Using electronic health records to drive discovery in disease genomics

Isaac S. Kohane

Abstract | If genomic studies are to be a clinically relevant and timely reflection of the relationship between genetics and health status — whether for common or rare variants — cost-effective ways must be found to measure both the genetic variation and the phenotypic characteristics of large populations, including the comprehensive and up-to-date record of their medical treatment. The adoption of electronic health records, used by clinicians to document clinical care, is becoming widespread and recent studies demonstrate that they can be effectively employed for genetic studies using the informational and biological 'by-products' of health-care delivery while maintaining patient privacy.

There is a growing and as yet unmet need in disease genomics research to obtain larger populations that are accurately clinically characterized and for which genomic characteristics can be measured affordably[1]. For example, very large populations will be important for crucial goals such as understanding the biological implications of rare variants and the disease risk profiles of common and rare variants, and their interactions[2–6]. In addition, human health and medicine are rapidly changing. An example is the current obesity epidemic in many countries and the novel therapies that are being used to treat its consequences, such as diabetes mellitus. As a result, there is a pressing need to address questions in disease genomics at the population scale and to answer them in just a few months rather than decades. Furthermore, with increasing financial pressures on the scientific and health-care establishments, these large and timely population studies of unprecedented size now have to be performed at much lower costs per subject. Many cost-savings have already been made in the area of genomic measurements[7] (see Genome.gov DNA sequencing costs) but the clinical characterization and sample acquisition costs have thus far been stubbornly resistant to similar declines[8–10].

The increased functionality of electronic health records (EHRs), linked biorepositories and the increased performance of natural language processing (NLP) provide an opportunity to use the informational and biological products of clinical care itself to drive large cohort studies on a wide range of diseases. At its core, the use of EHR data for genomic research enables two workflows in the investigation of population-wide genomic characterization. I refer to these types of workflow as EHR-driven genomic research (EDGR). In the first workflow, patients with characteristics matching those of interest — for example, a disease category such as rheumatoid arthritis or a lack of clinical response to serotonin-specific reuptake inhibitor (SSRI) antidepressants — are selected via EHRs using a combination of structured, codified and narrative text. The populations of patients thereby characterized are then recruited to provide samples or have their discarded clinical samples analysed for genomic research. Conversely, in the second workflow, EHRs are used to provide additional clinical characterization or to fill in missing details on subjects whose samples have already been collected for either a biobank or other cohort study. The individuals are usually selected for inclusion in a biobank or cohort as part of a broad population characterization effort or as part of a study into a specific disease[11]. However, those clinical characteristics that are not easily obtained during the recruitment into the cohort (for example, medications prescribed or laboratory values obtained during clinical care) can then be acquired from EHRs to augment the phenotypic characterization of the subject. The first workflow can be described as EDGR for phenotyping, selection and sample acquisition, whereas the second can be described as EDGR for phenotypic augmentation. In either workflow, the goal is to provide a rich set of clinical characterizations of the patient's state and lifetime trajectory at a low cost and a high degree of timeliness, matched to the corresponding DNA samples. These solutions have the potential to move a study from the stage of posing a question to that of having collected samples, at orders of magnitude lower

Biorepository
A biological materials repository that collects, processes, stores and distributes biospecimens to support future scientific investigation.

Natural language processing
(NLP). A field of computer science and linguistics concerned with the interactions between computers and human (natural) languages. NLP techniques allow the text in electronic medical records to be transformed from a clinical narrative to a set of codified terms or tags that are more readily subject to computational and statistical analysis.

Harvard Medical School, 10 Shattuck Street, Boston, Massachusetts 02115, USA.
e-mail: isaac_kohane@harvard.edu
doi:10.1038/nrg2999

