Rapporteurs: Sheethal Jose
and Melpomeni Kasapi,
Program Analysts, NHGRI

eMERGE & Beyond: The Future of Electronic Medical Records (EMR) and Genomics
October 30, 2017 – Rockville, MD

Welcome, Introduction and Opening Remarks (Eric Green and Rongling Li)

Co-chairs Dan Masys and Sharon Plon summarized the workshop objectives: to review the current work and accomplishments of the Electronic Medical Records and Genomics (eMERGE) Network and discuss future directions of genomic research and clinical care using the electronic medical record (EMR).

Eric Green thanked the participants for providing substantial input in helping the National Human Genome Research Institute (NHGRI) influence the genomic medicine landscape. NHGRI's strategic vision is based on a plan published in 2011, and as the Institute is thinking about its next stage, workshops such as this one aid in enriching and directing NHGRI's future efforts. Eric urged everyone to think forward and leverage the synergies and collaborations to best situate the next phase of eMERGE. Rongling Li highlighted NHGRI's appreciation to the workshop planning committee, external scientific experts, and the eMERGE investigators for their help with the workshop preparation.

NHGRI Genomic Medicine Portfolio (Teri Manolio)

NHGRI's Genomic Medicine portfolio includes the Undiagnosed Diseases Network (UDN), the Newborn Sequencing in Genomic Medicine and Public Health (NSIGHT) program, the Clinical Sequencing Evidence-generating Research (CSER) consortium, eMERGE, the Implementing Genomics in Practice (IGNITE) program, the Clinical Genome (ClinGen) Resource, and investigator-initiated projects in clinical sequencing research, HIV/AIDS drug response and co-morbidities, and serious adverse drug reactions. NHGRI's Genomic Medicine programs span the spectrum of genomic medicine implementation, ranging from those with an individual patient focus testing multiple models of clinician-patient interactions, to those like eMERGE that involve working system-wide to generate evidence of clinical utility of genomic medicine implementation. Regarding related programs, CSER focuses on the clinical utility of genomic sequencing and barriers to integrating genomic data into clinical care, with emphasis on the clinical encounter, while like eMERGE including aspects of EMR integration, clinical impact of return of results (ROR) and data sharing concerns. IGNITE's next phase focuses on pragmatic clinical trials in diverse, non-expert clinical settings to assess the clinical utility of established genomic medicine interventions, while like eMERGE including aspects of EMR integration, cost-effectiveness, and patient/clinician education.

eMERGE Program Overview (Rex Chisholm)

In eMERGE Phase I (2007-2011), the Network demonstrated the utility of linking EMRs to biorepositories for genomics research, allowing eMERGE to identify novel associations through Genome Wide Association Studies (GWAS) as well as explore the ethical, legal, and social issues surrounding this kind of research. In eMERGE Phase II (2011-2015), two additional adult sites and three pediatric sites joined the Network, and the scope broadened from GWAS using electronic phenotypes (e-phenotypes) to piloting clinical implementation studies including methods development for integrating genomic information into the EMR and supporting clinical

decision-making. Pharmacogenomics (PGx) work was also initiated during this phase, along with an effort to enrich the EMR to improve and inform clinicians' decision-making on drug selection and dosing.

The current phase of eMERGE consists of 9 clinical sites, 2 sequencing centers and 1 coordinating center working synergistically to sequence and assess clinically relevant genes in about 25,000 individuals, assess the phenotypic implications of these variants, integrate genetic variants into EMRs for clinical care, and create resources for the scientific community. eMERGE has collected genotypic and phenotypic data from >110,000 participants so far and by the end of this phase the dataset is expected to reach 136,000 participants. The eMERGE Record Counter enables users to search demographic, eMERGE phenotypic, and billing code data (the International Statistical Classification of Diseases and Related Health Problems (ICD), and Current Procedural Terminology (CPT)) to obtain preliminary counts of affected individuals for potential studies. The Sequence and PHenotype INtegration Exchange (SPHINX) tool is a search catalog of 82 pharmacogenes by genes, drugs, variant identifiers (rsID or chromosomal position) and pathways. Users can view variants, pathways and drug interactions for each gene, and for each variant users can view SNP info, categories, and frequencies. The dataset encompasses European, African, and Asian ancestry allelic data for disease associations.

Present deliverables include the imputation and merging of GWAS data from all three eMERGE phases and analyses utilizing these data. The consortium has developed and is now performing sequencing of eMERGE participants on an eMERGEseq platform consisting of 109 genes and hundreds of single nucleotide variants (SNVs). In this platform, clinical reports are generated based on the eMERGE "Consensus Actionable List" which includes the genes recommended by the American College of Medical Genetics and Genomics (ACMG) in July 2013 for 56 genes, other genes requested by individual sites and agreed to by the network, and some genes for reporting as an incidental finding. To date, 14,077 samples have been sequenced and 3,716 reports issued. The Phenotype KnowledgeBase (PheKB) has created a collaborative environment to build and validate electronic algorithms. The tools and processes enable computational and algorithm development in collaborations around the world. So far, the computational algorithm library holds 37 publicly available phenotyping algorithms that have been validated within the consortium before publishing. The eMERGE phenotype-wide association study (PheWAS) has developed methods for large-scale genotype/phenotype analyses and has implemented them across the entire Network.

