CONCEPT CLEARANCE FOR RFA

eMERGE (electronic Medical Records and Genomics) Phase II

NHGRI Advisory Council, May 2010

Purpose

NHGRI proposes two RFAs to speed incorporation of current genomic knowledge, combined with the electronic phenotyping and privacy protection methods generated in eMERGE phase I, into clinical research and practice in a Phase II eMERGE expansion. The goals of this initiative are to build upon the eMERGE experience in defining phenotypes from electronic medical records (EMRs), conducting genome-wide association (GWA) studies using those phenotypes, reducing risks to patient privacy from the sharing of EMR data, and developing consent and community consultation procedures for conducting such research. It will achieve these goals by: 1) expanding and validating the eMERGE "electronic phenotyping" library from 14 phenotypes to roughly 40; 2) expanding the number and diversity of participating eMERGE sites to include under-represented populations such as racial/ethnic minorities and children; and 3) incorporating GWA genotyping information into EMRs, where feasible, for improving genetic risk assessment, prevention, diagnosis, treatment, and/or accessibility of genomic medicine.

Background

eMERGE phase I is a 4-year program begun in late 2007 through RFA-HG07-005, "Genome-Wide Studies in Biorepositories with Electronic Medical Record Data." A "biorepository" in this program is a resource that collects and stores biospecimens from which DNA can be isolated, and which are linked to electronic health information. In eMERGE phase I, five biorepositories are conducting GWA studies of six site-specific phenotypes in roughly 18,000 subjects (Table), and a network-wide phenotype of resistant hypertension in an additional ~1,800 subjects. These 20,000 subjects will also be "electronically" phenotyped and analyzed for hypothyroidism. Six additional network-wide EMR phenotypes are being assessed in genotyped subjects through support made possible by the American Recovery and Reinvestment Act (ARRA).

eMERGE is unique in defining phenotypes from EMRs through structured data extraction, natural language processing and other informatics tools. Developing these definitions is an arduous process, requiring in-depth clinician review of records identified by preliminary EMR algorithms. Refinement of these algorithms to boost positive predictive value above 95% can require many iterations of code modification, record re-extraction, and clinician re-review. The feasibility of using EMR-derived phenotypes for GWA research within a single biorepository was recently demonstrated by the Vanderbilt eMERGE group (Ritchie M et al, *Am J Hum Genet* 2010) and has also been documented by the Marshfield and Mayo groups.

Each of the eMERGE sites is working to transport and validate its primary phenotyping algorithms in one or more additional eMERGE sites, and then to implement them within the diverse EMR systems at all five sites, thus increasing sample sizes at minimal extra cost. The external validity of these algorithms will also be assessed in settings outside eMERGE, such as the Veterans Administration and Kaiser Northern California. An EMR-based phenotyping library containing the "pseudocode" (syntactical description outlining the general steps of an algorithm rather than the actual computer code) for these algorithms is available through the eMERGE website (www.gwas.org). Using diagnostic codes extracted from EMRs, already-genotyped patients in eMERGE can be classified for several hundred additional preliminary phenotypes in

a phenome-wide scan, or "PheWAS" (Denny JC et al, *Bioinformatics* 2010). Phenotypes showing suggestive SNP-trait associations can then be refined through the eMERGE algorithm development process for more reliable GWA-based discovery studies.

While mining the vastness of EMR data greatly expands the potential for conducting GWA studies and related genomic research, it also expands the possible risks to patients' privacy as well as concerns about widespread data sharing. eMERGE is at the forefront of addressing the potential for re-identifiability in EMR data through bioinformatics research to determine the potential for linking large amounts of standardized clinical information, such as ICD-9 codes, back to patients' identifying information in their EMRs. eMERGE investigators have developed methods to extract potentially linkable clinical characteristics and modify them (by grouping or suppression) to minimize threats to the confidentiality of a patient's genomic information, while maximizing the EMR information preserved (Loukides G et al, *PNAS* 2010).

This and other concerns about genomic research linked to EMRs and widespread data sharing are being addressed by eMERGE's Consent and Community Consultation group, which explores consent and privacy concerns with participants and community advisory groups. Participant surveys have demonstrated: 1) generally favorable views of sharing; 2) support for sharing where it enhances the research value of the biorepository; and 3) some uneasiness over sharing with commercial entities. Participants have also expressed interest in receiving results of their genomic studies, even if findings are not actionable, but strong preferences and even expectations are expressed for receiving information that could affect clinical care.

Important limitations to the existing eMERGE network include the low proportions of minority participants and the lack of pediatric or other specialty settings. Returning results to patients or for use in clinical care is not addressed (or may even be proscribed) in some consent forms and is complicated by lack of CLIA certification for the eMERGE phase I genotyping process.

Opportunities presented by a continuation of eMERGE include: 1) expanding the phenotype library and ensuring its transferability; 2) increasing the diversity of participants; and 3) incorporating GWA results in these participants into their EMRs to begin to explore the value of this information in clinical decision-making. In addition, the advent of federally incentivized "meaningful use" of EMRs in clinical care through the HITECH Act will drive harmonization of EMR data as they are generated and recorded, rather than after the fact. eMERGE is thus well-positioned to promote standardization of EMR data specifically for genomic research.

Productivity to date has been high, with 19 papers accepted or published,15 submitted, and 31 in development. Software tools for analyzing privacy risk (VDART) and harmonizing data dictionaries (eleMAP) are available though the eMERGE website, as is model consent language for genomic research in biobanks, which is also posted on the NHGRI consent website.