Table 1 | **Relative properties of EDGR and conventional cohort studies**

	EDGR	Conventional cohort-driven research
Timeliness	Contemporary, reflective of current exposures	Varies with how often the cohort is 'refreshed'
Cost	Marginal cost in addition to processes already used for health-care delivery	Significant cost of parallel research infrastructure, including annotation and sample acquisition
Populations studied	Representative of a clinical care population with up-to-date treatments	Determined by the study/research protocol
Ability to assess comorbidity of genomic associations and phenome-wide scans	Made possible by a comprehensive health record	Only possible if very broad and detailed phenotyping is undertaken[17]
Family history information	Fragmentary at best	Systematically obtained
Environmental exposure	Variable quality for smoking or substance abuse history. Dietary and exercise history are largely absent	Some cohorts have an accurate dietary history; exercise and other exposures are not routinely assessed in health care
Medication and pharmacogenetics information	Up-to-date and comprehensive. Includes recently introduced medications	Up-to-date to the extent that there is synchronization with health-care data
Data accuracy	Can only be reliably ascertained for a subset of relevant patients	Systematic quality control for those data elements that are gathered
Range of data included	Broad reflection of clinical states and exposures	Determined by the protocol. Can be broadened but with a significant time-lag
Controls	Can be selected for every new study from the entire population and filtered incrementally	Can be contaminated with cases, especially when merging multiple groups
Identification of confounders	In addition to genomic stratification, ethnic self-identification, socio-economic status and geographic location, confounders in clinical care can be ascertained	Ancestry-informative genetic markers are routinely obtained but other clinical confounders are systematically identified in only a few cohorts
Consent regimen	Opt-out and opt-in regimens that are currently used may not be generally or durably acceptable	Cohort-specific regimens have withstood several decades of scrutiny and public debate

EDGR, electronic health record-driven genomic research.

cost and more quickly than the current standard practice. Currently, the recruitment, testing and questioning of subjects is driven by study personnel and occurs in population genomic studies that are largely conducted outside the confines of the health-care delivery process.

Biobanking procedures and policies have been reviewed extensively[11–17]. Consequently, although this Review addresses both EDGR workflows, it focuses mainly on the first workflow (EDGR for phenotyping, selection and sample acquisition). Discussions of the second workflow (EDGR for phenotypic augmentation of subject data in biobanks) are restricted to those aspects that entail the use of EHRs. EHR-driven genome-scale studies are also not without their critics or limitations, but with the substantial governmental investments and incentives in EHR adoption, these studies are likely to become increasingly common. Here, I discuss in more detail the advantages and limitations of EDGR, review the components of the approach — including enrolment, sample acquisition and consent regimens — and contrast it with existing methodologies. The successes of EDGR in replicating conventionally constructed genome-wide association (GWA) studies and developing novel findings are summarized. I also provide a near-term roadmap and set of challenges for the ongoing development of EDGR.

Biobank
A cryogenic storage facility used to archive biological samples for use in research and experiments. Ranging in size from individual refrigerators to warehouses, biobanks are maintained by institutions such as hospitals, universities, non-profit organizations, pharmaceutical companies and national biorepositories. More recently, the term biobank has been used to signify a population cohort study with stored biological samples.

Advantages of EDGR

Clinical relevance and cost considerations. The most important motivation for EDGR is that of clinical relevance, with cost a close second (TABLE 1). Much has been written about the insatiable appetite of genome-scale research for ever larger cohorts[1,4,5]. This is true not only for obtaining significance for weakly penetrant alleles but also for identifying rare alleles that have large biological effects and their interactions. However, just as important as scale is the relevance to the clinical care of the patient populations studied. All the complexity of polypharmacy and environmental influences that are present in clinical practice are not necessarily considered in studies designed to detect genetic associations or in clinical trials[18–20].

As well as answering clinical questions, new approaches are needed to answer combinations of questions that no cohort was designed to answer — and key to this is addressing questions in a timely manner. For example, take a hypothetical case in which it is reported within months of a new oral hypoglycaemic drug coming to the market that there is an increased incidence of myocardial infarction, but not in all those taking the drug. Can we identify, in a cost-effective manner, those genetic variants and clinical characteristics that together identify those individuals who are not at risk? To answer questions of this type requires

measuring population characteristics, obtaining samples and characterizing genotypes or sequences within a relatively short time period for tens of thousands of selected individuals across millions of patients. That is, if we are to achieve a clinically relevant genomically based medicine, at least one part of our collective genomic discovery process has to be able to nimbly and cost-effectively 'instrument' (that is, comprehensively characterize) the clinical populations that we hope to help in their 'natural' state.

Indeed, timeliness is essential for characterizing patients in a health-care system in constant evolution. Medications and procedures are changing every year and at an accelerating pace, particularly those informed by molecular and genetic signatures[21]. The distribution of disease burden across world populations is also shifting, as exemplified by the current worldwide epidemic of diabetes mellitus. There are also highly acute changes such as those caused by infectious disease[22]. In addition, changing financial and discovery-driven pressures alter the nature of how health care is delivered (for example, which drugs are used) and how patients are described[23]. Consequently, a phenotypic characterization obtained in the previous years (whether 20 years ago or sometimes even just the previous year) could be totally inadequate. What is required is a current clinical snapshot of the cohort being studied so as to be most relevant to a genomic–phenotypic analysis. For example, it is likely that had a genome-wide scan for diabetes-associated variants been possible 50 years ago, the genetic variants implicated would not correspond well to those that underlie susceptibility today — not least because the disease has changed by virtue of a change in environment.

