Leveraging Epidemiologic and Clinical Collections for Genomic Studies of Complex Traits

Dana C. Crawford^a Robert Goodloe^b Eric Farber-Eger^b Jonathan Boston^b
Sarah A. Pendergrass^c Jonathan L. Haines^a Marylyn D. Ritchie^c William S. Bush^a

^aDepartment of Epidemiology and Biostatistics, Institute for Computational Biology, Case Western Reserve University, Cleveland, Ohio, ^bCenter for Human Genetics Research, Vanderbilt University Medical Center, Nashville, Tenn., and ^cCenter for Systems Genomics, Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, Pa., USA

Key Words

Electronic medical records · African Americans · Hispanics · Genetic association studies · EAGLE BioVU · NHANES · Bias

Abstract

Background/Aims: Present-day limited resources demand DNA and phenotyping alternatives to the traditional prospective population-based epidemiologic collections. **Methods:** To accelerate genomic discovery with an emphasis on diverse populations, we – as part of the Epidemiologic Architecture for Genes Linked to Environment (EAGLE) study – accessed all non-European American samples (n = 15,863) available in BioVU, the Vanderbilt University biorepository linked to de-identified electronic medical records, for genomic studies as part of the larger Population Architecture using Genomics and Epidemiology (PAGE) I study. Given previous studies have cautioned against the secondary use of clinically collected data compared with epidemiologically collected data, we present here a characterization of EAGLE BioVU, including the billing and diagnostic (ICD-9) code distributions for adult and pediatric patients as well as comparisons made for select health metrics (body mass index, glucose, HbA1c, HDL-C, LDL-C, and triglycerides) with the

population-based National Health and Nutrition Examination Surveys (NHANES) linked to DNA samples (NHANES III, n = 7,159; NHANES 1999–2002, n = 7,839). **Results:** Overall, the distributions of billing and diagnostic codes suggest this clinical sample is a mixture of healthy and sick patients like that expected for a contemporary American population. **Conclusion:** Little bias is observed among health metrics, suggesting this clinical collection is suitable for genomic studies along with traditional epidemiologic cohorts.

© 2015 S. Karger AG, Basel

Introduction

For several decades, linkage analysis was the workhorse study design for identifying genes and mutations linked to human diseases and traits. The linkage study design is most suited for diseases that display a Mendelian pattern of inheritance within families. These diseases involve single genes and are caused by mutations that are rare in the general population. Examples of successful linkage studies for diseases that fit these criteria include Huntington's disease [1] and cystic fibrosis [2–4]. Linkage study designs were also applied to more common diseases such as type

2 diabetes [5] and hypertension [6], but with few exceptions such as familial forms of breast cancer [7], the study design failed to identify the responsible genes.

It is now well-appreciated that diseases common in an outbred population do not follow Mendelian patterns of inheritance in families and are influenced by common and rare genetic variants across multiple genes and environmental exposures. Given that multiple genetic variants are expected to influence risk of disease, each variant is expected to confer a small genetic effect size. The genetic association study design, now often conducted as case-control genome-wide association studies (GWAS), has been extremely successful in the last few years identifying genetic variants associated with common disease. To date [8], 2,006 GWAS have been performed and cataloged by the National Human Genome Research Institute's GWAS Catalog [9, 10].

The ultimate goals of identifying genetic variants that influence disease risk or human traits include the desire to better understand basic biology, which may lead to the development of clinical advances and personalized medicine [11]. While early GWAS have established the trail from genetic discovery to clinical translation, most studied diseases do not yet have the complete catalogue of susceptibility variants. Indeed, contemporary GWAS such as the recent anthropometry and lipid studies [12–16] require hundreds of thousands of samples to detect common and less common susceptibility alleles. Large sample sizes are also needed to identify complex gene-gene and gene-environment interactions [17]. For gene-environment interactions studies, large cohort studies are preferred to establish a temporal relationship between exposure and outcomes.

The United States currently do not have a national prospective population-based study linked to DNA samples. While some have made the case for the establishment for a national cohort study [18], others have argued this endeavor is too costly and unnecessary given the availability of already established cohorts with DNA samples [19]. Cohorts such as the Women's Health Initiative [20], the Coronary Artery Risk Disease in young Adults (CARDIA) [21], and the Cardiovascular Health Study (CHS) [22] have provided valuable opportunities to contribute towards GWAS; however, most prospective studies available for genomic studies are based on specific outcomes such as common cancers or cardiovascular disease. Also, most of these studies ascertain participants based on specific demographic characteristics [23] such as postmenopausal women in the Women's Health Initiative.

The prospective cohort study is the preferred study design for a variety of genetic epidemiologic experiments

[17], but the need to better represent all Americans across a wide range of outcomes in a cost-effective manner requires alternative models of data acquisition. One such model is the biorepository linked to electronic medical records (EMRs) [24]. With feasibility driven in part by the Health Information Technology for Economic and Clinical Health (HITECH) Act as part of the American Recovery and Reinvestment Act (ARRA), several medical centers have begun the process of collecting DNA samples from patients to conduct genetic association studies for outcomes extracted from EMRs. Biorepositories have several advantages over the traditional epidemiologic cohort, including the relative ease of ascertainment, cost, potential size, and the potential for immediate clinical translation of findings [24]. Biorepositories, however, are not without challenges in both implementation and downstream analysis, as evolving ethical issues for consent [25–27], privacy [28], and return of research results [29, 30] become more apparent.

Biorepositories linked to EMRs also have the potential for bias. It is often assumed and some data suggest [31, 32] that the dense clinical data available for patients in the biorepository represent only the sickest of patients, and these data may not be generalizable to a larger population. This bias is of particular concern for meta-analyses that include a mix of clinical and traditional epidemiologic collections. With this backdrop, we present here a clinical collection of almost 16,000 non-European Americans linked to de-identified EMRs for genomic studies. We describe the demographics of this dataset and explore the potential bias of the available outcome data. Overall, we demonstrate that large biorepositories can mimic population-based prospective epidemiologic studies primed for genomic discovery.

Methods

Study Population

The study population described here was derived from BioVU, the Vanderbilt University Medical Center (VUMC) biorepository linked to de-identified EMRs. The establishment of this opt-out clinical collection, including ethical and legal issues, has been detailed elsewhere [25, 33]. In brief, DNA is extracted from discarded blood drawn for routine clinical care in an outpatient setting. DNA collection on adult and pediatric patients began in 2007 and 2010, respectively.