The PGx project that began in Phase II builds on genetic sequencing of 82 PGx genes in 9,010 participants and collects both utilization and outcomes data. Multi-sample calling was implemented on the original PGRNseq with all 9,010 binary sequence alignment files (BAMs) re-aligned to the same genome reference. In addition, eMERGE has created infrastructure and tools to enable genomic medicine implementation through improved knowledge representation and clinical decision support (CDS) in the EMR. Examples include the InfoButton, a decision support tool to provide context specific links within the EMR to relevant genomic medicine content, and the Clinical Decision Support KnowledgeBase (CDS_KB) to catalog and share CDS implementation artifacts and design considerations for genomic medicine programs from a broad community of institutions. CDS_KB, an open resource created from a collaboration between eMERGE and IGNITE, currently contains 60 artifacts submitted by a variety of healthcare organizations. Finally, network-wide analysis using DNAnexus, a secure, cloud-based storage and analytic service, enables multiple sites to create common tools that can be applied to different datasets. It also provides a novel way for conducting genomic analysis using cloud computing, since every site doesn't have to create its own dataset and pipeline, and

facilitates innovations through built-in applications. Sustainability and keeping these tools up to date are ongoing challenges.

Publication tracking in August 2017 has shown a total of 633 projects, both site and network-wide with topics in genomics and phenotyping being the most impactful. Future deliverables of the Network include the database of Genotypes and Phenotypes (dbGaP) submissions, return of clinical results and EMR integration at all sites, outcomes analysis for effect of return of results on patients and providers, and creation and deployment of 27 new e-phenotypes, as well as ongoing GWAS PheWAS analyses.

All of Us **Program Synergy with eMERGE (Stephanie Devaney)**

The *All of Us* (AOU) Research Program aims to deliver a national resource of deep clinical, environmental, lifestyle, and genetic data from 1 million participants who will be consented and engaged to provide data on an ongoing longitudinal basis. Additional considerations include reaching out to communities that have been underrepresented in biomedical research, as well as making research tools and resources easily accessible to a diversity of researchers from citizen scientists to premier university labs. There are several opportunities for convergence between the AOU program and the eMERGE Network as both programs are building large datasets that involve both EMR and genomic data. However, there are also differences between the two programs that allow for shared learning and leveraging of tools.

One area where eMERGE can learn from AOU is data access and having datasets in a shared integrated cloud platform, a direction in which eMERGE wants to move to improve efficiency. AOU is developing a tiered data access approval system and the access will be highly transparent and researcher-based, which allows researchers to launch as many studies as they want once they get access. There will be a public topical dataset that is fully open for anyone to access without a login, a registered dataset that requires a data use agreement and approval, and a controlled dataset that requires registered access and an institutional signing official.

An aspect that AOU can learn from eMERGE is the synthesis of disparate types of health data for research use, in which eMERGE has extensive experience and success. Although AOU has been learning a lot from eMERGE on how to effectively integrate health data with EMR, one step further is to build technology to harmonize data from disparate sources.

AOU's "Sync 4 Science" (S4S) protocol will allow participants to choose to share their EMR data from a S4S-enabled participant portal directly to the research program. This will enable participants joining the program as direct volunteers rather than from a specific healthcare provider to share their EMR data, and enable access to their EMR data from multiple providers.

AOU has substantial focus on participant engagement, retention, and return of results. AOU is working on several digital engagement tools to retain participants and keep them engaged. One approach to participant retention is to return information and specific results to the participants. AOU plans to return individual health information, comparative survey data, EMR data, claims data, research results, etc., but the biggest challenge will be the return of genetic results. eMERGE has done a lot of research on this and can share lessons learned with the AOU program. The CDS tools and patient educational resources that eMERGE has already developed will also benefit AOU.

Lastly, AOU plans to do e-phenotyping and the integration of genomic results into EMRs for clinical research and care. Given that the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) is the shared information model adopted by both AOU and eMERGE, AOU will be taking advantage of the well-validated and published OMOP CDM-based phenotyping algorithms that eMERGE has developed.

VA Million Veterans Program Phenomic Science (Michael Gaziano)

The Million Veteran Program (MVP) aims to enroll 1 million veterans from the Veterans Health Administration (VHA) into an observational mega-cohort. About 610,000 veterans have been recruited to date. Whole genome sequencing will be performed in 100,000 veterans. Metabolomic, proteomic, and microbiome pilot projects will also be included in the program.

The VHA is one of the largest integrated healthcare systems in the United States with around 24 million patients, greater than 7 billion lab results, and 3 billion clinical notes. Currently, most of the health data are inconsistently organized with modest amounts of structured data. The MVP has 3 main core teams: (1) Phenomics Core extracts the data and creates the library of all the phenotypes; (2) Data Analytics & Management Core does simple, structured data curation; and (3) Applied Bioinformatics in Clinical Research Core does the complex phenotyping and development of automated components.

