Published in final edited form as:
Genet Med. 2016 September ; 18(9): 906–913. doi:10.1038/gim.2015.187.

The Geisinger MyCode Community Health Initiative: an electronic health record-linked biobank for Precision Medicine research

David J. Carey[*], **Samantha N. Fetterolf**, **F. Daniel Davis**, **William A. Faucett**, **H. Lester Kirchner**, **Uyenlinh Mirshahi**, **Michael F. Murray**, **Diane T. Smelser**, **Glenn S. Gerhard**[¶], and **David H. Ledbetter**
Geisinger Health System, 100 N. Academy Avenue, Danville, PA 17822

Abstract

Purpose—Geisinger Health System (GHS) provides an ideal platform for Precision Medicine. Key elements are the integrated health system, stable patient population, and electronic health record (EHR) infrastructure. In 2007 Geisinger launched MyCode®, a system-wide biobanking program to link samples and EHR data for broad research use.

Methods—Patient-centered input into MyCode® was obtained using participant focus groups. Participation in MyCode® is based on opt-in informed consent and allows recontact, which facilitates collection of data not in the EHR, and, since 2013, the return of clinically actionable results to participants. MyCode® leverages Geisinger's technology and clinical infrastructure for participant tracking and sample collection.

Results—MyCode® has a consent rate of >85% with more than 90,000 participants currently, with ongoing enrollment of ~4,000 per month. MyCode® samples have been used to generate molecular data, including high-density genotype and exome sequence data. Genotype and EHR-derived phenotype data replicate previously reported genetic associations.

Conclusion—The MyCode® project has created resources that enable a new model for translational research that is faster, more flexible, and more cost effective than traditional clinical research approaches. The new model is scalable, and will increase in value as these resources grow and are adopted across multiple research platforms.

Keywords

biobank; genomics; electronic health records; genetic association

[*]corresponding author: David J. Carey, PhD, Weis Center for Research, 100 N. Academy Avenue, Danville, PA 17822-2601, Phone: 570-271-6659, djcarey@geisinger.edu.

[¶]current address: Department of Medical Genetics and Molecular Biochemistry, Temple University School of Medicine, Philadelphia, PA 19140

Supplementary information is available at the *Genetics in Medicine* website

INTRODUCTION

Geisinger Health System (GHS), an integrated health system located in north central and northeastern PA, possesses a unique combination of resources to accelerate clinical translational research [1–3]. As an integrated system GHS incorporates within a single not-for-profit institution a large primary care and specialty group practice, more than 70 care sites (including 2 tertiary-quaternary care hospitals and other inpatient facilities and a network of community-based clinics), and an insurance operation (see Online Supplementary Information for additional details). This integration creates a more seamless approach to care and more complete capture of episodes of care. Much of the population served by GHS is relatively non-transient, with low rates of migration into or out of the area, a large number of life-long residents, and many multi-generation families. GHS was an early adopter of electronic health record (EHR) systems (beginning in 1996); its EHR is fully implemented across all sites of care. To enable use of these data for clinical care and research GHS created an enterprise data warehouse that consolidates data from the Epic EHR and other sources. While not individually unique, the combination of integrated health system, stable patient population, and health information technology provide a powerful platform for precision medicine, an approach to treatment and prevention that takes into account individual variation in genes, environment, and lifestyle [4,5].

To harness these resources to investigate the molecular and genetic bases of health and disease, in 2007 GHS launched a project now known as the MyCode® Community Health Initiative (MyCode®) to create a system-wide biorepository of blood, serum and DNA samples for broad research use, including genomic analysis. Data obtained from analysis of MyCode® samples can be linked to information in participants' digital health records. Use of these existing data provides enormous flexibility in the types of research questions that can be investigated, and at much lower cost and accelerated time scale compared to traditional approaches.

Here we describe the creation of the MyCode® biorepository and its operation, and examples of how it can be used for translational genomics research.

MATERIALS AND METHODS

Focus groups and survey

Before MyCode enrollment began focus groups with potential participants were held to assess their knowledge, attitudes and likely participation in a biobanking program. Focus group results were validated by means of a self-administered questionnaire. Focus groups of MyCode participants were held in 2013 to discuss return of research results and placement of research results in the EHR. Details on the focus groups and survey are provided in the online Supplementary Information.

MyCode® participation

During an outpatient visit to a Geisinger clinic eligible patients meet with a research assistant or a member of the clinic staff who explains the project, answers questions, and invites them to consider participating in MyCode®. Interested patients sign a written consent/HIPAA authorization. By enrolling in MyCode® participants agree to provide blood samples for broad research use and permit access to data in their EHR for research use. The consent form also states that 1) participation in the program is voluntary; 2) patients may derive no direct benefit from participation; 3) their decision regarding participation will have no impact on their medical care at Geisinger; 4) research done with their banked samples could include analysis of their genes; and 5) MyCode® investigators will take steps to protect their privacy and security of their information. Early versions of the MyCode® consent form contained a check box to indicate whether participants were willing to be re-contacted regarding the biobanking program or other research projects. More than 90 percent of consenting patients agreed to be re-contacted. Later versions of the consent form eliminated the check box, but stated that consenting participants agreed to be re-contacted. In 2012 the protocol was amended to allow enrollment of pediatric patients, with parental or guardian consent, and child assent for enrollees over age 7. An addendum to the consent allows family members to be linked for research purposes.