DNA samples are linked to a de-identified version of the EMR termed the Synthetic Derivative (SD). The VUMC SD contains approximately 20 years of clinical data representing >2.1 million patients of which approximately 1 million have detailed longitudinal data. The SD is generated and refreshed routinely using a cryptographic (one-way) hash function known as the secure hash algorithm 512 [34] applied to the EMRs. This cryptographic hash func-

tion allows the input data (e.g., a patient's medical record number) to be converted to a message digest or digest (e.g., a 128-character research unique identifier). The resulting research unique identifier is unique to the input and links the EMR data through time. In computer science, it is well-understood that one-way hashes are easy to compute for each input but hard to invert. The de-identification process applied to the SD also removes protected health information and randomly shifts dates in the EMR by up to 364 days backwards (but consistently within a patient's record). Data contained within the SD are considered limited datasets as defined by the Health Insurance Portability and Accountability Act (HIPAA) and are in accordance with provisions of Title 45, Code of Federal Regulations, part 46 (45 CFR 46) that define criteria for 'non-human subjects' research [33].

The specific population described here was selected for and genotyped on the Illumina MetaboChip [35, 36] for the Epidemiologic Architecture for Genes Linked to Environment (EAGLE) study, a study site of the larger Population Architecture using Genomics and Epidemiology (PAGE) I study [37]. The PAGE I study, supported by the National Human Genome Research Institute (NHGRI) at the National Institutes of Health (NIH), sought to characterize European-identified genetic associations in diverse populations for various common human diseases and traits [37]. All BioVU DNA samples from non-European Americans as of 2011 ($n = 15,863$) were selected for genotyping and are described and referred to here as 'EAGLE BioVU'. Race/ethnicity in BioVU is third-party or administratively assigned information, and previous work has suggested a high concordance between administratively assigned race/ethnicity and genetic ancestry for European and African Americans [38, 39]. However, a reduced concordance between administratively assigned race/ethnicity and genetic ancestry has been documented for Hispanics and other groups in BioVU [39].

We also accessed data from the epidemiologic National Health and Nutrition Examination Surveys (NHANES). NHANES are cross-sectional surveys conducted by the National Center for Health Statistics (NCHS) at the Centers for Disease Control and Prevention (CDC). Each NHANES cycle collects population-based data on Americans regardless of health status including demographics, health, and lifestyle data. Race/ethnicity is self-identified in NHANES and includes the following categories: non-Hispanic white, non-Hispanic black, Mexican American, and other. DNA was collected during phase 2 of NHANES III (1991–1994; $n = 7,159$) and NHANES 1999–2002 ($n = 7,839$) and is collectively referred to here as 'Genetic NHANES.' Serum high-density lipoprotein cholesterol (HDL-C), triglycerides, and total cholesterol were measured in NHANES using standard enzymatic methods. Low-density lipoprotein cholesterol (LDL-C) was calculated using the Friedewald equation, with missing values assigned for samples with triglyceride levels >400 mg/dl. Plasma glucose was measured using standardized procedures as described by the NHANES III reference manuals [40]. Serum insulin was measured by radioimmunoassay, and glycated hemoglobin (HbA1c) was measured as the percentage of all subfractions of HbA1 using high performance liquid chromatography [40]. Body mass index (BMI) was calculated from height and weight measured in the Mobile Examination Center by CDC medical personnel.

All procedures were approved by the CDC Ethics Review Board, and written informed consent was obtained from all participants. Because no identifying information was accessed by the

investigators, Vanderbilt University's Institutional Review Board determined that this study met the criteria of 'non-human subjects'. Data presented here represent epidemiologic data from Genetic NHANES III and 1999–2002. Analyses were conducted using SAS v9.2 (SAS Institute, Cary, N.C., USA).

Electronic Phenotyping

EMRs contain structured codes [International Classification of Diseases or ICD-9 (billing) codes and current procedural terminology (procedure) codes]; clinical laboratory test results, semi-structured, and unstructured (clinical notes) data, all of which can be used for electronic phenotyping. The data types currently available in the SD that can be accessed for electronic phenotyping include narratives (such as clinical notes, discharge summaries, history and physicals, problem lists, surgical reports, progress notes, and letters), ICD-9 codes, current procedural terminology codes, forms (intake, assessment), reports (pathology, electrocardiograms, and echocardiograms), clinical communications, laboratory values and vital signs, medication orders, TraceMaster (electrocardiograms), and the tumor registry [41].

Electronic phenotyping has been previously described for BMI in EAGLE BioVU [42]. To extract lipid and type 2 diabetes traits from EAGLE BioVU EMRs, laboratory measurements were queried for glucose, HbA1c, HDL-C, insulin, LDL-C, total cholesterol, and triglycerides. Records were also queried for calculated LDL-C. Prescription medication is available in the SD through MedEx [43], an algorithm that extracts medications and their signature mentions from free-text entries available in the EMRs.

For each adult patient (>18 years), median values were calculated for: (a) measurements taken when no medications are prescribed ('pre-medication' values) and (b) measurements taken at first mention of medication and after the mention of medication ('post-medication' values). We used the following medication classes and lists to identify lipid measurements determined at the time or after patients were prescribed lipid-lowering medications:

- Statins [also known as HMG CoA reductase inhibitors, atorvastatin (Lipitor®), fluvastatin (Lescol®), lovastatin (Mevacor®, Altoprev™), pravastatin (Pravachol®), rosuvastatin calcium (Crestor®), simvastatin (Zocor®), lovastatin + niacin (Advicor®), atorvastatin + amlodipine (Caduet®), and simvastatin + ezetimibe (Vytorin™)].
- Selective cholesterol absorption inhibitors [ezetimibe (Zetia®)].
- Resins [cholestyramine (Questran®, Questran® Light, Prevalite®, Locholest®, and Locholest® Light), colestipol (Colestid®), and colesevelam Hcl (WelChol®)].
- Fibrates [gemfibrozil (Lopid®), fenofibrate (Antara®, Lofibra®, Tricor®, and Triglide™), and clofibrate (Atromid-S)].
- Niacin.

Results

Demographics

VUMC is located in Davidson County of Nashville, Tenn., USA. VUMC serves Middle Tennessee (primarily Davidson, Montgomery, Rutherford, and Williamson counties), southern Kentucky, and northern Alabama.