Currently, the phenotyping process is manual with a 3-tier process of algorithm development and validation. However, it is moving towards semi-automated algorithm development which combines features of manual and automated phenotype development. Laboratory and medication adjudication processes have been used to validate results. They will be deploying natural language processing (NLP) in their Department of Energy computational workspace. MVP is exploring a cloud-based platform to share their data and to allow greater access to researchers. On November 17th, 2017, the Oak Ridge National Laboratory (ORNL) will make all the MVP data available on a cloud-based platform.

PANEL 1: Electronic Phenotyping for Genomic Research

eMERGE Presentation (George Hripcsak)

eMERGE has focused on phenotype sharing across the sites and reducing the time it takes to produce the phenotyping algorithms. The Network is adopting the Observational Health Data Sciences and Informatics (OHDSI)'s OMOP CDM that converts the current eMERGE site project data to the same schema and vocabulary to accelerate the phenotyping process.

eMERGE investigators identified challenges in developing and validating phenotypes. A significant challenge was using billing codes alone, an example of how research was impacted by the imperfect collection of health information. NLP was identified as an important component to develop effective phenotypes. It is also important to be aware of two different goals of phenotyping: (1) knowledge discovery through GWAS, which needs high positive predictive value and (2) knowledge deployment for decision support, which needs high sensitivity. Other lessons learned include the complexity of effective phenotype definitions, the use of tools to improve phenotyping, and the need for validation across multiple sites.

To do more sophisticated phenotyping in the future, it will be important to produce high-fidelity phenotyping algorithms. Currently, eMERGE is largely using binary phenotypes (disease present/absent), but will need to move towards more graded phenotypes. This includes factors

such as degree and severity of a condition and innovative methods to infer phenotypes. The high-fidelity phenotypes should also account for the biases in the healthcare process. Machine learning and other advanced computational tools can be used for this process. In eMERGE, the Harvard site has effectively and efficiently applied machine learning algorithms to a large population to accurately phenotype patients. The Cincinnati Children's Hospital Medical Center (CCHMC) and Geisinger sites have combined NLP and machine learning for automatic prediction of phenotypes. The Marshfield site has used these methods to reduce workload.

Reaction Presentation (Ken Kawamoto)

The main recommendation to the eMERGE phenotyping effort is to learn from and synergize with related efforts as e-phenotyping is a common problem encountered in the genomics area and beyond. It is crucial to create standards that EMR vendors will be likely to adopt. Ken noted that we should invite EMR vendors to these workshops. The major trend in the industry now is the use of the Health Level-7 (HL7) Clinical Quality Language (CQL). eMERGE should work with the EMR vendors to support the HL7 standards. There are two important ongoing initiatives to integrate data across platforms: HL7 Clinical Information Modeling Initiative (CIMI) and Healthcare Services Platform Consortium (HSPC). CIMI is based on detailed clinical models for true interoperability. The HSPC includes efforts to be interoperable by using Fast Healthcare Interoperability Resources (FHIR) interfaces. FHIR provides specifications for exchanging healthcare information electronically. HSPC plans to utilize FHIR interfaces to seamlessly integrate electronic data between EMR systems. This will allow for full functionality of the EMR, such as maintaining accurate health information and standardized administrative processing, to allow health providers to provide timely service to their patients. With regard to machine learning, Ken recommends eMERGE focus on basic approaches that are easiest to scale, such as rule-based processing of structured data. For low-resource settings, the Network should make judicious use of NLP. There could also be a focus on areas with gold standards, as establishing standards for phenotypes can be very costly. Other recommendations include leveraging NLP-facilitated phenotyping, studying alternate approaches to manual phenotype validation, and leveraging increasing EMR consolidation through the development of phenotype validation approaches that are optimized for the most frequently used EMR systems.

Panel Discussion and Summary (Josh Denny and Marylyn Ritchie)

In eMERGE, there has been a progression from defining research cases and developing PheKB published algorithms to use of e-phenotypes in clinical care, both for finding important clinical populations within the institutions for research and for clinical decision support.

The panel discussed the following issues:

- PheKB is a great resource. The Phenotype Execution and Modeling Architecture's (PheMA) Phenotype Authoring Tool is being integrated into PheKB as a beta version in an effort to streamline phenotype implementation. Such an effort has the potential of having measures identified in PheKB evolve into electronic Clinical Quality Measures (eCQMs). eMERGE also has OMOP modules as clinical quality measures (CQM) that are now searchable. These are examples of how eMERGE promotes the reuse of computable tools.
- eMERGE has its disparate phenotype databases at the sites, but it also has a central data resource such as the eMERGE Record Counter (eRC). eMERGE has also been depositing a selected set of variables into the central genome-wide association (GWA) database which can be shared among the eMERGE sites.

