Prepared Statement of Bradley Malin, Ph.D.
Associate Professor of Biomedical Informatics & Computer Science, Vanderbilt University
National Committee on Vital & Health Statistics; Subcommittee on Privacy, Confidentiality, & Security
Hearing on Next Steps for Community Data Use: Beyond Data Use Agreements
Tuesday, April 17, 2012

Good morning and first, I wish to thank the committee for the opportunity to present testimony this morning. My name is Bradley Malin, and I work as an Associate Professor of Biomedical Informatics in the School of Medicine and an Associate Research Professor of Computer Science in the School of Engineering at Vanderbilt University in Nashville, Tennessee. I have experience in developing, applying, and evaluating policies and technologies for the governance of electronic medical record systems and biobanks. I was asked to provide several remarks on behalf of the Coordinating Center for the Electronic Medical Records and Genomics (eMERGE) Network[1] [MCC+11], of which I am a member, regarding the topics of governance models, data use agreements (DUAs), and alternatives to DUAs as they pertain to local communities and biomedical research. Before commencing, it will benefit the Subcommittee to learn that eMERGE is sponsored by the National Human Genome Research Institute of the National Institutes of Health. The network functions as a national consortium, established to develop, disseminate, and apply approaches to research that combine DNA-based biorepositories with data derived from electronic medical record (EMR) systems for large-scale, high-throughput, genetic research. Initially formed in 2007, the network currently consists of seven member sites distributed across the United States: 1) Essentia Institute of Rural Health (in partnership with Marshfield Clinic), 2) Geisinger Health System, 3) Group Health Cooperative of Puget Sound (in partnership with the University of Washington), 4) Mayo Clinic, 5) Mt. Sinai Hospital (in partnership with Columbia University), 6) Northwestern University, and 7) Vanderbilt University.

When discussing governance issues in this context, it is helpful to consider the lifecycle of data management. The following characterization is an oversimplification of the lifecycle, but it is useful to envision three interconnected steps: 1) initial data collection, 2) data utilization, and 3) data dissemination.

In the first step of the process, data is solicited from the community. Though eMERGE is a consortium, it was recognized from the outset that each site is situated in a different locale, with disparate policies

[1] Further information available online at http://www.gwas.net

and procedures, as well as populations from which data is collected. Each member site of eMERGE is thus provided with the opportunity to consult with its constituent populations to establish basic principles for data collection and research oversight. At the same time, eMERGE sites share best practices, through frequent phone calls and scholarly publications, to inform governance strategies within and beyond the consortium, such as through public workshops, when possible. [CSF+10] Despite the loose relations among eMERGE sites at this point in the lifecycle, a driving principle common across the sites is that the community should be involved in the planning and continued oversight of the biomedical research program. Appendix A (attached) provides a summary of the variety of activities that eMERGE sites have dedicated to community involvement. Let me provide several examples to illustrate the various manners in this involvement has been realized in eMERGE:

Community Engagement Models: All eMERGE sites have utilized community engagement models in one way or another. [MBD+11] Table 1 (derived from [MBD+11]) summarizes various methods used and topics addressed with participants and community advisory boards by the first round of funded eMERGE sites.

For example, when the Mayo Clinic began its joint EMR-biobanking efforts, it adopted a deliberative community engagement model, based on the principles of deliberative democracy[2]. In doing so, Mayo engaged community members in open dialogue over four days of activities. Deliberants were initially provided with background materials on various issues around biobanking, biomedical research, and efforts locally at Mayo. [Mayo12] Additionally, they were afforded the opportunity to interact with a variety of domain-specific experts, such as scientists involved in genetics research and patient privacy advocates. Then, with the aid of facilitators, the deliberants spent the remaining days debating the issues and formulating and refining recommendations. The recommendations were stratified across four main areas: 1) interaction with donors, 2) community involvement, 3) options for participation, and 4) sample sharing and accountability. Further details can be found in [HHK08].

To highlight another example, Northwestern has used focus groups and surveys to gauge the community's views on issues such as biobanking research, consent, and data sharing [LWB+10]. Additionally, they have used follow-up surveys to gauge the extent to which research subjects understood the informed consent measures post-participation [OCH+09].

Community Advisory Boards: Community advisory bodies are in place at all eMERGE sites, such as the Community Advisory Group (CAG) at Marshfield [MWG+05] and the Community Advisory Committee (CAC) at Northwestern [LWB+10].

[2] Deliberative democracy is a form of representative democracy which involves groups of citizens who discuss and decide policy issues; an approach focused on enhancing the nature and form of political participation. Further details can be found in [BR97].