Table 1. Population characteristics for Davidson County, BioVU, and EAGLE BioVU

	Davidson County (n = 626,684)	BioVU (n = 162,716)	EAGLE BioVU (n = 15,863)
Females	51.55%	51.93%	65.35%
Adults			
18–64 years	68.06%	57.66%	71.87%
≥65 years	10.23%	24.83%	15.61%
European Americans	60.48%	81.07%	–
African Americans	28.43%	8.65%	73.06%
Hispanics	10.04%	1.32%	10.87%
Asians	3.10%	0.83%	7.12%

Davidson County statistics are based on the 2010 census. BioVU statistics are based on a query performed in April 2013. EAGLE BioVU patient DNA samples were selected for genotyping in 2011. Race/ethnicity in the Davidson County 2010 census was self-identified and administratively assigned in BioVU and EAGLE BioVU.

The 2010 US Census counted 1,244,799 residents across the four Middle Tennessee counties with half (626,684) of the residents in Davidson County. As of April 8, 2013, BioVU contained 162,716 DNA samples. With the assumption that the majority of BioVU samples were obtained from Davidson County residents, BioVU represents approximately up to one quarter of the county (26%). Despite representing many of the actual residents that make up Davidson County, BioVU does not reflect the overall population demographics of the county. That is, BioVU underrepresents minorities and overrepresents older adults (≥65 years of age) compared with Davidson county (table 1). The overrepresentation of older adults in BioVU is consistent with the fact that BioVU draws samples from a clinical population, as an older age group tends to access health care more often than younger adults. Also, according to the 2011 National Health Interview Survey conducted by the CDC [44], Hispanics were least likely to visit a health care provider compared with European Americans and African Americans, and this national trend is reflected somewhat in BioVU. The lower racial/ethnic diversity observed in BioVU compared with Davidson County may also be explained by the fact that VUMC serves other less diverse Middle Tennessee counties as well as southern Kentucky and northern Alabama. Given that patient location is de-identified in the SD, it is difficult to distinguish amongst these possibilities.

Overall, non-European American patients tend to be younger female adults, and these demographics are similar to non-European Americans ascertained by CDC for

NHANES III (1991–1994), 1999–2000, and 2001–2002 (table 2). The majority of DNA samples in EAGLE BioVU (genotyped on the MetaboChip) are from African Americans (73%) followed by Hispanics (~11%). In the 2010 US Census, more than half of the Hispanic population in Davidson County was Mexican (61%) or ‘other’ Hispanic (31%). In comparison, data mining of EAGLE BioVU clinical notes suggests that 55.9% of all DNA samples are from Mexican patients followed by patients from Honduras (10.3%), Guatemala (9.6%), Cuba (3.7%), Peru (2.7%), Ecuador (2.5%), and Nicaragua (2.5%) [45].

Distribution of Diagnostic and Billing Codes

An important concern associated with the use of DNA samples and data collected in the clinic is whether or not the patients and phenotypes are more extreme compared with participants and health data collected from the general population, as this could pose challenges in harmonizing phenotypic data with other epidemiologic collections. Indeed, groups of DNA samples and clinical data collected from specific medical specialties or clinics will be biased towards extreme phenotypes. For example, if EAGLE BioVU were limited to DNA samples from patients visiting the bariatric clinic (Vanderbilt Center for Surgical Weight Loss), the median and range of BMI values would be higher and not representative of a general population in Tennessee and the United States.

EAGLE BioVU did not select patients and DNA samples that represent specialty clinics or specific types of clinic visits. DNA samples collected for BioVU as a whole are in fact collected from a variety of outpatient clinic visits and are not collected during inpatient visits (although the EMRs will contain health data from inpatient visits). The EAGLE BioVU clinical data represent on average ($\pm$ standard deviation) 81.8 ($\pm$ 107.8) clinic visits per patient (range: 1–1,456 clinic visits/patient). The average patient has 147.3 ($\pm$ 230.4) ICD-9 codes (range: 1–3,617).

Characterization of the extensive EAGLE BioVU data suggest that the health status for the majority of the patient population with DNA samples is similar to that reported for the general population of the United States. For example, the three most commonly mentioned ICD-9 codes among African-American adults in EAGLE BioVU include hypertension (401.9 and 401.1), type 2 diabetes (250), and end-stage renal disease (585.6). These conditions are known to be prevalent in African Americans, and their frequency in EAGLE BioVU is consistent with data from NHANES and other national surveys, where almost 40 and 20% of non-Hispanic black participants are hypertensive [46] and/or have type 2 diabetes [47], re-

Table 2. Study population characteristics for African Americans and Hispanics from EAGLE BioVU and Genetic NHANES

	African Americans		Hispanics/Mexican Americans	
	EAGLE BioVU (n = 11,521)	Genetic NHANES (n = 3,458)	EAGLE BioVU (n = 1,714)	Genetic NHANES (n = 3,950)
Females	65.7%	55.5%	63.6%	51.1%
Median age $\pm$ SD, years	42.30 $\pm$ 20.45	38.00 $\pm$ 18.75	26.08 $\pm$ 19.28	38.00 $\pm$ 18.74
Median BMI $\pm$ SD	27.17 $\pm$ 7.20	27.16 $\pm$ 7.10	25.24 $\pm$ 7.11	27.10 $\pm$ 5.67
Median HDL-C level $\pm$ SD, mg/dl	50 $\pm$ 16.99	51 $\pm$ 16.5	46 $\pm$ 15.4	47 $\pm$ 13.44
Median LDL-C level $\pm$ SD, mg/dl	99 $\pm$ 35.34	116 $\pm$ 38.46	101 $\pm$ 35.64	118 $\pm$ 33.66
Median triglyceride level $\pm$ SD, mg/dl	98 $\pm$ 69.14	87 $\pm$ 75.04	121 $\pm$ 97.8	125 $\pm$ 106.8

BMI and lipid levels were extracted from the laboratory records in the EMRs for adults >18 years of age as described in Methods. For patients with multiple laboratory records, the median was calculated, and an overall median of the values is presented here for the patient population stratified by administratively assigned race/ethnicity.