Prior to October, 2013, the consent form stated that results of research done with participant's samples would not be placed in their medical record. The protocol and consent did allow for the "small chance that researchers could discover something that might be important" for their medical care, and in that case they would be contacted "to see if you want to learn more". In 2013 the protocol and consent were amended to allow return of medically actionable findings.

Sample collection and processing

When a participant enrolls in MyCode® this information is entered into their GHS EHR. This creates an automatic order for the collection of MyCode® samples that is activated when the participant has blood drawn for clinical testing in a GHS outpatient lab. The MyCode® blood order is triggered in response to future outpatient blood draws (maximum 12 times per year), resulting in serial sample collection. MyCode samples are transported to a central processing lab in the Geisinger Department of Laboratory Medicine and then to the genomics core laboratory for final processing. For the initial MyCode blood draw 4 ml of EDTA-whole blood and two 4 ml serum-separator tube samples are obtained. For subsequent blood draws only serum is collected. One ml aliquots of whole blood are used for DNA extraction on a Qiagen QiaSymphony robot. DNA is eluted into 2-dimensional barcoded tubes; purity and yield of DNA are determined by ultraviolet spectroscopy. Samples are given a unique study identification number. A secure key linking the sample identification number to a specific patient is maintained by the MyCode® team. Additional details are provided in the online Supplementary Information.

Linking samples to clinical data

Clinical data are linked to samples or molecular/genomic data obtained from analysis of samples by means of the unique MyCode® identification number. For most studies, de-identified data are used. The linking of MyCode® samples or data to clinical data for research studies is accomplished through the use of a data broker. A data broker is empowered to work with identified data and to provide them to investigators in a manner that conforms to Institutional Review Board (IRB) and other approvals, and is bound to maintain the privacy of the personal information. A research data core was created to model EHR, billing, and administrative data in Geisinger's data warehouse and other sources, extract data for use by researchers, de-identify data when necessary, and develop and validate phenotypes based on these data. Phenotype algorithms corresponding to clinical traits of interest are developed using concepts from various source vocabularies to define the presence, progression, treatment, and response of various diseases.

Genomic data and analysis

Genotype data from from MyCode participants were used for genetic association analyses of previously reported single nucleotide polymorphisms (SNPs) for cardiovascular disease [6–10], type 2 diabetes [7,11–14], and obesity [7,15–19]. Clinical phenotypes were determined using validated phenotype algorithms that use CPT codes, ICD-9 codes, laboratory results, and vital signs to define cases, controls, and excluded individuals (see online supplementary information for additional details). Associations were calculated by logistic regression, controlling for sex and current age, using an additive genetic model. To determine effects of rare *APOC3* variants on blood lipid traits [20,21], mean lifetime values for triglycerides, low density lipoprotein cholesterol, and high density lipoprotein cholesterol were determined using EHR-derived lab values. For individuals on lipid-lowering medications mean lipid values before the start of therapy were used. One-way analysis of variance with Dunn's multiple comparison test was used to calculate two-tailed p values for *APOC3* variant carriers and non-carriers.

RESULTS

Developing a system-wide biobank for broad research use

The goal was to create a central biorepository of blood, serum, and DNA samples from GHS patients that could be linked to information in the EHRs of the sample donors, under conditions that would allow the samples and data to be used for broad, future research, including genetics. In light of the ethical and legal considerations and the logistical challenges associated with creating a sustainable project of this type, the program was developed in stages, using participant engagement to develop and guide consenting strategies. That engagement began with initial assessment of patient attitudes, followed by development and evaluation of a pilot program, and finally large-scale recruitment (Figure S2).

Patient Attitudes about Biobanking and Genomics Research

Prior to initiating the biobanking program a focus group study of randomly selected Geisinger patients was conducted to explore awareness and attitudes toward health care research and genomic research using biobanked specimens, reactions to proposed consent language for such studies, and use of health information for research.

A summary of observations from the focus groups is in Table S1. Overall, the participants were highly supportive of medical research, and took pride that such research was conducted in their community. Support was expressed regardless of whether they derived direct short-term benefit, and no compensation was expected. Potential concerns were that participation in the research be voluntary, that there be no repercussions for not participating, and that safeguards to protect confidentiality be implemented. Most participants, including more than 80 percent of women, expressed a desire to be contacted if findings related to their own health arose from the research. The most negative comments were directed at the proposed consent form language (Table S2). These were judged to be unnecessarily complex given the simplicity of their involvement in the research.