EDGR is uniquely positioned to comprehensively characterize populations in our health-care delivery systems in close to real time. In EDGR the re-phenotyping of patients can occur day by day and can be studied with respect to the mostly unchanging DNA samples, which are available recurrently from discarded samples from each individual. Even if somehow one could rapidly measure the health status of sufficient numbers of relevant individuals using conventional population study techniques to address the types of questions outlined above, it would be quite unaffordable to do so. Indeed, just 4 years ago the US Secretary's Advisory Committee on Genetics, Health and Society argued for the need for a million-strong population genetic study but estimated that this would cost at least US$3 billion[24]. Of this cost, as detailed in Murphy *et al.*[25], only a fraction is for genomic analysis. The bulk is for the study infrastructure, research staff and obtaining patient biosamples with detailed phenotypic characterization (that is, including such important details as clinical outcomes, medications and laboratory tests), which typically costs $1,000 per sample and often ten times that amount[8–10,26]. In contrast to conventional cohort studies, EDGR-based approaches are at least one order of magnitude cheaper per sample and run two orders of magnitude more rapidly for sample collection. This cost analysis includes per-study costs comprising the optimization of NLP phenotyping, statistical modelling, and incremental computational and biobanking infrastructure. When this cost is spread over the millions of patients analysed in even a single hospital system, the performance difference per sample is due to the enormous investment in data gathering and sample acquisition that is already 'baked' into the cost of health care. That is, the efficiency advantage of EDGR comes from maximizing the research utility of that clinical-care investment such that it is only a fraction of what it would take to create a *de novo* research cohort pipeline[6]. This 'productive parasitism' is a strategy that is likely to grow with increased penetration of the EHRs into clinical care[27]. Such cost differentials represent the difference between achievable and unrealizable research. If we can study a million patients for a total of $100 million instead of $3 billion[25] because of a marginal cost per sample of $20 (rather than >$1,000) and with phenotyping costs well under $0.01 per patient for large populations, then the very nature and breadth of the research that can be accomplished will be influenced directly.

Other advantages of EDGR. As noted by Collins *et al.*[28], writing on behalf of the US National Human Genome Research Institute (NHGRI) in 2003, obtaining adequate representation of minorities and underserved populations is a priority for future genomic research. EDGR-identified cohorts tend to reflect the populations cared for; this is in contrast to most clinical research recruiting, in which the efficiency of recruiting underserved populations (and therefore their representation in the resulting studies) has typically been low. For example, African Americans are less represented in clinical research studies than their population size would suggest[29–31]. By contrast, they are often overrepresented in the populations cared for in emergency medical departments[32–34], which are often managed by the same large health centres that are leading the implementation of EDGR. 'Capturing' the clinically cared for populations also allows us to enrich for rare diseases requiring specialized treatment, which are typically concentrated in tertiary health-care centres that serve as referral basins for these populations.

There are other less-significant advantages of EDGR that are worth mentioning. One is the ability to perform in-depth comorbidity analyses to determine whether genetic variants predispose to just a single disease or rather a set of diseases with common pathways. Indeed, not captured in most genome-wide studies is the fact that the majority of patients above age 65 have more than one chronic condition[35]. Also, just as it has been shown that health-care systems can be used to detect adverse drug events in a timely fashion using EHRs[36,37], so can a timely pharmacogenetics study be implemented using EDGR[38]. The EDGR-based identification of genetically (and clinically) characterized subpopulations may allow a more nuanced understanding of where contemporary drugs can be of true benefit and where these benefits are overshadowed by greater potential harm for specific individuals.

EDGR also helps to address the thorny matter of stratification by continent of origin. Genomic studies typically and consistently can identify the continent of origin of a biological sample using a relatively small number of variants[39], which is used to control for population stratification. Dumitrescu *et al.*[40] have shown that ancestry, defined by informative SNP markers, does not differ significantly from EHR-recorded (including self-reported) ancestry in European Americans and African Americans. Nonetheless, they recognize that there may be other factors that influence self-reported ancestry. These factors might be proxies for other forms of stratification (for example, socio-economic) that are not captured by genetic markers but that are potential confounders. That is, given two individuals with the same ancestry-informative markers, their biology might be more confounded by their socio-economic status and present geography — which are readily determined through EDGR — than they are by their genetic markers. More generally, there are other confounders, such as medical treatments, that are rarely captured in cohort studies but can be identified through EDGR.