Age was extracted for each patient as age of the first clinic visit. Sample size shown represents maximum sample size regardless of age. Genetic NHANES represents data from participants ascertained between 1991–1994 and 1999–2002, who consented to providing a biologic sample for genetic studies. Data presented here for Genetic NHANES are for adults >18 years of age.

spectively, both of which are major known causes of end-stage renal disease [48, 49]. Rounding out the most frequent billing codes in African-American adults are sequestrectomy (77), long-term current use of other medications (v58.69), other malaise/fatigue (780.79), chest pain (786.5), heart failure (428), and kidney transplant (v42.0). In contrast, the three most commonly mentioned ICD-9 codes among Hispanic adults include sequestrectomy (77), supervision of other normal pregnancy (v22.1, typically coded for a second or subsequent pregnancy associated with routine outpatient prenatal visits when no complications are present), and end-stage renal disease (585.6). Among Asian adults, the three most commonly mentioned ICD-9 codes include supervision of other normal pregnancy (v22.1), hypertension (401.9, 401.1), and long-term current use of other medications (v58.69). The high rate of end-stage renal disease in this Hispanic population is consistent with the known high prevalence of type 2 diabetes in this population (250; the sixth most common code in this sample of Hispanic adults). Other common billing codes observed for Hispanic adults include outcome of delivery; single liveborn (v27.0); unknown, default (000.00); supervision of normal first pregnancy (v22.0), and long-term current use of other medications (v58.69). For Asian adults, the other common billing codes include diabetes mellitus (250), routine general medical examination at a health care facility (v70.0), kidney transplant (v42.0), sequestrectomy (77), supervision of normal first pregnancy (v22.0), and mixed hyperlipidemia. The common codes among Hispanic and

Asian adults, such as the pregnancy codes, suggest that this sample of Hispanics and Asians is relatively speaking as healthy as individuals ascertained in a general population setting.

As expected, the distribution of billing codes differs in the pediatric population of EAGLE BioVU compared with the adult population. The code for routine infant or child health check (v20.2) is the most frequent code mentioned for both African-American and Hispanic pediatric populations, and it is the second most frequent code mentioned for Asian pediatric populations in EAGLE BioVU. Of note is the code for acute lymphoid leukemia, in remission (204.01). Code 204.01 is the second most common code among Hispanic children and the sixth most common code for African-American children. Thus, the children in EAGLE BioVU represent a complex mix of healthy and sick children, not unexpected, given that samples come from both outpatient settings as well as the Monroe Carell Jr. Children's Hospital at Vanderbilt [25]. Other frequently mentioned codes among pediatric patients in EAGLE BioVU include fever and other physiologic disturbances of temperature regulation (780.6), as might be expected for this age range.

The ten most frequent codes mentioned for African-American children included routine infant or child health check (v20.2); Hb-SS disease without crisis (282.61); fever and other physiologic disturbances of temperature regulation (780.6); asthma, unspecified (493.9); incision, excision, and division of other bones (77); acute lymphoid leukemia, in remission (204.01);

unknown, default (000.000); other convulsions (780.39); acute respiratory infections of unspecified site (465.9), and ADHD (314.01). For Hispanic children, the most common billing codes included routine infant or child health check (v20.2); acute lymphoid leukemia, in remission (204.01); fever and other physiologic disturbances of temperature regulation (780.6); delayed milestones (783.42); unknown, default (000.00); anemia, unspecified (285.9); incision, excision, and division of other bones (77); acute upper respiratory infections of unspecified site (465.9), and other convulsions (780.39). Finally, for Asian children, the ten most common billing codes include acute lymphoid leukemia, in remission (204.01); routine infant or child health check (v20.2); fever and other physiologic disturbances of temperature regulation (780.6); esophagomyotomy (v42.7); unknown, default (000.000); lymphoid leukemia (204); encounter for antineoplastic chemotherapy (v58.11); anemia, unspecified (285.9); incision, excision, and division of other bones (77), and feeding difficulties and mismanagement (783.3).

Other Health Metrics

To further assess the potential for bias, we compared the commonly measured health metrics BMI and lipid traits (HDL-C, LDL-C, and triglycerides) between the nationally representative NHANES to EAGLE BioVU (table 2). On average, African-Americans and Hispanics from EAGLE BioVU as well as non-Hispanic blacks and Mexican Americans from Genetic NHANES are overweight (median BMI ≥ 25 ; table 2). When restricted to adults with lipid traits available, African-Americans from EAGLE BioVU on average are obese (median BMI >30) compared with non-Hispanic blacks from Genetic NHANES (median BMI 27.14), whereas Hispanics from EAGLE BioVU have a comparable median BMI (27.99) compared with Mexican Americans from Genetic NHANES (27.10). The high rate of overweight and obesity among EAGLE BioVU African-Americans presumably from the Nashville and surrounding area is consistent with the high rates of obesity observed among African-Americans (particularly women) from the southeastern United States compared with other groups and states [50, 51]. Genetic NHANES is cross-sectional and contains data from 1991 to 1994 (NHANES III), when obesity rates were lower overall compared with more contemporary obesity trends [52, 53].

Among lipid traits, HDL-C is similar between the clinical and epidemiologic collections for both African Americans and Hispanics. Triglyceride levels are higher among

African Americans from EAGLE BioVU (median = 98 mg/dl) compared with non-Hispanic blacks from Genetic NHANES (median = 87 mg/dl) but similar between Hispanics (121 mg/dl) and Mexican Americans (125 mg/dl) from the two collections. Of the three lipid traits, median LDL-C levels are lower in the clinical collection (medians = 99 and 101 mg/dl) compared with the epidemiologic collection (medians = 116 and 118 mg/dl) for both African Americans and Hispanics, respectively. When stratified by usage of lipid-lowering medication, overall EAGLE BioVU median LDL-C pre-medication levels (105 ± 38.9 mg/dl) are more similar to Genetic NHANES compared with median post-medication levels (96 ± 38.6 mg/dl).

To further compare EAGLE BioVU to NHANES, we extracted laboratory measures used to monitor the development and control of type 2 diabetes: glucose, insulin, and HbA1c. Overall, both mean and median glucose levels are higher among both African-American and Hispanic men and women compared with NHANES (table 3). HbA1c levels are also on average higher in both sexes of EAGLE BioVU compared with NHANES (table 3). Too few EAGLE BioVU samples had insulin levels to make meaningful comparisons with NHANES.