To provide details on a specific advisory body, I will highlight community-related activities at Vanderbilt, whose research program was predicated on the establishment of an opt-out, de-identified environment. [RPB+08, PCB+10] The Vanderbilt Institutional Review Board (IRB) agreed that the plan did not meet the criteria of human subjects research, but given the anticipated scale of the project and its potential impact on the local community, it advised additional safeguards. These included ongoing institutional and IRB oversight; evaluation by the medical center's ethics committee; and establishment of Ethics, Scientific, and Community Advisory Boards (EAB, SAB, and CAB). Regarding the latter recommendation, the CAB was established to ensure community involvement and input into the design and function of the repository's operations, with the goal of evaluating and ultimately supporting acceptance among broader medical and lay audiences. Initially twelve members, the CAB represents a cross-section of the local community, based on employment, parenting activities, church groups, civic groups, educational activities, or extracurricular activities. A familiarity with science and genetics was not expected. The CAB is ongoing and meets several times per year to evaluate the conduct of the repository's operations in the context of established security and privacy measures, voice issues raised in the community relating to the use of genetic information for research, and identify practical measures toward resolving ethical or social dilemmas.

Dissemination of Research & Findings to the Community: The previous two examples demonstrate how community stakeholders may be engaged, educated, and involved to assist in the oversight of biomedical research. However, eMERGE sites have further shown it is also important to keep patients aware of how their participation facilitates biomedical research. The Marshfield Clinic, for instance, uses quarterly newsletters to inform the public about specific research projects and how data is being shared with a wider research community. Research conducted by eMERGE investigators have shown that such activities have various positive benefits, such as affirming the value of research participation, informing participants about research conducted based on broad consent, educating patients and the public, and building trust in the research enterprise. [BBF+12]

After data has been collected from a community of patients, eMERGE sites then turn to the second step in the lifecycle: utilization. At this point, when the data is studied locally (i.e., at the eMERGE site which it was collected), investigators access the data through traditional mechanisms, such as IRB-approved research protocols. And, certain sites may have additional requirements beyond the IRB approval and oversight process. For instance, Vanderbilt provides researchers with access to de-identified records, but it is recognized there is residual risk of identification [MLB+11] when working with such records. As such, local investigators are required to enter into a data use agreement with Vanderbilt in which they agree not to try to identify previously de-identified EMR data and DNA-based biospecimens or data. Additionally, each IRB-approved study is assigned its own research number, such that investigators must register each phenotype under investigation with biorepository managers – even if the same cohort is being studied. Though this process does not prevent investigators from using data or biospecimens, it is designed to remind investigators about their responsibilities to

perform research in a manner that adheres to the expectations of the repository and hold them accountable to their actions.

Upon completion of research, the findings may be returned to the community through the dissemination of aggregated findings as mentioned earlier. When specific research findings are ready to be returned to specific patients, the manner by which this is accomplished is defined by each site's policies and procedures. I will note that the return of research results is a complex problem, composed of scientific as well as legal and regulatory issues. NHGRI has formed a separate expert consortium, and the moral aspect of whether results should be returned is outside the scope of my testimony today. However, I refer the Subcommittee to [FWB+12] for details on how eMERGE has deliberated and addressed this issue.

However, one challenge associated with the return of research results to a community or specific individual I wish to comment on is the extent to which a finding can be scientifically validated. In this regard, eMERGE leverages the networked aspect of its mandate to facilitate multi-site experiments and quality control (e.g., [CBC+11, DCR+11, KHR+12, PRB+12]). For instance, one of the charges of eMERGE is to develop algorithms to detect patients with a specific phenotype of interest from existing EMR data. In doing so, one site can propose a method to detect cases and controls for a phenotype. The fidelity of the approach is then evaluated at each site. This is critical to assess how local practices of clinical care documentation influence the performance of such specifications, which enables informaticians to develop techniques that are neither overly-specific nor overly-general. Moreover, eMERGE sites may perform joint studies on the same phenotype, whereby each site provides a portion of the study cohort. In such cases, it is critical to ensure cases (and controls) at one site are comparable to those at another site.

