research Protections. 2008. *Guidance on research involving coded private information or biological specimens*. U.S. Department of Health and Human Services (HHS), Washington, DC. http://www.hhs.gov/ohrp/policy/cdebiol.html

Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, Balser JR, Masys DR. 2008. Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin Pharmacol Ther* **84:** 362–369.

Trinidad SB, Fullerton SM, Bares JM, Jarvik GP, Larson EB, Burke W. 2010. Genomic research and wide data sharing: Views of prospective participants. *Genet Med* **12:** 486–495.

Trinidad SB, Fullerton SM, Ludman EJ, Jarvik GP, Larson EB, Burke W. 2011. Research practice and participant preferences: The growing gulf. *Science* **331:** 287–288.

U.S. Dept. of Health and Human Services. 2009. Protection of human subjects. In *Code of federal regulations, Title 45, Part 46*. U.S. Government Printing Office, Washington, DC.

Wolf WA, Doyle MJ, Aufox SA, Frezzo TF, Smith ME, Kibbe KE, Chisholm RL. 2003. DNA banking study in an ethnically diverse urban university hospital. *Am J Hum Genet* **73:** 423.

Received January 5, 2011; accepted in revised form April 28, 2011.