Discussion

The data presented here suggest that large biorepositories linked to clinical data can mimic traditional epidemiologic collections. More specifically, we demonstrated that EAGLE BioVU, a clinical collection of almost 16,000 non-European descent samples linked to de-identified EMRs is similar to Genetic NHANES, a population-based survey of Americans ascertained regardless of health status, for several health metrics such as BMI and lipid levels. We also described the distributions of billing and diagnostic codes for this clinical collection of adult and pediatric patients. Overall, these data suggest that EAGLE BioVU represents a mixture of healthy and diseased patients much like that expected for the contemporary American population.

Much excitement has been generated towards the use of biorepositories given their potential for cost-effectiveness and large sample sizes compared with the ascertainment and maintenance of traditional epidemiologic collections [24]. However, secondary use of existing data such as that collected primarily for clinical use is not without problems and potential biases. Indeed, Weiskopf et al. [32] have suggested that EMRs are non-randomly com-

Table 3. Study population characteristics for African Americans and Hispanics from EAGLE BioVU and Genetic NHANES by sex

	African Americans				Hispanics/Mexican Americans			
	EAGLE BioVU		Genetic NHANES		EAGLE BioVU		Genetic NHANES	
	males (n _{max} = 3,952)	females (n _{max} = 7,569)	males (n _{max} = 1,539)	females (n _{max} = 1,919)	males (n _{max} = 624)	females (n _{max} = 1,090)	males (n _{max} = 1,930)	females (n _{max} = 2,020)
<i>BMI</i>								
Samples, n	2,680	5,497	1,339	1,688	259	595	1,758	1,832
Mean	27.73	29.31	27.29	30.12	27.97	28.48	27.58	28.71
Median ± SD	27.31±5.94	28.88±6.8	26.60±5.69	29.2±7.66	27.45±5.57	27.44±6.15	27.10±4.85	28.00±6.08
<i>Glucose, mg/dl</i>								
Samples, n	3,325	5,672	980	1,304	337	525	1,286	1,327
Mean	110.00	104.63	101.83	102.17	108.82	102.28	107.97	103.77
Median ± SD	100.00±34.26	95.00±34.09	94.00±35.69	90.80±47.20	102.00±28.02	95±28.86	97.9±42.84	92.50±42.07
<i>HbA1c, %</i>								
Samples, n	1,006	1,915	1,079	1,398	61	153	1,319	1,340
Mean	6.93	6.72	5.70	5.69	6.80	6.20	5.70	5.62
Median ± SD	6.40±1.83	6.25±1.68	5.50±1.11	5.40±1.36	6.20±1.8	5.70±1.58	5.40±1.23	5.30±1.25
<i>Insulin, μU/ml</i>								
Samples, n	32	245	980	1,303	1	12	1,287	1,328
Mean	28.84	31.24	14.16	16.04	11.30	33.00	14.56	15.83
Median ± SD	15.55±35.12	16.40±84.46	9.33±17.38	11.61±18.23	11.30	22.70±42.23	10.63±19.00	11.39±27.26

BMI, glucose, HbA1c, and insulin levels were extracted from the laboratory records in the EMRs for adults >18 years of age as described in Methods. For patients with multiple laboratory records, the median was calculated, and an overall median of the values is presented here for the patient population stratified by administratively assigned race/ethnicity.

NHANES represents data from participants ascertained between 1991–1994 and 1999–2002, who consented to providing a biologic sample for genetic studies. Data presented here for Genetic NHANES are for adults >18 years of age.

plete and that the densest data are associated with the sickest patients. Furthermore, Rusanov et al. [31] have suggested that algorithms requiring sufficient laboratory results and medication orders in a patient's EMR for study subject selection result in a biased or sicker collection of patients compared with patients not selected for study.

Indeed, EMRs are fraught with limitations when applied in a research setting, including data quality, missing data, and bias [24]. EAGLE BioVU does not seem to display the biased qualities observed in other datasets. In addition to the data presented here, PAGE I studies suggest that genetic effect size estimates generated from the clinical resource BioVU are not different to those calculated from epidemiologic data [54, 55]. Formal tests of heterogeneity have been performed across PAGE I study sites for meta-analyses of various cancers [56–60], and, to date, only one PAGE I study of multiple primary cancers has demonstrated significant heterogeneity or bias [61] likely resulting from survival bias for incident and prevalent cancer data in BioVU (where DNA samples are

available only for patients who survived between 2007 and the present). Recently, we have also compared cardiac structure and function data extracted from African-American patients in EAGLE BioVU with epidemiologic data collected by Jackson Heart Study and the Atherosclerosis Risk in Communities Study [62]. These data suggest that the data extracted from the clinical EAGLE BioVU for African Americans is comparable in both sample size and distributions for these measures of cardiac structure and function [62]. Similar comparable results were observed between EAGLE BioVU-extracted ages at menopause and menarche compared with NHANES [63].

The potential for bias lurking in biorepositories linked to EMRs extends beyond genomic discovery. As research moves from discovery to clinical translation [64], clinical data is being explored for other research applications such as comparative effectiveness research with an emphasis on genomic testing. The lack of such data to support clinical applications has been extensively documented (for example Malinowski and Naylor [65]); neverthe-

less, genomic testing in the clinic is on the rise [66, 67], prompting contemporary cost-effective studies of clinical utility and validity given the expense and time required for gold standard randomized clinical trials and well-designed longitudinal cohort studies [68]. Biorepositories linked to extensive longitudinal EMRs may provide the much needed data to bridge the evidence gap for these genomic tests in a real-life clinical setting. However, much like its use in genomic research, the use of EMRs for comparative effectiveness research is limited to the data quality, completeness, and bias inherent in clinical and billing practices [69] and should be used with the appropriate caution.