These observations were validated in a self-administered questionnaire that was mailed to 500 randomly selected Geisinger patients. Similar to the focus group results, attitudes towards research were highly favorable. Seventy-five percent of respondents agreed or strongly agreed that creating a biobank for research was a good idea, and 77 percent supported genetic research at Geisinger (Table S3).

The MyCode® Initiative

Based on these results and consultation with the Geisinger IRB, a pilot biobanking program was initiated in 6 Geisinger outpatient clinics. Experience from the pilot study was used to design the MyCode® protocol. Although some operational details have been modified over the course of the project, the overall process has remained essentially the same, and is outlined in Figure 1. A goal was to use when possible existing infrastructure or processes, especially for sample collection and participant tracking, to maximize operational efficiencies and minimize costs.

MyCode® enrollment has been ongoing since early 2007, and is based on opt-in informed consent obtained in most cases during a face-to-face conversation with a research consenter in a Geisinger primary care or specialty clinic. MyCode® participant accrual is shown in Figure S3. By the end of September, 2015 more than 90,000 Geisinger patients had enrolled, including more than 3,600 pediatric patients enrolled through parental or guardian consent. The rate of enrollment increased in 2014 as a result of a deliberate scale-up; currently, approximately 4,000 new participants are added per month. The consent rate of patients who are approached for participation is high, with an overall consent rate of more then 85 percent. The protocol permits participants to withdraw from the project at any time. Since the inception of the project approximately 2% have withdrawn. The age distribution of consented adult participants (Figure S4) approximates that of the GHS outpatient population, but with under-sampling of adults less than 30 years old, and over-sampling of patients in the 60–89 age ranges. Compared to the age distribution of the regional population, individuals over age 50 years are enriched in the GHS outpatient and MyCode cohorts. Because eligibility to participate in MyCode® does not depend on a particular condition or diagnosis and participants have been enrolled from a large number of diverse clinics, and the consent rate for participation is high, MyCode® participants provide a reasonably good sampling of the Geisinger adult patient population. Table S4 shows the most common diagnoses in the GHS outpatient and MyCode cohorts. The rank order of frequency is nearly identical in the two cohorts, although MyCode participants are enriched for most diagnoses.

The most significant change to MyCode® occurred in 2013 when it was realized that analysis of MyCode® samples provided opportunities for finding medically actionable results, and that sharing such results with participants was consistent with Geisinger's health care mission. Before a systematic return of results program was initiated a series of participant focus groups were held that probed attitudes about this topic. These revealed a strong consensus favoring the return of results to participants and their clinicians, and placement of medically actionable results in the participant's EHR, with appropriate educational support to healthcare providers and patients.

In light of these considerations the protocol and consent were amended in 2013 to state that "researchers may find information that could be specifically important to your health care", and that if such information is found "we may share that information with both you and your doctor and place it in your medical record", and that educational materials and clinical support would be provided to clinicians and participants when results were returned. The consent also stated that non-medically actionable results would not be returned.

Nearly all MyCode® blood samples are obtained at the time of an outpatient clinical blood draw using existing clinical infrastructure. This process has several advantages: 1) it eliminates the need for an extra venipuncture to collect research samples; 2) it allows collection at nearly all Geisinger care sites and transportation of samples to a central processing laboratory; 3) the cost for sample collection reflects only incremental resources used to collect the research samples after clinical samples are obtained; 4) blood sample collection is done under a Clinical Laboratory Improvement Amendment (CLIA)-certified process and with quality controls consistent with clinical sample collection; and 5) serial samples are obtained whenever a participant has an outpatient clinical blood draw.

MyCode® samples are retrieved, processed, stored and tracked in the Geisinger genomics laboratory using standardized processes and quality control measures. A unique study identification number is assigned to all consented MyCode® participants and used to identify samples in the biobank. Beginning in January, 2015 DNA extraction was carried out under CLIA protocols, to allow the samples to be used for validation of clinically actionable findings.

Access to biobanked samples or data for specific research projects is determined by the MyCode® Governing Board, which has representatives from Geisinger research and clinical departments and non-scientist/non-physician members. The Governing Board evaluates requests on the basis of scientific merit and potential impact of the research and availability of samples. As of July, 2015 54 requests for MyCode® samples or data use were approved; more than 50,000 samples have been used for molecular analyses.

Using samples and data for association studies

Consent to participate in MyCode® allows information collected during clinical encounters at a Geisinger care site to be used for research and linked to data obtained from analysis of MyCode® samples. The duration of EHR records for current MyCode participants is a median of 12.0 years, with a range of 0–221 months (Figure 2A). The number of clinical encounters recorded in the EHR for MyCode participants is a median of 60, with a range of 1–1,153 (Figure 2B, Table 1). The number of encounters is greater for participants older than 55 years. Consistent with the large number of clinical encounters are many records for clinical lab values and vital signs (Table 1).