Finally, the selection of controls for genomic population studies has been recognized as an ongoing problem, particularly for the more common diseases for which control sets 'borrowed' from other studies may actually be contaminated with cases[41]. Because EDGR allows any control population to be cost-effectively rescanned for their full set of clinical phenotypes, some of these challenges in the selection of controls can be mitigated.

What studies are under way?

Arguably the best known and largest biobanks and cohorts are those in Europe and Japan, with the deCODE Icelandic project serving as an early archetype[42]. For the most part the phenotypic characterizations of these patients have been acquired by research assistants and questionnaires separate from the health-care record (TABLE 2). If the health record was used, it was only for phenotypic augmentation.

By contrast, studies specifically driven by recruitment through EHRs (that is, EDGR) are mainly taking place in the United States, despite the highly fragmented health-care system and health information technology infrastructure. Several health-care systems in the United States have started accruing large EHR databases, often augmented with NLP characterization (see below and FIG. 1) and linked to discarded clinical biosamples. Between them, the different repositories represent well over 100,000 individuals. Indeed, international biorepositories appear poised to follow suit by linking their samples to EHR data in the near future (TABLE 2). Notable among the efforts in the United States are the Harvard University/Partners Healthcare system i2b2 effort[25,43], the Vanderbilt BioVu[44] effort and the multicentre, NHGRI-funded eMERGE Network[45,46]. As discussed in detail later, genomic studies have already been conducted and published by these researchers, including validation of previous conventional cohort-driven GWA studies[47,48] by using populations characterized and selected using EHR data. That is, in addition to novel discoveries, these studies have replicated the direction and order of magnitude of the odds ratios of prior GWA studies.

Components of EHR-driven research

Enrolment. Although there are many ways in which EDGR can be conducted, a simplified perspective may be helpful (FIG. 2). Patients or subjects enter the system in two main ways: through enrolment at the outset of joining a clinical-care system as a patient; or enrolment and/or recontact once it has been determined that a patient's data and biosamples would be useful. Either way, their corresponding EHR data are then loaded into a study-specific or system-wide database. The level of phenotypic characterization of patients in this database is dependent on the nature of the health-care data. Many health-care systems frequently maintain codified, structured formats for laboratory studies and for billing diagnoses and less so for immunizations, in-patient medications and allergies. However, other clinical data types — such as outpatient medications, adverse effects, fine-grained diagnostic assessments, smoking history or family history[49] — are rarely codified and, if present at all, are buried in narrative texts. Examples of these texts include clinic notes, discharge summaries, radiology reports, electroencephalography (EEG) reports and pathology reports. Computer science techniques that are collectively designated as NLP[50] must be used to transduce these textual files into structured data in controlled vocabularies[51] such as the Systematized Nomenclature of Medicine (SNOMED) or the International Classification of Diseases, 10th revision (ICD10) (FIG. 1). However, none of the commercial or academic (free of charge) NLP packages works with the classification accuracy necessary to avoid the problems of phenotypic misclassification that have been documented in genomic studies[41]. Rather, a multi-month iterative refinement of the NLP queries — involving a team of computer scientists, statisticians, and clinical and genomic experts — is required to attain positive predictive values of at least 95% for the intended phenotypes[52,53]. Reassuringly, early studies across different health-care centres have demonstrated that NLP characterizations developed in one institution can be rerun within other health-care systems with remarkable fidelity, as judged by a panel of clinical experts at those other institutions comparing the NLP phenotyping to their own 'gold standard' reading of the patients' records (D. Masys, personal communication). Once the NLP processing is completed, researchers are left with an unambiguous set of codified phenotypic characterizations that are used as in conventional GWA studies.

Although NLP does require considerable effort and a systematic, statistically informed approach, the costs of rerunning these NLP procedures are marginal. In total, including the original NLP refinement and tuning process, NLP procedures cost one or two orders of magnitude less than carrying out the phenotyping for conventional clinical or research databases, in which such information about human subjects is gathered during the study process[25]. Furthermore, because EHR systems tend

Population stratification
The presence of a systematic difference in allele frequencies between subpopulations from a larger population, possibly owing to different ancestry, especially in the context of association studies. (Population stratification is also referred to as population structure in this context.) If not properly accounted for in association studies, population stratification can lead to spurious associations.