The present study has both limitations and strengths. This comparison between EAGLE BioVU and Genetic NHANES is limited to select demographic variables (age, sex, and race/ethnicity) and health metrics (BMI, lipid traits, and type 2 diabetes-associated traits). The selection of variables for comparison was driven, in part, by the overlap between variables available in NHANES and high-quality outcomes and risk factors extracted from EAGLE BioVU. NHANES contains hundreds of laboratory and environmental measures (e.g., inflammation markers, C-reactive protein [70], and fibrinogen [71, 72]) not routinely ordered in a clinical setting. Conversely, EAGLE BioVU contains hundreds of clinically documented (as opposed to self-reported) outcomes in greater numbers given the older average age compared with NHANES (e.g., myocardial infarction). For overlapping variables with sufficient counts, it is possible that more extensive comparisons will uncover major differences between the clinical and epidemiologic collections. Indeed, we observed even at this cursory level higher HbA1c and glucose levels in EAGLE BioVU compared with Genetic NHANES. This observed difference in type 2 diabetes surveillance laboratory records may be due to a complex mix of age and other risk factor differences between the two data collections. Using the electronic MEDical Records and GENomics (eMERGE) Network type 2 diabetes algorithm [73], we identified 1,672 and 101 potential cases of type 2 diabetes in EAGLE BioVU representing 14.5 and 6% of African Americans and Hispanics, respectively. In contrast, only 7.4% of non-Hispanic blacks in Genetic NHANES have type 2 diabetes, while 8.5% of Mexican Americans in Genetic NHANES have type 2 diabetes [74]. The higher proportion of type 2 diabetes cases among African Americans in EAGLE BioVU can explain, in part, the higher HbA1c levels compared with Genetic NHANES; however, the reciprocal relationship

between the clinical and epidemiologic studies for Hispanics/Mexican Americans is more complex. Another limitation of the present study is the fact that the analyses performed in Genetics NHANES were not weighted for survey design, consistent with our previous analyses and meta-analyses with PAGE I studies (e.g., Dumitrescu et al. [75]).

Despite these limitations, there are several strengths. EAGLE BioVU is a relatively large collection of non-European Americans and promises to add towards genomic discovery in relatively understudied minority populations in the United States. EAGLE BioVU patients were selected for study based on a single variable: race/ethnicity. This broad selection process coupled with the lengthy EMR history (dating more than 20 years for some patients in BioVU) most likely contributed to the unbiased features of this dataset compared with previously published datasets extracted from the EMRs [31, 32, 76].

Conclusions

Alternative models of data collection for genomic studies have highlighted the potential uses of EMRs in genomic discovery and genomic medicine settings. Previous studies revealing problems related to data quality, completeness, and bias, however, have tempered the excitement surrounding the use of EMRs in replacing or augmenting traditional epidemiologic collections. We described here a diverse clinical cohort for genomic studies and made comparisons with a nationally representative cross-sectional study. Collectively, these data suggest that EAGLE BioVU is essentially a random sample of the general population, making this resource more than comparable to traditional epidemiologic collections at least in the context of genetic association studies.

Acknowledgements

This work was supported in part by NIH grant U01 HG004798 and its ARRA supplements. The dataset(s) used for the analyses described were obtained from VUMC's BioVU, which is supported by institutional funding and by the National Center for Research Resources, grant UL1 RR024975-01 (which is now at the National Center for Advancing Translational Sciences, grant 2 UL1 TR000445-06). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or the CDC. The Vanderbilt University Center for Human Genetics Research, Computational Genomics Core provided computational and/or analytical support for this work.

References

- Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, Tanzi RE, Watkins PC, Ottina K, Wallace MR, Sakaguchi AY, Young AB, Shoulson I, Bonilla E, Martin JB: A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* 1983;306:234–238.
- Knowlton RG, Cohen-Haguenaer O, Van Cong N, Frezal J, Brown VA, Barker D, Braman JC, Schumm JW, Tsui L-C, Buchwald M, Donis-Keller H: A polymorphic DNA marker linked to cystic fibrosis is located on chromosome 7. *Nature* 1985;318:380–382.
- Tsui L-C, Buchwald M, Barker D, Braman JC, Knowlton R, Schumm JW, Eiberg H, Mohr J, Kennedy D, Plavsic N, Zsiga M, Markiewicz D, Akots G, Brown V, Helms C, Gravius T, Parker C, Rediker K, Donis-Keller H: Cystic fibrosis locus defined by a genetically linked polymorphic DNA marker. *Science* 1985;230:1054–1057.
- Wainwright BJ, Scambler PJ, Schmidtke J, Watson EA, Law H-Y, Farrall M, Cooke HJ, Eiberg H, Williamson R: Localization of cystic fibrosis locus to human chromosome 7cen-q22. *Nature* 1985;318:384–385.
- Altmuller J, Palmer LJ, Fischer G, Scherb H, Wjst M: Genomewide scans of complex human diseases: true linkage is hard to find. *Am J Hum Genet* 2001;69:936–950.
- Binder A: A review of the genetics of essential hypertension. *Curr Opin Cardiol* 2007;22:176–184.
- Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, Huey B, King MC: Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* 1990;250:1684–1690.
- Hindorf LA, MacArthur J, Morales J, Junkins HA, Hall PN, Klemm AK, Manolio TA: A Catalog of Published Genome-Wide Association Studies, 2014. www.genome.gov/gwastudies (accessed October 24, 2014).
- Hindorf LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA: Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci* 2009;106:9362–9367.
- Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorf L, Parkinson H: The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 2014;42:D1001–D1006.
- McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA, Hirschhorn JN: Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* 2008;9:356–369.
- Global Lipids Genetics Consortium: Discovery and refinement of loci associated with lipid levels. *Nat Genet* 2013;45:1274–1283.
- Speliotes EK, Willer CJ, Berndt SI, et al: Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet* 2010;42:937–948.
- Wood AR, Esko T, Yang J, et al: Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* 2014;46:1173–1186.
- Teslovich TM, Musunuru K, Smith AV, et al: Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* 2010;466:707–713.
- Berndt SI, Gustafsson S, Magi R, et al: Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nat Genet* 2013;45:501–512.
- Manolio TA, Bailey-Wilson JE, Collins FS: Genes, environment and the value of prospective cohort studies. *Nat Rev Genet* 2006;7:812–820.
- Collins FS: The case for a US prospective cohort study of genes and environment. *Nature* 2004;429:475–477.
- Willett WC, Blot WJ, Colditz GA, Folsom AR, Henderson BE, Stampfer MJ: Merging and emerging cohorts: not worth the wait. *Nature* 2007;445:257–258.
- The Women's Health Initiative Study Group: Design of the Women's Health Initiative clinical trial and observational study. *Control Clin Trials* 1998;19:61–109.
- Friedman GD, Cutter GR, Donahue RP, Hughes GH, Hulley SB, Jacobs DR Jr, Liu K, Savage PJ: CARDIA: study design, recruitment, and some characteristics of the examined subjects. *J Clin Epidemiol* 1988;41:1105–1116.
- Fried LP, Borhani NO, Enright P, et al: The Cardiovascular Health Study: design and rationale. *Ann Epidemiol* 1991;1:263–276.
- Collins FS, Manolio TA: Merging and emerging cohorts: necessary but not sufficient. *Nature* 2007;445:259.
- Kohane IS: Using electronic health records to drive discovery in disease genomics. *Nat Rev Genet* 2011;12:417–428.
- McGregor TL, Van Driest SL, Brothers KB, Bowton EA, Muglia LJ, Roden DM: Inclusion of pediatric samples in an opt-out, de-identified linking DNA to de-identified medical records: pediatric BioVU. *Clin Pharmacol Ther* 2013;93:204–211.
- Master Z, Nelson E, Murdoch B, Caulfield T: Biobanks, consent and claims of consensus. *Nat Meth* 2012;9:885–888.
- Pulley J, Clayton E, Bernard GR, Roden DM, Masys DR: Principles of human subjects protections applied in an opt-out, de-identified biobank. *Clin Transl Sci* 2010;3:42–48.
- Malin B, Loukides G, Benitez K, Clayton E: Identifiability in biobanks: models, measures, and mitigation strategies. *Hum Genet* 2011;130:383–392.
- Kocarnik JM, Fullerton SM: Returning pleiotropic results from genetic testing to patients and research participants. *JAMA* 2014;311:795–796.
- Jarvik GP, Amendola LM, Berg JS, et al: Return of genomic results to research participants: the floor, the ceiling, and the choices in between. *Am J Hum Genet* 2014;94:818–826.
- Rusanov A, Weiskopf N, Wang S, Weng C: Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research. *BMC Med Inform Decis Mak* 2014;14:51.
- Weiskopf NG, Rusanov A, Weng C: Sick patients have more data: the non-random completeness of electronic health records. *AMIA Annu Symp Proc* 2013;2013:1472–1477.
- Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, Balser JR, Masys DR: Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin Pharmacol Ther* 2008;84:362–369.
- Dang Q: Recommendation for Applications Using Approved Hash Algorithms. Washington, US Department of Commerce, National Institute of Standards and Technology, 2012.
- Buyske S, Wu Y, Carty CL, et al: Evaluation of the metabochip genotyping array in African Americans and implications for fine mapping of GWAS-identified loci: the PAGE Study. *PLoS One* 2012;7:e35651.
- Voight BF, Kang HM, Ding J, et al: The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet* 2012;8:e1002793.
- Matise TC, Ambite JL, Buyske S, Carlson CS, Cole SA, Crawford DC, Haiman CA, Heiss G, Kooperberg C, Le Marchand L, Manolio TA, North KE, Peters U, Ritchie MD, Hindorf LA, Haines JL: The Next PAGE in understanding complex traits: design for the analysis of Population Architecture Using Genetics and Epidemiology (PAGE) Study. *Am J Epidemiol* 2011;174:849–859.
- Dumitrescu L, Ritchie MD, Brown-Gentry K, Pulley JM, Basford M, Denny JC, Oksenberg JR, Roden DM, Haines JL, Crawford DC: Assessing the accuracy of observer-reported ancestry in a biorepository linked to electronic medical records. *Genet Med* 2010;12:648–650.
- Hall JB, Dumitrescu L, Dilks HH, Crawford DC, Bush WS: Accuracy of administratively-assigned ancestry for diverse populations in an electronic medical record-linked biobank. *PLoS One* 2014;9:e99161.
- Centers for Disease Control and Prevention (CDC): Plan and operation of the Third National Health and Nutrition Examination Survey, 1988–1994. Series 1: programs and collection procedures. *Vital Health Stat* 1994;1:1–407.