After data has been utilized locally, we move into the third step of the lifecycle. At this point, the data may be shared beyond the local institution that collected it. Here, I wish to highlight how data is shared to other eMERGE sites for additional research purposes and then how data is shared more broadly. To facilitate the process between eMERGE sites, the consortium employs the assistance of a Coordination Center (CC). In addition to administrative support, the CC assists the sites in standardizing, harmonizing, and performing quality control on the clinical and genomic data collected by the sites for their studies. Additionally, the CC has assisted in the establishment of an overarching DUA, which is designed to facilitate data sharing among the sites.[3] The DUA defines the principles behind guiding data sharing in the consortium, the responsibilities of the parties involved in the sharing and receiving of the data, a statement of confidentiality (to ensure that data is not shared beyond eMERGE members and affiliates), and limitations of data use. The principles of data sharing in this context are oriented to ensure that all sharing adheres to

[3] Please see the appendix in [NRC11]

1) the terms of the consent agreed to by participants at the local sites,
2) applicable laws and regulations, and
3) the principle that individual sites have final authority regarding whether their site's data will be used or shared on a per-project basis.

Once the DUA is established, eMERGE sites may share data with one another directly or it may be disseminated via the CC.

The DUA as initially drawn up required that any two eMERGE sites execute the contract jointly in order to share. In time, it was recognized this limited the speed with which eMERGE could incorporate new sites into the consortium. As such, eMERGE recently migrated to a more general DUA that enables greater flexibility regarding who can join (or leave) the consortium.

In addition to harmonizing data and facilitating its exchange, the CC also serves as a clearinghouse for all new research studies that members of the consortium may undertake. Briefly, when an investigator (or site) is interested in conducting research on a new phenotype, it drafts and posts a concept sheet to the CC. This sheet documents the topic to be studied, the investigators proposing the study, and the dataset(s) to be studied. The CC then notifies all of the sites about the proposed study, and each site can then propose to be a part of the study as they choose. And, as noted in principle #3 of the data use agreement, each site may decline to include their data in the proposed study. The goal of the concept sheet and CC facilitation is to ensure transparency in the behavior, as well as equity in the intellectual capital, of the eMERGE members.

Finally, since eMERGE is sponsored by the National Institutes of Health, it is subject to its data sharing policies. Thus, all genomic and clinical data generated by, or studied to substantiate findings, are shared beyond the eMERGE sites to support validation of findings and enable novel biomedical investigations. Currently, this data, along with appropriate documentation on patient consent, is sent in a de-identified form to the database of Genotypes and Phenotypes (dbGaP) at the National Center for Biotechnology Information (NCBI). General information about what type of information is available per dataset is made available to the public (e.g., number of patients in the study, type of data in the study, number of patients who provided consent per type of research, etc.). When a non-eMERGE investigator wishes to access and use the data, they must receive local IRB approval and submit a request to a data access committee (DAC) at the NIH, whose role it is to ensure that the proposed research study has sufficient merit and that appropriate protections are in place to warrant access to patient-level, but still de-identified records.

On behalf of the eMERGE Coordinating Center member sites, I thank the Subcommittee for its attention to an important policy issue. I would also like to thank Teri Manolio and Rongling Li of the NIH; Melissa Basford, Ellen Wright Clayton, Jonathan Haines, and Dan Roden of Vanderbilt, Rex Chisholm and Maureen Smith of Northwestern for their assistance in assembling this statement. Please

feel free to contact us at any time for further clarification of the issues we have raised. I would be pleased to try to answer any questions that you might have.

Cited References

[BBF+] Beskow LM, Burke W, Fullerton SM, Sharp RR. Offering aggregate results to participants in genomic research: opportunities and challenges. *Genetics in Medicine*. 212; 14(4): 490-496.

[BR97] Bohman J, Rehg W, eds. Deliberative democracy: essays on reason and politics. MIT Press. 1997.

[CBC+11] Conway M, Berg R, Carrell D, et al. Analyzing the heterogeneity and complexity of electronic health record oriented phenotyping algorithms. *Proceedings of the American Medical Informatics Association Annual Symposium*. 2011: 274-283.

[CSF+10] Clayton EW, Smith M, Fullerton SM, et al. Confronting real time ethical, legal, and social issues in the eMERGE (Electronic Medical Records and Genomics) consortium. *Genetics in Medicine*. 2010; 12(10): 616-620.

[DCR+11] Denny JC, Crawford DC, Ritchie MD, et al. Variants near FOXE1 are associated with hypothyroidism and other thyroid conditions: using electronic medical records for genome and phenome-wide studies. *American Journal of Human Genetics*. 2011; 89(4): 539-42.

[FWB+12] Fullerton SM, Wolf WA, Brothers KB, et al. Return of individual research results from genome-wide association studies: experience of the Electronic Medical Record and Genomics (eMERGE) Network. *Genetics in Medicine*. 2012; 14(4): 424-431.