Research Scope and Objectives

In eMERGE Phase II, 5-8 biorepositories would be supported to expand and validate the eMERGE "electronic phenotyping" library, expand the number and diversity of participants and sites, and incorporate GWA genotyping information into EMRs, where feasible, for improving clinical care. Sharing of expertise and experience within and outside eMERGE will continue to be a key goal, with the intent of raising the standards for genomic research in biorepositories and its incorporation into medical care in general. A separate RFA will be issued to support a Coordinating Center; these functions are currently performed by one of the clinical sites but a more independent Coordinating Center appears scientifically and administratively preferable.

The Phase II eMERGE Steering Committee will define, validate, and disseminate EMR phenotypes; develop informatics tools for enhancing genomic research in biorepositories; and address consent and community concerns related to this research. A key component of eMERGE phase II will be examination of concerns regarding return of genotyping data to participants, their physicians, and their EMRs for use in clinical care. Such concerns might include constraints in existing consent documents, potential benefits and harms, lack of CLIA certification, and need to repeat some or all genotyping through CLIA-approved processes.

Within the first year of eMERGE phase II, an initial set of GWA-defined variants potentially useful in clinical practice for purposes such as assessment of very high (>99th percentile) genetic risk for complex disorders or selection or dosing of drugs, along with the levels of evidence supporting them, would be agreed upon by the Steering Committee. Informatic procedures for linking genotyping data on these variants to a participant's clinical EMR and appropriate decision-support tools (such as warning of an *SLCO1B1* variant when prescribing a statin, and recommending lower doses and more frequent monitoring), as well as policy decisions regarding need for CLIA-certification and potential re-consent, should also be developed within this period so that use of genotyping data in clinical care can be initiated by the beginning of year 2. Consent and certification constraints may limit the number of variants that can be returned to those that can be efficiently re-genotyped in accordance with accepted processes, either before or after a decision-support tool identifies their relevance to an individual patient. Limited funding for genotyping a subset of variants and/or participants will be made available. The process of defining clinically useful variants, the evidence supporting them, and the approvals necessary will continue in years 2-4. Initial experience from year 2 will be used to inform the approval process and minimize, where possible, the need for repeat genotyping.

Selection of eMERGE II sites will be based on: performance in phase I, as applicable; population diversity, particularly in regard to children and to minority populations with important health disparities; availability of high-quality GWA genotyping data in 3,000-4,000 patients from platforms testing at least 550,000 SNPs; consent for sharing individual-level data through dbGaP; completeness and usability of the available EMR data; and public health importance of the traits to be studied. Applicants to be "new" sites must demonstrate their ability to implement existing eMERGE phenotypes reliably; impute their existing GWA data as needed to combine in analyses with existing eMERGE data; conduct appropriate consultations with participants, their communities, and other relevant stakeholders on issues related to incorporating genomic data into individual participants' EMR; and develop new electronic phenotypes and disseminate them within and outside eMERGE. To be considered for inclusion in eMERGE phase II, existing eMERGE sites must demonstrate their ongoing productivity and contributions to eMERGE collaborative efforts. Other considerations will include the diversity of phenotypes and populations studied, availability of additional genotyped samples to be contributed to eMERGE, and completeness and flexibility of the site's EMRs for incorporating decision-making tools.

Mechanism of Support

This initiative would use the NIH U01 (Cooperative Agreement) award mechanism. Five to eight applicants would be selected for study sites and one for the Coordinating Center. Support for re-genotyping would be achieved through purchase agreements or a separate RFA.

Funds Available

NHGRI will commit roughly $30M over four years to support these cooperative agreements; roughly $5M of this will support repeat genotyping as needed for use in clinical care.

Table. Primary phenotypes, sample sizes, and EMR characteristics of eMERGE sites.

Institution	Primary Phenotype	Repository Size	GWA Study Size	EMR Description	Phenotyping Methods
Group Health, Seattle, WA	Alzheimer's Disease and Dementia	~4000; >96% EA	3,370; 97% EA	Vendor-based EMR since 2004; 20+ yrs pharmacy 15+ yrs ICD9	Structured data extraction, Mining free-text via regular expressions, Manual chart review
Marshfield Clinic, Marshfield, WI	Cataracts and HDL-Cholesterol	~20,000; 98% EA	3,968; 99% EA	Internally developed EMR since 1985; 75% ppts have 20+ yrs medical hx	Structured data extraction, NLP, Intelligent Character Recognition
Mayo Clinic, Rochester, MN	Peripheral Arterial Disease	3,500; >96% EA	3,412; 99% EA	Internally developed EMR since 1995; 40 yrs data extraction	Structured data extraction, NLP
Northwestern University, Chicago, IL	Type 2 Diabetes	9,200; 12% AA 8% Hispanic	3,564; 52% AA	Vendor based Inpt and Outpt EMR since 2000; 20+ yrs ICD9	Structured data extraction, text searches
Vanderbilt University, Nashville, TN	QRS Duration	100,000; 11% AA	3,061; 16% AA	Internally developed EMR since 2000; 35+ yrs medical hx	Structured data extraction, NLP

ICD9 = Ninth International Classification of Diseases
NLP = Natural Language Processing.
Structured data extraction = retrieving data that have been stored in a predefined format