MyCode® DNA samples have been used to generate high-density genotype data (Table S4). As part of a collaboration with the Regeneron Genetics Center DNA samples are also used for exome sequencing. The genomic data are stored in a secure database and made available to investigators, contingent on approval by the MyCode® Governing Board.

To demonstrate the utility of MyCode® data for genetic association studies we replicated previously reported associations of SNPs with obesity, cardiovascular disease and diabetes. Cases and controls for these studies were identified using EHR data (details are provided in the online supplemental information). As shown in Table 2, SNPs in the 9p21 [6–10], *TCF7L2* [7,11–14], and *FTO* [7,15,16] and *MC4R* [17–19] loci were significantly associated with cardiovascular disease, type 2 diabetes, and body mass index, respectively. The calculated odds ratios were of the same magnitude and direction as previously reported.

We also examined the relationship between previously reported rare variants in the *APOC3* gene and blood lipid levels [20,21]. The R19X stop-gain mutation (rs76353203), IVS2+1G/A splice site variant (rs138326449), and A43T missense variant (rs147210663) were identified in 15 (0.13%), 52 (0.45%) and 4 (0.03%) of 11,449 individuals, respectively, with both genotype and blood lipid data (Figure 3 and Table S5). In the MyCode cohort 1 in 160 individuals were heterozygous carriers of one of these variants, which is similar to the prevalence of 1 in 150 reported in an earlier study[20]. Consistent with previous reports, heterozygous carriers of R19X or IVS2+1G/A variants had significantly lower serum triglyceride levels than non-carriers; triglyceride levels in A43T carriers were lower than the mean non-carrier value, but this did not reach statistical significance because of the small sample size. R19X and IVS2+1G/A carriers had significantly higher HDL-cholesterol levels than non-carriers. None of the variants had a significant effect on LDL-cholesterol levels.

DISCUSSION

The resources created by MyCode® provide a powerful platform for translational research. At the core is a large, central repository of biological samples from participants who consent to the use of their samples for broad, future research use. A timeline that highlights key events in the creation of MyCode® is shown in Table S7.

The ability to link existing, large sets of molecular and clinical data creates an efficient and flexible vehicle for the discovery and validation of molecular and genetic factors associated with clinical traits. A wide range of research questions can be studied faster and at reduced costs compared to conducting the same studies using traditional approaches. The analyses reported here, and others that have been published [22], were completed in a matter of weeks, as opposed to years that would be required using conventional approaches. The value of these resources will increase as the MyCode® biobank and molecular data, and associated clinical data continue to grow. The long-term goal is to invite every active GHS patient to consider participation in MyCode®, which would create a cohort of more than 500,000 individuals. An EHR-linked biorepository provides an attractive model to advance the goals of the “Precision Medicine Initiative”, unveiled in early 2015 by the White House and government research leaders. Precision Medicine is defined as “an approach to disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle in each person” [5], based on research with a cohort of individuals who “give consent for extensive characterization of biologic specimens . . . linked to their electronic health records” [4].

This model for translational research is scalable. As similar biorepositories and EHR data become widely available they can be combined to create very large data sets to address a larger and more diverse range of clinical research questions. The feasibility of such cross-platform scaling is demonstrated by the success of the NIH-funded eMERGE (electronic Medical Records and GEnomics) Network [23–25]. The creation of similar large-scale biobanks is occurring in the U.S. and elsewhere [26–29].

The processes adopted to create the MyCode® biobank have several advantages. Leveraging existing infrastructure, such as health information technology to identify and track eligible participants and create automatic blood orders, and clinical infrastructure to collect, transport and track samples, creates substantial increases in efficiency and reductions in cost compared to stand-alone processes. It also facilitates collection of serial blood samples, which is valuable for studies that investigate, for example, changes in biomarkers related to a clinical event.

The use of opt-in consent for broad research use allows a greater range of activities to be conducted under a single protocol. The MyCode® consent provides permission to re-contact, which enables research studies that require data that cannot be obtained from the EHR, such as environmental exposures, nutrition information, or physical activity measures. In 2013 the MyCode® protocol and consent were updated to allow for the return of medically actionable research findings to participants and their medical providers.

In light of this broad range of activities it remains important to engage participants to elicit their perspectives on ethical and practical questions regarding research and integration of genomics into clinical practice. This is especially important in light of the fact that GHS serves a mostly rural population with little other direct exposure to medical research. Based on the overwhelmingly positive responses in focus groups and survey results, we are confident we are respecting the wishes of the participants. Community acceptance of the program is also reflected in the high rates of consent by individuals invited to participate. High levels of participant support for similar projects have been reported by others [30,31].

Internal oversight of the GHS biobanking program is provided by the MyCode® Governing Board. An additional layer of independent oversight is provided by an Ethics Advisory Council made up of external experts in genetics and ethics as well as members from the local community who are MyCode® participants, and a separate Return of Results Oversight Committee comprised of experts in genetics, clinical medicine, and bioethics. MyCode samples and data can be shared for collaborative research studies. The MyCode Governing Board reviews and approves all uses of MyCode samples and data, with additional review and approval of the Geisinger IRB, if needed.