Controlled vocabularies
A controlled vocabulary only includes terms that have been selected by the group that created the vocabulary. The goal of such a vocabulary is to standardize and simplify the organization of data and knowledge in a particular domain.

Table 2 | **Overview of EDGR efforts, with selected cohorts and biobanks included for comparison**

System (name/site)	EHR-driven patient/subject selection	NLP extraction of phenotype	Prospective research annotation	Restrictions	Consent	Study-specific versus system-wide database	Data identified or de-identified	Size
EDGR efforts								
eMERGE network (Group Health Cooperative, Marshfield Clinic, Mayo Clinic, Northwestern University, Vanderbilt University)	Yes	Not at all sites	Occasionally	Some sites do not include children	Mixed full consent and opt-out	Study-specific databases (some of which are extracted from system-wide databases)	De-identified	The aggregate studies of the consortium exceed 30,000 subjects
Kaiser Permanente Research Program on Genes, Environment, and Health (RPGEH)	Yes	Not at present	Sparse	No children <18 years old but they may be included in future	Varies by study: full consent to waived consent per IRB. Opt-in is default for biorepository	System-wide database and study datamarts supported	De-identified or identified as per the study design and the IRB review	>140,000, to grow to 400,000 in the near future
i2b2 (Partners Healthcare System/Harvard Medical School)	Yes	Extensive use to select from 4 million patients using 50 million documents	Sparse	No children <18 years old	Patient consent-to-care includes use of anonymized data	Study-specific databases	De-identified	Studies of 2,000–12,000 subjects/patients selected 'on demand'. Eight studies are underway to date
BioVU (Vanderbilt University)	Yes	Extensive use across 120 million documents on 2 million patients	Sparse	None	Opt-out	System-wide database with derived datamarts	De-identified	>104,000 patient/subjects
Cohorts and biobanks								
Estonian Genome Center	Not yet	Planned. Not at present	Extensive	No children <18 years	Full consent	National system	De-identified	52,000 'donors' to date. Growing steadily
UK Biobank	No	Not at present	Extensive	No children (includes ages 40-69 years)	Full consent	Single, centrally managed cohort study	Only de-identified data will be released	500,000 subject target reached
Biobank Japan Project	Not yet	Planned. Not at present	Extensive	None, but access to children's data is more tightly controlled	Full consent	System-wide database	De-identified	>200,000 subjects
Danish National Biobank	Yes (including combined EHR and Danish registry searches)	Currently only in select textual extracts but likely to include the full database in the near future	For specific disease groups	All citizens including children	Opt-out	Mostly study-specific datamarts	Depending on study design, includes both identified and de-identified data sets	Size depends on which cohorts are included, but >100,000 subjects

EDGR, electronic health record-driven genomic research; EHR, electronic health record; eMERGE, electronic Medical Records and Genomics; IRB, institutional review board; NLP, natural language processing.