- eMERGE can provide some important use cases to EMR vendors, such as Epic Systems Corporation (Epic) and Cerner, and to their client stakeholders. The EMR vendors may use them, along with HL7 and other relevant groups, to create a standard. The degree to which it is supported by the EMR vendors will be critical.
- Phenotype definitions should contain outcome assessment tools, such as time course of development and progression of a condition.
- There seems to be a difference between the amount of phenotyping needed to drive research discovery versus the phenotyping needed in clinical settings. It might be necessary to study whether this difference exists as a general phenomenon and, if so, calibrate how different the phenotyping needs are in the two settings. This study may be carried out by surveying the different eMERGE sites which have a research repository that requires a full abstraction of the EMR and which are using their EMR as it is.
- Epic has been talking about creating a conglomerate of Epic EMR instances called Cosmos, which pulls all the EMR data from all the different Epic systems and stores them in a data repository. Unfortunately, the participants were not aware of any eMERGE sites that have participated in such conglomerates to date and we do not know if they are useful for research. However, it may be noted that research often needs data that are more consistent across sites than some clinical use cases.
- For projects like eMERGE, the primary goals are discovery and implementation. The amount of effort that it takes to promulgate standards may exceed the available resources. But in the future, it can be approached from studying the costs and gaps of not creating standards.
- eMERGE has to think about a way to combine the different phenotyping approaches by experimenting with alternative phenotyping strategies that are more efficient and faster at least for a few phenotypes.
- Summary
    - In PheKB, there are 154 phenotypes, half of which are eMERGE's 75 multi-site validated phenotypes; the Phase III phenotypes are yet to be added to PheKB.
    - It will be important to assess whether eMERGE wants to focus on fewer, more complicated phenotypes that have the greatest impact on health or focus on adding more phenotypes using simpler algorithms.
    - Since there is no quality standard that supports NLP, using NLP as part of multimodal phenotypes (i.e., including multiple EMR components to improve performance) has been the hallmark of many eMERGE phenotypes.
    - eMERGE is still challenged with the ability to analyze data across sites and how to transition these discoveries to clinical practice.

PANEL 2: Evidence Generation for Genomic Medicine

eMERGE Presentation (Gail Jarvik and Marc Williams)

eMERGE sites have different actionability and report preferences with most sites following the eMERGE list of genes and SNVs while a few sites are making additions and/or exclusions to this list. Of the 109 genes sequenced on the panel, 68 are thought by the eMERGE investigators to be clinically actionable. In addition to these 68 actionable genes being sequenced, 14 SNVs from this platform are also being returned by most sites. Even though all sites will return clinically actionable variants, the process of return is site-specific, with no two sites following the same practice.

eMERGE has found more “likely pathogenic” and “pathogenic” variants than expected in individuals without known disease. This has raised the question of whether these variants are truly pathogenic but low penetrance versus being truly benign. Re-phenotyping of individuals with such incidental findings can shed light on pathogenicity and penetrance. Another method to help assess penetrance is through family cascade testing followed by segregation analysis. A current challenge is that family history data are not consistently collected by providers nor captured well in most medical records. Thus, a standardized collection and formatting across all sites would be useful. Leveraging other types of data, such as geocoding and family history, can inform gene-environment analyses.

eMERGE generates evidence by following three different outcome types: 1) process outcomes that look at potential changes in healthcare utilization related to returning genetic information; 2) intermediate or surrogate outcomes, such as a biomarker indicating benefit/harm or adherence to a recommendation; and 3) clinical outcomes which include the actual impact on a health condition in a patient who receives an intervention based on the genomic result. Evidence that a process or surrogate outcome has a direct impact on health outcomes of interest can be strong, intermediate, or weak; to generate this evidence, standardizing outcome measurements is essential. The eMERGE Outcomes workgroup has developed and disseminated standardized data collection forms to all sites for this purpose. Challenges include the time it takes for outcomes to develop that will likely be longer than the duration of the eMERGE Phase III program, use of a single time point for outcomes assessment which might not be relevant to the genetic finding (6 months post-ROR), timing of sequencing and reporting (ending late in the course of eMERGE Phase III), and inferred attribution of outcome to ROR (relying on subjective assertions by each site). Future directions should address: 1) the potential for long-term follow-up of patients with ROR, to identify conditions or genomic results where health outcomes are more likely to accrue in a longer timeframe; 2) the possibility of obtaining sequencing results faster to allow longer follow-up; and 3) the development and testing of methods for attributing outcomes to ROR.

Reaction Presentation (Eric Boerwinkle)

The spectrum of evidence includes human genetics discovery, experimental discovery, and translation from research/discovery to clinical application. To transition genomics into clinical practice, the experimental data from discovery are essential to generating evidence needed to prove the value in patient care. Vast amounts of sequencing are done in the US, Europe and Asia which can be leveraged to address both the discovery and translation ends of the evidence spectrum. Leveraging the large amount of data being generated from patient care and adopting standardized data for clinical reporting will expand the data available and generate evidence.

Panel Discussion and Summary (Sharon Plon)

The panel discussed the following issues:

- Longitudinal data are essential to be able to understand the natural history and penetrance of different genotypes. The pediatric population could be an area of opportunity, since this information is being obtained at a young age and is readily available. However, many of the current actionable genes/variants do not express phenotypes until adulthood. Re-phenotyping and cascade testing are other areas where data are being generated and can be used to understand the natural history and penetrance.
- Network-wide clinical utility and cost effectiveness/economic outcomes and models are limited because the Network is not configured to be able to easily look across

institutions; instead, there are site-specific projects. Integration of clinical and genomic data will be a critical step to enable this type of analysis.