While strengths of the MyCode® project are summarized above, several limitations should be noted. More than 95 percent of the regional population served by GHS is of white European ancestry. Thus, MyCode® provides limited opportunities to study health disparities among racial and ethnic groups, or differences in genetic variant frequencies and their impact on health-related traits.

Nearly all phenotype data used for studies that utilize MyCode® samples or data are derived from data collected during participant's clinical encounters with the health system. While this provides enormous breadth and flexibility with respect to research questions that can be addressed, it also requires special care to account for "noise" in these data, caused, for example, by misclassification (e.g. through incorrect use of diagnostic codes), data entry errors, and missing data. In some cases important information is available only in text-based notes or other unstructured sources, which requires the use of natural language processing to extract the data. The use of rigorous and validated phenotype algorithms is therefore needed to reduce or eliminate effects of these data limitations. The validity of this approach has been well documented. The eMERGE Network, of which GHS is a participant, has pioneered the use of EHR data for electronic phenotyping for genomics research [23–25].

Embedding these research processes into a health care system helps reduce barriers between research and clinical activities. GHS has adopted the Learning Health System concept, which strives to use the system's resources to drive continuous improvement and innovation in health and health care, "with best practices seamlessly embedded in the delivery process and new knowledge captured as an integral by-product of the delivery experience" [32]. Inherent to MyCode® is the use of information obtained during the health care delivery process. The resources created by MyCode® also enable the facile utilization and testing of genomic and biomarker data to improve health.

Relying on existing clinical infrastructure also places constraints on the MyCode® process. MyCode® consenting and sample collection are designed to leverage, but not interfere with, clinical care activities. Thus, the logistics of patient enrollment are tailored to existing workflows in clinics where consenting is occurring. Also, use of a "passive" sample collection process (where obtaining samples is dependent on a clinical blood draw order) often creates a lag between consenting and research sample collection. Under the current MyCode® process, samples are obtained from 40 percent of participants within 1 month of consent; the average time between consent and sample collection is about 3 months.

The samples and molecular data generated by MyCode® have been used in a large number of research studies. They have also been leveraged to generate external research funding, and enabled Geisinger's participation in research collaborations. These include the eMERGE Network, a consortium to conduct research that combines DNA biorepositories with EHR systems for genetic research, and a collaboration with the Regeneron Genetics Center to do exome sequence analysis of MyCode® participants. These further increase the value of MyCode® to address important clinical research questions.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors gratefully acknowledge Ryan Colonie for help in developing and maintaining the MyCode database, and Dr. Marylyn Ritchie for reading and making helpful comments on the manuscript. Initial financial support for the creation of the MyCode biobank came from a grant from the Ben Franklin Technology Development Fund of the Pennsylvania Department of Community and Economic Development. Expansion of the MyCode biobank since the beginning of 2014 was made possible by funds provided by the Regeneron Genetics Center, Tarrytown, NY.

References

1. Slotkin JR, Casale AS, Steele GD Jr, Toms SA. Reengineering acute episodic and chronic care delivery: The geisinger health system experience. Neurosurg Focus. 2012; 33(1):E16.doi: 10.3171/2012.4.FOCUS1293 [PubMed: 22746233]
2. Paulus RA, David K, Steele GD. Continuous innovation in health care: Implications of the geisinger experience. Health Aff (Millwood). 2008; 27(5):1235–1245. [PubMed: 18780906]
3. Casale AS, Paulua RA, Selna MJ, et al. "ProvenCare^{SM}" A provider-driven pay-for-performance program for acute episodic cardiac surgical care. Annals of Surgery. 2007; 249:613–623. [PubMed: 17893498]
4. Collins FS, Varmus H. A new initiative on precision medicine. N Engl J Med. 2015; doi: 10.1056/NEJMp1500523
5. Precision medicine initiative. www.nih.gov/precisionmedicine. Updated 2015
6. Helgadottir A, Thorleifsson G, Manolescu A, et al. A common variant on chromosome 9p21 affects the risk of myocardial infarction. Science. 2007; 316(5830):1491–1493. 1142842 [pii]. [PubMed: 17478679]
7. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007; 447(7145):661–678. nature05911 [pii]. [PubMed: 17554300]
8. Helgadottir A, Thorleifsson G, Magnusson KP, et al. The same sequence variant on 9p21 associates with myocardial infarction, abdominal aortic aneurysm and intracranial aneurysm. Nat Genet. 2008; 40(2):217–224. DOI: 10.1038/ng.72 [PubMed: 18176561]
9. Samani NJ, Erdmann J, Hall AS, et al. Genomewide association analysis of coronary artery disease. N Engl J Med. 2007; 357(5):443–453. NEJMoa072366 [pii]. [PubMed: 17634449]
10. Schunkert H, Gotz A, Braund P, et al. Repeated replication and a prospective meta-analysis of the association between chromosome 9p21.3 and coronary artery disease. Circulation. 2008; 117(13):1675–1684. DOI: 10.1161/CIRCULATIONAHA.107.730614 [PubMed: 18362232]
11. Grant SF, Thorleifsson G, Reynisdottir I, et al. Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. Nat Genet. 2006; 38(3):320–323. ng1732 [pii]. [PubMed: 16415884]
12. Groves CJ, Zeggini E, Minton J, et al. Association analysis of 6,736 UK. subjects provides replication and confirms TCF7L2 as a type 2 diabetes susceptibility gene with a substantial effect on individual risk. Diabetes. 2006; 55(9):2640–2644. 55/9/2640 [pii]. [PubMed: 16936215]
13. Lyssenko V, Lupi R, Marchetti P, et al. Mechanisms by which common variants in the TCF7L2 gene increase risk of type 2 diabetes. J Clin Invest. 2007; 117(8):2155–2163. DOI: 10.1172/JCI30706 [PubMed: 17671651]
14. Florez JC, Jablonski KA, Bayley N, et al. TCF7L2 polymorphisms and progression to diabetes in the diabetes prevention program. N Engl J Med. 2006; 355(3):241–250. 355/3/241 [pii]. [PubMed: 16855264]
15. Frayling TM, Timpson NJ, Weedon MN, et al. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science. 2007; 316(5826):889–894. 1141634 [pii]. [PubMed: 17434869]