- 41 Bush WS, Boston J, Pendergrass SA, Dumitrescu L, Goodloe R, Brown-Gentry K, Wilson S, McClellan B, Torstenson E, Basford MA, Spencer KL, Ritchie MD, Crawford DC: Enabling high-throughput genotype-phenotype associations in the Epidemiology Architecture for Genes Linked to Environment (EAGLE) project as part of the Population Architecture using Genomics and Epidemiology (PAGE) study. *Pac Symp Biocomput* 2013;18:373–384.
- 42 Goodloe R, Farber-Eger E, Boston J, Crawford DC, Bush WS: Reducing clinical noise for body mass index measures due to unit and transcription errors in the electronic medical record. Submitted.
- 43 Xu H, Jiang M, Oetjens M, Bowton EA, Ramirez AH, Jeff JM, Basford MA, Pulley JM, Cowan JD, Wang X, Ritchie MD, Masys DR, Roden DM, Crawford DC, Denny JC: Facilitating pharmacogenetic studies using electronic health records and natural-language processing: a case study of warfarin. *J Am Med Inform Assoc* 2011;18:387–391.
- 44 Schiller JS, Lucas JW, Peregoy JA: Summary health statistics for U.S. adults: National Health Interview Survey, 2011. National Center for Health Statistics. *Vital Health Stat* 10 2012;256:1–218.
- 45 Farber-Eger E, Goodloe R, Boston J, Bush WS, Crawford DC: Extracting country-of-origin from electronic medical records for gene-environment studies as part of the Epidemiologic Architecture for Genes Linked to Environment (EAGLE) study. Submitted.
- 46 Haves DK, Denny CH, Keenan NL, Croft JB, Greenlund KJ: Health-related quality of life and hypertension status, awareness, treatment, and control: National Health and Nutrition Examination Survey, 2001–2004. *J Hypertens* 2008;26:641–647.
- 47 Centers for Disease Control and Prevention: National Diabetes Fact Sheets: National Estimates and General Information on Diabetes and Prediabetes in the United States, 2011. Atlanta, Department of Health and Human Services, 2011.
- 48 Crews DC, Plantinga LC, Miller ER, Saran R, Hedgeman E, Saydah SH, Williams DE, Powe NR, for the Centers for Disease Control and Prevention Chronic Kidney Disease Surveillance Team: Prevalence of chronic kidney disease in persons with undiagnosed or prehypertension in the United States. *Hypertension* 2010;55:1102–1109.
- 49 de Boer IH, Rue TC, Hall YN, Heagerty PJ, Weiss NS, Himmelfarb J: Temporal trends in the prevalence of diabetic kidney disease in the United States. *JAMA* 2011;305:2532–2539.
- 50 Centers for Disease Control and Prevention: Behavioral Risk Factor Surveillance System Survey Data. Atlanta, US Department of Health and Human Services, 2013.
- 51 Ogden CL, Carroll MD, Kit BK, Flegal KM: Prevalence of childhood and adult obesity in the United States, 2011–2012. *JAMA* 2014;311:806–814.
- 52 Flegal KM, Carroll MD, Ogden CL, Curtin LR: Prevalence and trends in obesity among us adults, 1999–2008. *JAMA* 2010;303:235–241.
- 53 Flegal KM, Carroll MD, Kuczmarski RJ, Johnson CL: Overweight and obesity in the United States: prevalence and trends, 1960–1994. *Int J Obes Relat Metab Disord* 1998;22:39–47.
- 54 Bush WS, Dumitrescu L, Crawford DC: Assessment of bias in an EMR-derived cohort. American Society of Human Genetics, San Francisco, October 6–10, 2012.
- 55 Ritchie MD, Denny JC, Crawford DC, Ramirez AH, Weiner JB, Pulley JM, Basford MA, Brown-Gentry K, Balser JR, Masys DR, Haines JL, Roden DM: Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *Am J Hum Genet* 2010;86:560–572.
- 56 Lim U, Kocarnik JM, Bush WS, et al: Pleiotropy of cancer susceptibility variants on the risk of non-Hodgkin lymphoma: the PAGE consortium. *PLoS One* 2014;9:e89791.
- 57 Setiawan VW, Schumacher F, Prescott J, et al: Cross-cancer pleiotropic analysis of endometrial cancer: PAGE and E2C2 consortia. *Carcinogenesis* 2014;35:2068–2073.
- 58 Kocarnik JM, Park SL, Han J, et al: Replication of associations between GWAS SNPs and melanoma risk in the Population Architecture Using Genomics and Epidemiology (PAGE) Study. *J Invest Dermatol* 2014;134:2049–2052.
- 59 Park SL, Fesinmeyer MD, Timofeeva M, et al: Pleiotropic associations of risk variants identified for other cancers with lung cancer risk: the PAGE and TRICL Consortia. *J Natl Cancer Inst* 2014;106.
- 60 Cheng I, Kocarnik JM, Dumitrescu L, et al: Pleiotropic effects of genetic risk variants for other cancers on colorectal cancer risk: PAGE, GECCO and CCFR consortia. *Gut* 2014;63:800–807.
- 61 Park SL, Caberto CP, Lin Y, et al: Association of cancer susceptibility variants with risk of multiple primary cancers: the Population Architecture using Genomics and Epidemiology Study. *Cancer Epidemiol Biomarkers Prev* 2014;23:2568–2578.
- 62 Wells Q, Farber-Eger E, Crawford D: Extraction of echocardiographic data from the electronic medical record is a rapid and efficient method for study of cardiac structure and function. *J Clin Bioinforma* 2014;4:12.
- 63 Malinowski J, Farber-Eger E, Crawford DC: Development of a data-mining algorithm to identify ages at reproductive milestones in electronic medical records. *Pac Symp Biocomput* 2014;2014:376–387.
- 64 Green ED, Guyer MS: Charting a course for genomic medicine from base pairs to bedside. *Nature* 2011;470:204–213.
- 65 Malinowski J, Naylor H: A rapid evidence review for the inclusion of genetic data in clinical care for a common, complex disease. Submitted.
- 66 Green RC, Berg JS, Grody WW, Kalia SS, Korf BR, Martin CL, McGuire AL, Nussbaum RL, Daniel JM, Ormond KE, Rehm HL, Watson MS, Williams MS, Biesecker LG: ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet Med* 2013;15:565–574.
- 67 Gwinn M, Grossniklaus DA, Yu W, Melillo S, Wulf A, Flome J, Dotson WD, Khoury MJ: Horizon scanning for new genomic tests. *Genet Med* 2011;13:161–165.
- 68 Manolio TA, Chisholm RL, Ozenberger B, et al: Implementing genomic medicine in the clinic: the future is here. *Genet Med* 2013;15:258–267.
- 69 Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PR, Bernstam EV, Lehmann HP, Hripisak G, Hartzog TH, Cimino JJ, Saltz JH: Caveats for the use of operational electronic health record data in comparative effectiveness research. *Med Care* 2013;51:S30–S37.
- 70 Crawford DC, Sanders CL, Qin X, Smith JD, Shephard C, Wong M, Wittrak L, Rieder MJ, Nickerson DA: Genetic variation is associated with C-reactive protein levels in the Third National Health and Nutrition Examination Survey. *Circulation* 2006;114:2458–2465.
- 71 Jeff JM, Brown-Gentry K, Crawford DC: Identification of genetic modifiers within the fibrinogen gene cluster for fibrinogen levels in three ethnically diverse populations. *Pac Symp Biocomput* 2015;2015:219–230.
- 72 Jeff JM, Brown-Gentry K, Crawford DC: Replication and characterisation of genetic variants in the fibrinogen gene cluster with plasma fibrinogen levels and haematological traits in the Third National Health and Nutrition Examination Survey. *Thromb Haemost* 2012;107:458–467.
- 73 Kho AN, Hayes MG, Rasmussen-Torvik L, Pacheco JA, Thompson WK, Armstrong LL, Denny JC, Peissig PL, Miller AW, Wei WQ, Bielinski SJ, Chute CG, Leibson CL, Jarvik GP, Crosslin DR, Carlson CS, Newton KM, Wolf WA, Chisholm RL, Lowe WL: Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. *J Am Med Inform Assoc* 2012;19:212–218.
- 74 Haiman CA, Fesinmeyer M, Spencer KL, et al: Consistent direction of effect for established T2D risk variants across populations: the Population Architecture Using Genomics and Epidemiology (PAGE) Consortium. *Diabetes* 2012;61:1642–1647.
- 75 Dumitrescu L, Carty CL, Taylor K, et al: Genetic determinants of lipid traits in diverse populations from the Population Architecture using Genomics and Epidemiology (PAGE) Study. *PLoS Genet* 2011;7:e1002138.
- 76 Bowton E, Field JR, Wang S, Schildcrout JS, Van Driest SL, Delaney JT, Cowan J, Weeke P, Mosley JD, Wells QS, Karnes JH, Shaffer C, Peterson JF, Denny JC, Roden DM, Pulley JM: Biobanks and electronic medical records: enabling cost-effective research. *Sci Transl Med* 2014;6:234cm3.