- Understanding the source of the information used is significant for making inferences. We should be very careful in making generalizable inferences across sites unless the differences in sources of information is understood.
- eMERGE has the challenge of recruiting scientists outside of the Network to become involved with the analysis of the large volume of data generated. Everyone recognizes that engaging other groups, such as payers, and working with other genomic medicine consortia, such as CSER, are essential and might facilitate crowdsourcing data analysis.
- The current sequencing timeline of eMERGE has not allowed for the return of results in time to be able to follow patients for a long period of time. In future projects, it will be important to ensure that results are returned to the patient with enough time to allow for following the patient longitudinally. This is critical to be able to answer fundamental questions about pathogenicity and penetrance of variants.
- Assessing the impact of returning negative results has not been a focus for the network (although it is being studied at a couple of eMERGE sites), but is an opportunity for the future.
- The focus of sequencing and reporting only 109 genes limits the ability for discovery of other variants. For the next phase, to aid discovery, eMERGE could consider sequencing an exome or genome, but only reporting on a specific set of clinically validated genes. Exome or whole genome sequencing are emerging as the preferred technologies for studies of this type as evidenced by the investment by the VA's MVP and national efforts like GenomeEngland.
- There is cautious adoption from the primary care community due to a lack of consistent standards for how to manage the results of the ACMG's list of genes. Studies are needed that help provide the evidence to define guidelines for management of these variants.
- Resources generated by eMERGE should be viewed as durable and should not end with the funding period. As other hypotheses are generated the data should remain a source for scientists to use to answer questions. A continuous reinterpretation of the sequencing data should be available.
- The question of including epigenetic, methylation, and chromatin immunoprecipitation (ChIP) data was raised. However, it was agreed that in a healthcare system setting, we do not want to include parameters about which we know very little regarding their importance in clinical care.
- Summary
    - Consistent generation of genomic and clinical data across sites is important to generate evidence for utility of genomic information in clinical practice.
    - Standardized measures for genomic medicine, such as clinical outcomes, clinical utility, cost-effectiveness, and actionability to inform ROR, are also essential.
    - Costs of clinical sequencing tests in the public sector are declining rapidly and AOU is an example of increasing scale of research protocols that will include ROR.
    - Clinical Genome Resource (ClinGen) is also approaching ways to expand involvement of a larger community to speed curation and variant interpretation.
    - Planning for the next phase of eMERGE should have thoughtful approaches to scaling efforts.

PANEL 3: EMR Integration of Genomic Results and Automated Decision Support

eMERGE Presentation (Casey Overby and Sandy Aronson)

eMERGE has been a great test bed for genomic discovery and implementation, and in the future, it may become a great test bed for a learning healthcare system, where clinical practice and research are increasingly collaborative and synergistic. This involves measuring outcomes and improving care over time, providing the most up-to-date evidence to providers, and delivering genomic knowledge to patients and providers at the point of care. There are four main challenges in using CDS as part of a learning healthcare system: (1) reproducibility in creating CDS that can be used at multiple sites; (2) timing and data quality; (3) diversity of population; and (4) replicability in accounting for evidence changes from previous research or clinical interpretation. Currently, eMERGE has been sharing phenotyping algorithms across sites and transitioning to storing data on the cloud via DNAnexus and GeneInsight. If CDS models and rules can also be shared via the cloud, they will become broadly accessible allowing for reproducibility. However, since CDS differs in timing when support is provided (which includes before, during, and after a clinical decision is made), side-effect risk screening algorithms will be required to help standardize the timing of support. For risk screening algorithms to be useful in clinical care, however, they will need to be developed in a replicable (in terms of inputs and data quality requirements) manner, so that they can be deployed in different CDS systems to ensure reliable performance. Diversity plays a role in clinical and research practice as we need to develop digital strategies to recruit populations while also minimizing over or under sampling of a given population, especially among racial and ethnic populations. To address this sample disproportionality, support for a range of recruitment strategies that include the utilization of multiple levels of health literacy will be required. Lastly, replicability is crucial because at any point in time, it is important to know what data were used and what evidence was available to get to the same answer. Though replicability is prevalent in research settings, it is not as widely assessed in clinical practice and it is important in tracking changes in variant classifications.

It is critical for eMERGE to have a CDS focus in the future, especially in demonstrating the economic and clinical value of genetics, incorporating electronic clinical decision support (eCDS), and accelerating the development and deployment of genetic CDS. In eMERGE Phase III, the focus has been on developing the clinical IT infrastructure and harmonizing data from the two different sequencing laboratories. One of the key ideas is to develop display-based CDS (such as SMART or FHIR) or event-based CDS (such as CDS Hooks) for a specific clinical area where genetics could materially improve the care of patients.

eMERGE may represent our best opportunity to accelerate the introduction of genomic knowledge into clinical care. However, it is important to recognize the differences between development of functionality for research use vs. functionality for clinical use. For example, developing eCDS, using genomic information, to improve clinical decisions or alter clinical process flows is a multistep process with significant validation requirements at each step. Proper validation will be necessary to ensure patient safety. If eMERGE pursues the development of clinical functionality, it is critical that it constructs a framework under which this can be done safely. This will involve establishing mechanisms to ensure validation processes are adequately resourced and functionality is not released for clinical use before it is ready.