16. Hunt SC, Stone S, Xin Y, et al. Association of the FTO gene with BMI. Obesity (Silver Spring). 2008; 16(4):902–904. DOI: 10.1038/oby.2007.126 [PubMed: 18239580]

17. Loos RJ, Lindgren CM, Li S, et al. Common variants near MC4R are associated with fat mass, weight and risk of obesity. Nat Genet. 2008; 40(6):768–775. DOI: 10.1038/ng.140 [PubMed: 18454148]

18. Qi L, Kraft P, Hunter DJ, Hu FB. The common obesity variant near MC4R gene is associated with higher intakes of total energy and dietary fat, weight change and diabetes risk in women. Hum Mol Genet. 2008; 17(22):3502–3508. DOI: 10.1093/hmg/ddn242 [PubMed: 18697794]

19. Renstrom F, Payne F, Nordstrom A, et al. Replication and extension of genome-wide association study results for obesity in 4923 adults from northern sweden. Hum Mol Genet. 2009; 18(8):1489–1496. DOI: 10.1093/hmg/ddp041 [PubMed: 19164386]

20. Crosby J, Peloso GM, et al. TGHDL Working Group of the Exome Sequencing Project, National Heart Lung and Blood Institute. Loss-of-function mutations in APOC3, triglycerides, and coronary disease. N Engl J Med. 2014; 371(1):22–31. DOI: 10.1056/NEJMoa1307095 [PubMed: 24941081]

21. Pollin TI, Damcott CM, Shen H, et al. A null mutation in human APOC3 confers a favorable plasma lipid profile and apparent cardioprotection. Science. 2008; 322(5908):1702–1705. DOI: 10.1126/science.1161524 [PubMed: 19074352]

22. Mirshahi T, Murray MF, Carey DJ. The metabolic syndrome and DYRK1B. N Engl J Med. 2014; 371(8):784–785. DOI: 10.1056/NEJMc1408235#SA1 [PubMed: 25140973]

23. Crawford DC, Crosslin DR, Tromp G, et al. eMERGEing progress in genomics-the first seven years. Front Genet. 2014; 5:184.doi: 10.3389/fgene.2014.00184 [PubMed: 24987407]

24. Gottesman O, Kuivaniemi H, Tromp G, et al. The electronic medical records and genomics (eMERGE) network: Past, present, and future. Genet Med. 2013; 15(10):761–771. DOI: 10.1038/gim.2013.72 [PubMed: 23743551]

25. 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 Med Genomics. 2011; 4:13–8794–4–13. DOI: 10.1186/1755-8794-4-13 [PubMed: 21269473]

26. Olson JE, Bielinski SJ, Ryu E, et al. Biobanks and personalized medicine. Clin Genet. 2014; 86(1):50–55. DOI: 10.1111/cge.12370 [PubMed: 24588254]

27. Olson JE, Ryu E, Johnson KJ, et al. The mayo clinic biobank: A building block for individualized medicine. Mayo Clin Proc. 2013; 88(9):952–962. DOI: 10.1016/j.mayocp.2013.06.006 [PubMed: 24001487]

28. Sudlow C, Gallacher J, Allen N, et al. UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 2015; 12(3):e1001779.doi: 10.1371/journal.pmed.1001779 [PubMed: 25826379]