[HHK08] Hicks A, Hellyer JH, Koenig B. Involving the public in planning for the genomics revolution: an experiment in deliberative democracy. *University of Minnesota Examiner*. 2008; 11(2): 1-3.

[KHR+12] Kho AN, Hayes MG, Rasmussen-Torvik L, et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. *Journal of the American Medical Informatics Association*. 2012; 19(2): 212-218.

[LWB+] Lemke AA, Wolf WA, Herbert-Beirne J, Smith ME. Public and biobank participant attitudes toward genetic research participation and data sharing. *Public Health Genomics*. 2010; 13: 368-377.

[Mayo12] Mayo Clinic. DNA biobanking in Olmstead County: A deliberative community engagement. Booklet. Available online at: http://biobank.mayo.edu/upload/booklet.pdf. Last accessed: April 9, 2012.

[MBD+11] McGuire AL, Basford M, Dressler LG, et al. Ethical and practical challenges of sharing data from genome-wide association studies: the eMERGE consortium experience. *Genome Research*. 2011; 21(7): 1001-1007.

[MCC+11] McCarty CA, Chisholm RL, Chute CG, et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Medical Genomics*. 2011; 4: 13.

[MLB+11] Malin B, Loukides G, Benitez K, Clayton EW. Identifiability in biobanks: models, measures, and mitigation strategies. *Human Genetics*. 2011; 130(3): 383-392.

[MWG+05] McCarty CA, Wilke RA, Giampietro PF, et al. Marshfield Clinic personalized medicine research project (PRMP): design, methods, and recruitment for a large population-based biobank. *Personalized Medicine*. 2005; 2(1): 49-79.

[NRC11] National Research Council. Towards personalized medicine: building a knowledge network for biomedical research and a new taxonomy of disease. *National Academies Press*. 2011.

[OCH+09] Ormond KE, Cirino AL, Helenowski IB, et al. Assessing the understanding of biobank participants. *American Journal of Medical Genetics A*. 2009; 149A(2): 188-198.

[PCB+10] Pulley JM, Clayton EW, Bernard GR, Roden DB, Masys DR. Principles of human subjects protections applied in an opt-out, de-identified biobank. *Clinical and Translational Science*. 2010; 3(1): 42-48.

[PRB+12] Peissig P, Rasmussen LV, Berg RL, et al. Importance of multi-modal approaches to effectively identify cataract cases from electronic health records. *Journal of the American Medical Informatics Association*. 2012; 19(2): 225-234.

[RPB+08] Roden DM, Pulley JM, Basford MA, et al. Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clinical Pharmacology and Therapeutics*. 2008; 84(3): 362-369.

Other Selected References from eMERGE That May Interest the Committee

Zuvich RL, Armstrong LL, Bielinski SJ, et al. Pitfalls of merging GWAS data: lessons learned in the eMERGE network and quality control procedures to maintain high data quality. *Genetic Epidemiology*. 2011; 35(8): 887-898.

Pathak J, Pan H, Wang J, et al. Evaluating phenotypic data elements for genetics and epidemiological research: experiences from the eMERGE and PhenX network projects. *Proceedings of the AMIA Summits on Translational Science*. 2011: 41-45.

Turner S, Armstrong LL, Bradford Y, et al. Quality control procedures for genome-wide association studies. *Current Protocols in Human Genetics*. 2011; Chapter 1: Unit1.19.

Table 1. Characteristics of eMERGE biobank populations and phenotypes for GWAS (genome-wide association studies). Reproduced from [MBD+11]

Institution	**Biobank Population**	**Biobank Size & Demographics**	**Ongoing Participant Interactions**	**Primary GWAS Phenotypes**
Group Health Cooperative (Seattle, WA)	Disease specific: Adult Changes in Thought (ACT) Study cohort; 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 (Marshfield, WI)	Broad population: Personalized Medicine Research Project; 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)
Mayo Clinic (Rochester, MN)	Disease specific: Cases identified from noninvasive vascular lab database; controls identified from the Cardiovascular Health Clinic	1641 cases and 1604 controls Age: mean 66 +/_ 11 yr, cases; 61 +/_ 8 yr, controls 96% European ancestry	No	Peripheral Arterial Disease (n = 3335)
Northwestern University (Chicago, IL)	Broad population: NUgene Project; 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 (Nashville, TN)	Broad population: BioVU; use of discarded blood/non-human subjects linked to EMRs	>100,000 samples All ages 70% European ancestry 10% AA	N/A	QRS duration (n = 3192)