Reaction Presentation (Blackford Middleton)

Currently, there is adoption of EMRs and data flowing across EMR systems, but the potential value to EMRs from CDS is not yet achieved. There are problems with interoperability and usability, but the main problem lies in the fact that EMRs do not include the rich phenotype and genotype information needed to develop a standard CDS that can be deployed across multiple sites. This is both an implementation and knowledge-sharing issue. The sociotechnical context needs to be reviewed when developing CDS, which includes the quality of data, knowledge, presentation, inference, and actionability. PGx CDS has the potential to be the most impactful CDS. The biggest challenge is that the knowledge representation for PGx is not standardized in a manner that allows for consistent performance in utilizing PGx algorithms in the clinic. The main recommendations are to develop standards so that data are shareable, recognize the potential of networked knowledge, and standardize CDS PGx delivery and formatting of information for processing and display. There is a need to scale across multiple instances of an EMR and multiple EMRs because patients have multiple sites of care across time.

Panel Discussion and Summary (Howard McLeod)

The panel discussed the following issues:

- Diagnostic CDS is a different type of CDS that might be a focus of eMERGE. For example, a CDS rule could be fired when chronic renal disease (CRD) appeared on the problem list so that, if that patient has sequence data, the physician can look up all the putative pathogenic variants in genes associated with CRD to potentially arrive at a genetic etiology that could inform treatment. This has not been done before, and could be a use-case that can be explored in eMERGE.
- Currently, these broad superficial CDS systems that are based on simple rules have high failure rates related to inaccurate alerts being triggered leading to alerts being ignored by the clinician. Instead, if a CDS system is validated, accurate, and includes both genetic data and other kinds of clinical information, physicians may be more amenable to interacting with the system, although end-user engagement research is needed to optimize the intervention.
- Alert fatigue is an issue that may be solved by developing more web apps for managing genetic results that focus on innovative ways to treat patients, consult with specialists instantly, and improve care and efficiency such as utilizing crowdsourcing for diagnoses.
- A standardized display that enables the assessment of a condition will save a lot of time for clinicians who need to make a decision when going through the EMR. This will allow easier accessibility to clinical data so that clinicians do not need to spend 20 minutes to find the data they need to make a decision.
- To improve CDS, it might be helpful to go directly to the source or consumers of the CDS and ask them what they prefer. Asking the physicians directly about which areas they have problems with through qualitative and quantitative surveys of physicians can help to develop useful CDS.
- CDS has multiple stakeholders, such as IT, patients, physicians, and financial stakeholders, but there are no defined units or measures of success for CDS implementation. The easiest metric is the measure of adoption, in terms of whether they ignore the alert or not. The metrics are dependent on the use-cases.
- One of the problems with CDS is that there may be other information that negates the importance of genomics, such as patient notes and environmental information. When you use genomics as a predictor of a PGx response, or other phenotypes, a great tool would be to incorporate other patient-specific factors that might increase or decrease the

importance of the genomic information in the CDS so that the alert is relevant. This leads to the idea of taking into consideration patients' values and opinions. In this respect, the CDS can be viewed as a resource for developing talking points for the physicians when they meet their patients.

- Summary
    - If we choose to build genetic-specific apps for managing genetic results in the EMR, eMERGE can make great progress in increasing the availability of updated genetic results and energizing the creation and adoption of FHIR-based standards. However, this type of support is likely to be less powerful in the context of specific clinical scenarios.
    - Genetic-specific apps could be made more powerful by incorporating additional information, such as family history, environment, and patients' notes, but this functionality would increase development scope.
    - Stimulating shareable CDS is achievable, but this leads to a need to maintain these resources.
    - Developing measures of success is critical for improving the quality of CDS.
    - Reproducibility should be possible across different EMR systems.
    - CDS needs to be implemented for all stakeholders, including physicians and patients outside the system.
    - The importance of high quality and accurate clinical data should be emphasized.

PANEL 4: Novel and Disruptive Opportunities in Genomic Medicine

eMERGE Presentation (Heidi Rehm and Iftikhar Kullo)

The Network has been dealing with the challenge of how to keep physicians and patients up-to-date with genomics and to be able to interpret genetic information in the EMR. There are various genetic data resources and apps for genetic knowledge, however there is still a need for developing approaches for determining when and how to update genetic knowledge and alert physicians and patients. Building on standards for structured reporting of genetic test report content that works with all sources of genetic data is essential for enabling continuous updates. In addition, we need to enhance the study of variant interpretation and potentially have real-time integration of population data with both genotype and phenotype defined. Additional data could be obtained through surveys or other mechanisms, such as neural networks, which is a form of deep-learning technologies that enables computers to learn from observational data. EMRs can improve our ability to annotate genomic variants, by providing access to a wide range of phenotype data. A method for labs and clinicians to access population (and individual level) patient data from many sources is essential for patient care. To accomplish this goal, eMERGE could help establish standardized data models and data sharing approaches, methods for integrating patient evidence into variant interpretation, and a process to help clinicians manage the uncertainty that results from the dynamic nature of genomic interpretation.