29. Kang B, Park J, Cho S, et al. Current status, challenges, policies, and bioethics of biobanks. Genomics Inform. 2013; 11(4):211–217. DOI: 10.5808/GI.2013.11.4.211 [PubMed: 24465232]

30. Tomlinson T, De Vries R, Ryan K, Kim HM, Lehpamer N, Kim SY. Moral concerns and the willingness to donate to a research biobank. JAMA. 2015; 313(4):417–419. DOI: 10.1001/jama.2014.16363 [PubMed: 25626040]

31. Rahm AK, Wrenn M, Carroll NM, Feigelson HS. Biobanking for research: A survey of patient population attitudes and understanding. J Community Genet. 2013; 4(4):445–450. DOI: 10.1007/s12687-013-0146-0 [PubMed: 23605056]

32. olsen, L.; aisner, D.; McGinnis, JM., editors. Institute of medicine. The learning healthcare system. Workshop summary. http://www.iom.edu/Reports/2007/The-Learning-Healthcare-System-Workshop-Summary.aspx. Updated 2007

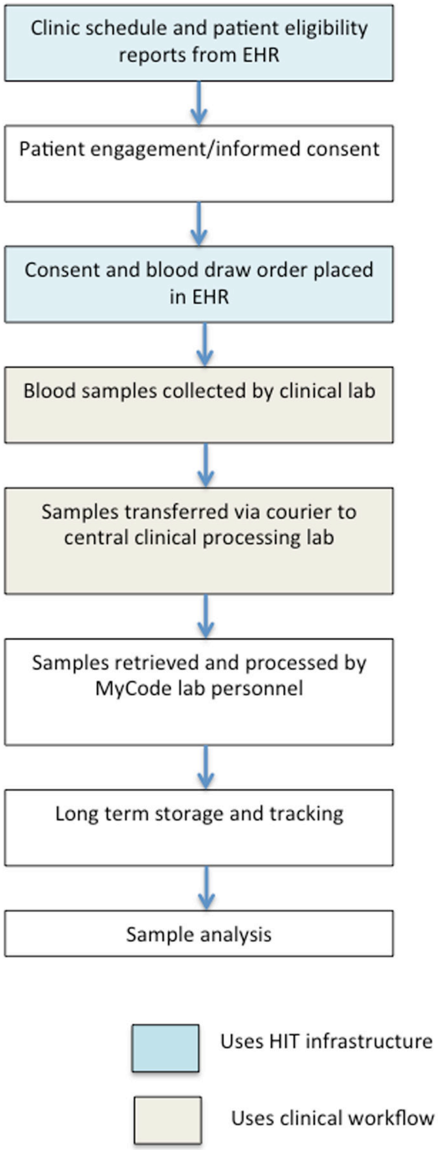
Clinic schedule and patient eligibility reports from EHR
Patient engagement/informed consent
Consent and blood draw order placed in EHR
Blood samples collected by clinical lab
Samples transferred via courier to central clinical processing lab
Samples retrieved and processed by MyCode lab personnel
Long term storage and tracking
Sample analysis
Uses HIT infrastructure
Uses clinical workflow

Figure 1. MyCode® enrollment and biobanking flow chart

Steps from determining patient eligibility to sample analysis are shown. Whenever possible, existing processes and infrastructure are utilized to maximize efficiency. Steps that use existing health information technology (HIT) or clinical work flows are indicated by blue and tan boxes.

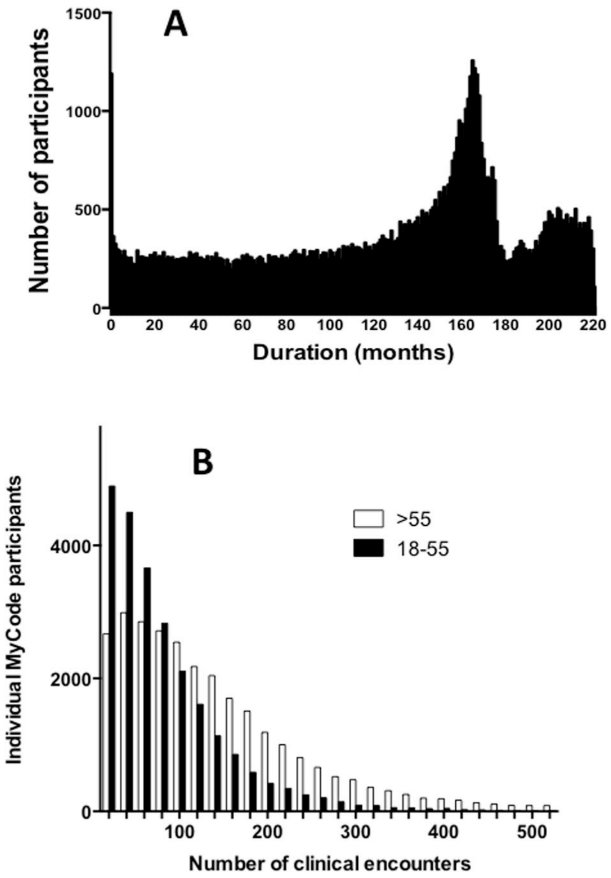

Figure 2. EHR data available for MyCode® participants

Panel A: The duration of available EHR data for 51,893 adult MyCode® participants, defined as the length of time between the most recent clinical encounter and the first encounter recorded for that individual in the GHS EHR; the spike at approximately 160 months corresponds to the completion of EHR implementation in GHS outpatient clinics; *Panel B:* the total number of clinical encounters recorded in the GHS EHR for the same MyCode participants, stratified as participants between 18 and 55 years (current age) or >55 years. The median number of encounters is 120 for age >55 years, and 50 for 18–55 years.