There are four novel sources of data that potentially have relevance to eMERGE research goals: direct-to-consumer (DTC) genomic test results, environmental variables (e.g. physical activity and dietary intake, geocodes), social media and crowdsourcing (especially for rare conditions), and patient-reported data (such as family history and medication adherence surveys). Ideally, a healthcare provider would be able to tap into all of these sources, as appropriate, through an application program interface (API) built on top of the EMR and other data sources. In addition, linking big 'omic' data to EMR-derived phenotypes can allow for new insights and discoveries. Stakeholders, namely patients, physicians, payers, and the public, can facilitate the collection of these data. Patient-centered rather than health care institution

centered data governance schemes may aid in addition of such data to EMRs There is an opportunity to improve integration of genomic CDS in EMRs, conduct economic modeling for cost-effectiveness, and develop strategies to impact public health genomics. This includes the linkage of eMERGE data to Health Information Exchanges (HIE) that enables doctors, nurses, pharmacists, other healthcare providers and patients to appropriately access and securely share a patient's medical information electronically.

Reaction Presentation (Matt Might)

With the present use of technology through social media, video-sharing platforms, and mobile apps, patients can receive or look for vast information related to their health. Self- or "peer-to-peer phenotyping", which is the concept of patients sharing their clinical manifestations via online social media tools (e.g., websites, Facebook, Twitter), has emerged and is being facilitated through these media. This internet/social media-driven case finding has created an online resource where patients share their cases and find others with the same experience; this is a pool of information that can be leveraged by the clinical and the research community. In addition, these new sources of data can provide novel data-driven genotype inferences. A good example is inferring Human Cytochrome P450 (CYP) variants from drug history, however we may also be able to infer genotypes from images or from patients' google search histories. Patients inquire about or share their symptoms and information through websites and these data could prove useful. Facebook and Twitter have already published studies on sentiment analysis of their users; these analyses can be expanded to additional mental or physical characteristics. Patient engagement is more prominent in genomic medicine than in other medical disciplines, and the development of disease-focused communities is thriving. We need to think of ways to leverage the efforts and impact of these groups in developing new therapies. Lastly, on the variant interpretation opportunities, deep learning can help us go from variants reported to protein structures. Having the actual molecular structure can be very useful for interpreting these variants of unknown significance (VUS), identifying toxicity predictors, and develop structural PGx that will allow for predictions of VUS interactions with drugs.

Panel Discussion and Summary (Dan Masys)

The panel discussed the following issues:

- The group inquired about the number of patient and disease-focused online communities. As far as we know there has not been a census of them, however in many instances there is more than one group per disease. An opportunity may exist for researchers to partner with Facebook and other social media providers to explore potential synergies.
- The role of the physician and the relationship with the patient should not be overlooked. So far, physician experiences with genomic medicine have not been optimal, therefore there is need for tools that aid the implementation of genomics in a way that does not burden the physician while benefitting patients. There should be different levels of tools, and engagement strategies that accommodate variable levels of genomic medicine literacy depending on the physician's specialty, training, and experience.
- There is opportunity for using social media and websites to find family history data. 23andMe has already started working on finding and connecting family members through their datasets.
- eMERGE has assembled a rich dataset of clinical and genomic data. However, there are a lot of other types of data that the consortium could receive from patients themselves (e.g., environmental interactions) that might explain issues such as variable penetrance.

- Efforts on scaling CDS are important. eMERGE is well-positioned to move this field forward.
- Walmart, CVS, and Walgreens have expanded their services to include providing primary care activities. These could be potential companies we could reach out to and engage the primary care world in genomics.
- Another opportunity would be to mine information from the large sequencing programs funded by NHGRI. The Centers for Common Disease Genomics consist of large numbers of individuals and samples that could be leveraged.
- There is an opportunity for participants to complete or correct a dataset that is known to be incomplete—the EMR. Patients usually know more than their record; therefore, we can update the EMR or EMR-derived research records with information we receive from them.
- Patient-reported outcomes from biobank participants could be useful and should be captured. In addition, building on relationships with patients and engaging them as partners, asking them what they want and what would help them understand their genomic information should be considered.
- Summary
    - Leveraging social media, video-sharing platforms, applications and the extensive use of online web tools by patients as novel sources of data can yield new findings or complement current methods.
    - Variant interpretation could be improved by monitoring patient data through surveys and potentially by using machine learning approaches to look for patterns in the data as they are being generated.

Conclusion and Recommendations (Dan Masys and Sharon Plon)

Electronic Phenotyping for Genomic Research

- Focus on developing better phenotyping methods and technologies, such as:
    - Increasingly automated phenotyping
    - Longitudinal phenotyping
    - Subtyping of diseases and disease outcomes
    - Continuum of disease severity rather than binary disease absent/present
    - Incorporating information on the time course of conditions to create more accurate phenotypes
    - Using innovative methods to infer phenotypes
    - Machine learning methods, including learning latent states using deep learning
    - Improved and more sharable natural language processing
    - Alternate approaches to manual phenotype validation
- Accommodate biases in healthcare data in an explicit and principled way
- Capture phenotypes produced by gene-environment interactions, such as massive hemolysis on exposure to drugs and foods in people with glucose-6-phosphate dehydrogenase deficiency.
- Study differences between phenotyping needed for research discovery vs. those needed for clinical care
- Focus on fewer phenotypes and experiment with alternative phenotyping strategies to improve speed and efficiency
- Increase collaboration across sites during phenotype development, exploiting common data models
- Ensure that other consortia are aware of the eMERGE phenotypes and how they might facilitate their research program

- Find more efficient ways to pool, normalize and analyze data across all eMERGE sites

Evidence Generation for Genomic Medicine

- Improve capture of standardized family history data across all sites; develop apps for collecting family history and incorporate the information into the EMR
- Serve as a source of evidence for ClinGen and other genomic medicine consortia
- Standardize or synthesize different study designs including ROR decisions because currently there is a wide spectrum of study designs across the different institutions of eMERGE, which reduces sample size and impairs joint analysis
- Seek appropriate balance between standardization and experimentation with different study designs since there is not yet enough information to standardize many aspects
- Develop and document best practices from the studies already completed in eMERGE
- Create data standards for new types of genomic medicine "data objects", such as genome sequencing data VCF formats, which incorporate quality metrics
- Study the value and impact of reporting negative results and study the definition of negative results in different contexts
- Learn from existing sites that are returning negative results to participants and develop methods to better educate patients on what negative results mean
- Assess longer-term outcomes of testing and results reporting
- Perform cascade testing and phenotyping of affected relatives

EMR Integration of Genomic Results and Automated Decision Support

- Work with electronic health record vendors for EMR standardization by inviting vendors to NIH workshops
- Develop tools and standardized displays to synthesize and present information at the point of care so that physicians do not have to hunt for information to make a decision
- Develop user-centered designs through both display-based and event-based eCDS
- Build foundations that promote shareable eCDS, which includes the knowledge representation of complex CDS and enhancement of existing knowledge repositories
- Narrow the scope of eMERGE to developing CDS in a few specific, high-priority clinical areas to avoid spreading resources too thin
- Assess the value of genomic CDS that embodies deep knowledge, which is the ability to recognize patterns from different types of datasets and make predictive analysis, rather than simple, superficial rules so as to confront alert fatigue and other usability issues
- Develop genomic apps as supplements to clinical systems and incorporate physician preferences into CDS
- Develop closed-loop CDS, which contains automated outcome assessment tools and allows determination of whether users followed the guidance to assess utility
- Explore patient-specific factors that might increase or decrease the importance of genomic information in the CDS to improve relevance of alerts to specific patients
- Correlate clinicians' attitudes and patients' beliefs with actions taken by patients after ROR
- Evaluate the effect of standard and nonstandard approaches to delivering results (e.g., direct-to-patient/participant) on physician-patient relationships including what information can be dispersed using alternative technologies vs. what needs to be conveyed through a genetic counselor
- Assess the financial impacts on health systems, patients, and physicians
- Identify and focus on a few high priority areas where genomics can improve patient care
- Address scaling of phenotyping/interpretation/reporting as its own research problem

- Develop roadmap for naïve adopters of genomic data/CDS implementation in EMR
- Develop standard extract of EMR for research

Novel and Disruptive Opportunities in Genomic Medicine

- Incorporate genomic data derived or inferred from external sources, such as patient-contributed data, environment (geocoding), direct-to-consumer data, social media, "peer-to-peer phenotyping" (patients sharing their clinical manifestations via social media), family history, and online disease-focused patient communities
- Perform real time variant interpretation that incorporates patient data as well as matches publicly available knowledge sources to the patient's variants
- Develop methods to efficiently re-interpret genomic results over time
- Develop methods to automate interpretation via analysis pipelines
- Enhance clinical methods for assessing pathogenicity and variant penetrance
- Collaborate efficiently with other research consortia, especially for pragmatic trials and rare variant characterization, and take advantage of opportunity for eMERGE to add "genomic dimensions" to their research programs, such as AOU, for the advancement of standards-based data exchange
- Address challenges in public health genomics, such as linkage of eMERGE data to HIE
- Develop formal approaches to representing and accommodating uncertainty in the analysis and interpretation of the data
- Link EMR-derived phenotypes with other classes of -omics data
- Infer genotypes from non-traditional data sources, such as drug experience, images, internet search history
- Apply deep learning techniques to the characterization of VUS, drug targets, and toxicity predictions associated with primary genomic data
- Communicate both data and their interpretation directly to patients, through family sharing, health literacy, incorporation of patient preferences, and automation
- Limit the use of genetic professionals to disclose positive results in high-penetrance genes and use short educational materials and tutorials to disclose negative results
- Embrace the idea of clinical reengineering and implementation science as opposed to iterating using current traditional processes and facing the same barriers
- Evaluate the impact of patient engagement with the science on patients' understanding their own disorders as a measurable outcome, as well as the impact of disorder-focused patient communities
- Partner with participants to support stakeholder-centered participatory design of RoR efforts
- Encourage patient-centered data governance and develop or encourage development of apps for patient self-phenotyping
- Develop innovative ways to present sequence information to general physicians, especially those early in training, by identifying problems that physicians are facing and creating pragmatic solutions
- Develop methods to increase health literacy in both patients and physicians
- Evaluate the legal and ethical implications of directly contacting relatives of patients potentially harboring deleterious variants rather than having to go through the patient
- Assess crowdsourcing of variant interpretation

The workshop presentation slides and video recordings have been posted online at https://www.genome.gov/27569445/.