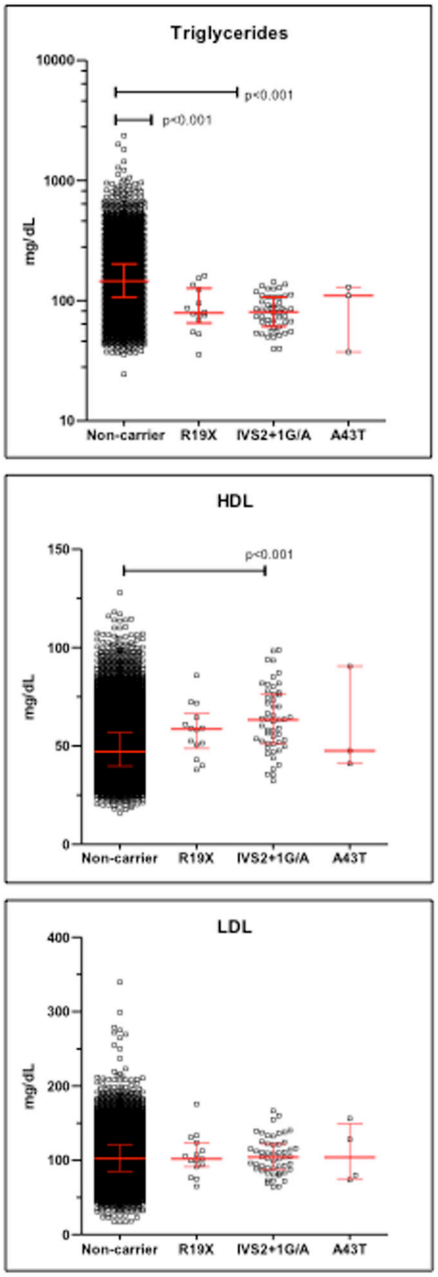

Figure 3. Lipid lab values of carriers and non-carriers of *APOC3* variants

Laboratory values for triglycerides, low density lipoprotein cholesterol (LDL), and high density lipoprotein cholesterol (HDL) were extracted from electronic health record data of 11,499 individuals with both array genotype and blood lipid data. Each point represents the mean value of an individual carrier or non-carrier of the indicated genomic variants. For individuals with no record of a lipid lowering medication a lifetime mean value was calculated; for individuals prescribed a lipid lowering medication, the pre-medication values were averaged. Bars indicate median and inter quartile ranges. *APOC3* variants were determined by array genotyping using the Illumina HumanExome array V1.1. The groups were compared by ANOVA and Dunn's multiple comparison test. Unless indicated, differences among groups were not significant.

Table 1

MyCode Participant Data Recorded in the EHR[a]

Measure	Median value	Range
Duration of EHR data	12.0 years	0 – 221 months
Clinical encounters	60	1 – 1,153
Clinical lab test results	455	1 – 52,041
Vital signs measurements	54	1 – 7,321

[a] 51,893 MyCode participants

Table 2

Genetic Association Analysis Using EHR-Derived Phenotypes[a]

Phenotype Variant (locus)	Number		Reported OR	Risk Allele	OR (95% CI)[b]	*P*[b]
	Cases	Controls				
CVD	1947	4824				
rs10757278 (9p21)			1.3[6,8]	G	1.2 (1.09–1.29)	0.000074
rs1333049 (9p21)			1.3[9,10]	C	1.2 (1.09–1.28)	0.000098
T2DM	2306	679				
rs4506565 (*TCF7L2*)			1.4[7,12] 1.5[11]	T	1.4 (1.18–1.60)	0.000017
rs7903146 (*TCF7L2*)			1.4[12]	T	1.4 (1.20–1.60)	0.0000067
Obesity	534	895				
rs9939609 (*FTO*)			1.3[7,15,17]	A	1.4 (1.18–1.63)	0.000064
rs17782313 (*MC4R*)			1.2[19]	C	1.3 (1.06–1.54)	0.0092

[a] Cases and controls were identified by applying validated phenotype algorithms to EHR data. CVD, cardiovascular disease; T2DM, type 2 diabetes mellitus; OR, odds ratio; CI, confidence interval

[b] Additive genetic model