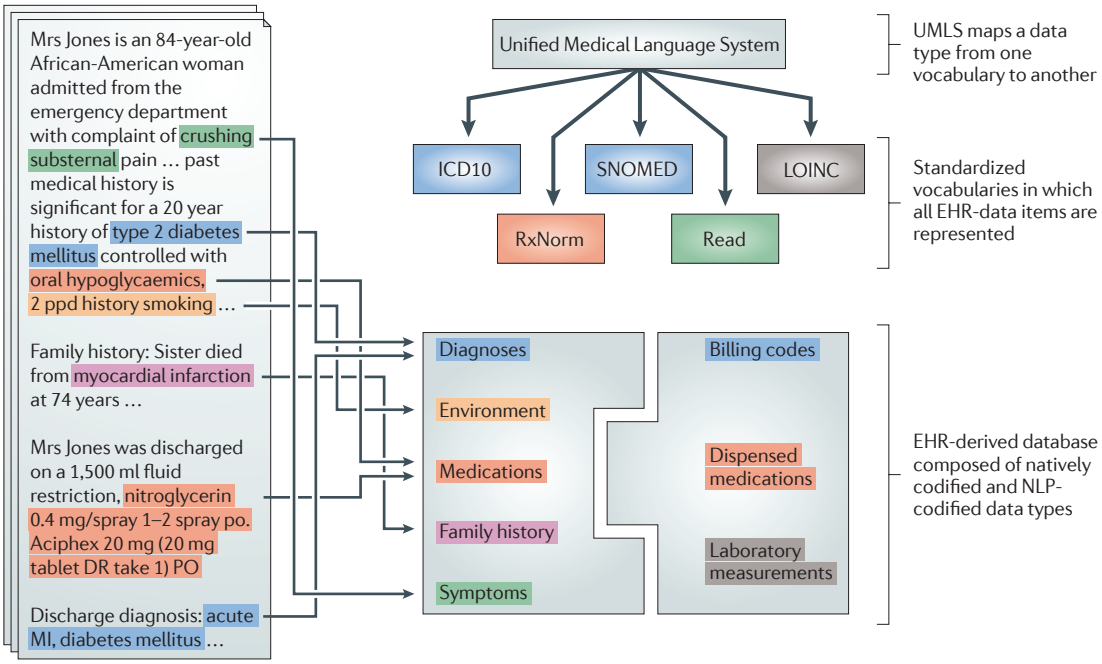


Figure 1 | **From clinical notes to structured phenotypes.** Natural language processing (NLP) identifies various concept types in the textual records that are associated with each patient for each medical record. For example, concept types could include discharge summaries and radiology reports. These concept types are then associated with a term selected from a standardized vocabulary (for example, RxNorm[93] for medications, Read vocabulary for symptoms[94] or the International Classification of Diseases, 10th revision (ICD10)[95] for diagnoses). These terms are then joined to the elements of the medical record that are already codified (for example, billing codes for diagnoses, or laboratory results). Not shown in this diagram is the identification of temporal and other relationships between concepts (for example, medication given to treat a diagnosis, or obesity onset after chemotherapy). The Unified Medical Language System (UMLS)[51,96] is a National Library of Medicine resource that maps the contents of one vocabulary system to another so that, for example, pathological diagnoses codified through NLP using Systematized Nomenclature of Medicine (SNOMED)[97] can be compared to diagnoses entered through a billing system in ICD10. EHR, electronic health record; LOINC, Logical Observation Identifiers Names and Codes.

to cover the full range of human disease, an interesting inversion of the characterization of genotype–phenotype associations can be implemented when NLP is used: that is, the phenome-wide association study (PheWAS). As articulated by Jones *et al.*[54] and implemented by Denny *et al.*[55], a full scan of all diseases and disease endopathotypes can be performed for significant association with any genetic variant or set of such variants. For example, a SNP implicated with genome-wide significance for diabetes mellitus in a standard GWA study can be promptly run against every EHR-derived phenotype, such as obesity, heart disease, smoking history or hypothyroidism. Such a scan may show that the original association is either an epiphenomenon of another pathology or part of a broader endopathotype[56]. EDGR thereby provides the opportunity to explore this broader range of pathological mechanisms across a range of disease types, which is not possible in single phenotype studies. Nor are all the relevant phenotypes likely to be documented *a priori* in even the broadest cohort studies because of the dynamic nature of clinical care and human epidemiology. Indeed, it could be argued that any genetic study should start with a full phenome scan to avoid a myopic perspective on the clinical implications of any set of genetic variants.

Sample acquisition. As mentioned above, although the costs of genome-wide characterization of genetic variation have consistently and exponentially decreased, the costs of acquiring samples to be characterized have been stubbornly slow to fall[10]. Owing largely to these high costs, sample acquisition rates have remained low. By contrast, two approaches used for EDGR now provide sample acquisition that is orders of magnitude faster and cheaper than previous approaches.

The first (exemplified by the i2b2 system[57]) requires that the patient cohort of interest is first selected *in silico,* as outlined above and in FIG. 2. A set of anonymous identifiers corresponding to this study-specific EHR 'datamart' of selected patients is then forwarded to the clinical laboratory information system. When the selected patients return to the health-care system and have a sample taken during clinical care, the clinical laboratory information system will forward the portion of the sample that is surplus, or not required for clinical measurements, to a study-specific biobank or biorepository. The numbers of patients for common diseases such as asthma can exceed 100,000, even in a small health system of <4 million patients[25]. Although the sample acquisition rate decreases with time as the selected cohort is depleted, the average over a period of 1 year is approximately 200–500 samples per week.

Phenome
The set of all phenotypes expressed by a cell, tissue, organ, organism or species.

Datamart
The entire stored data of an enterprise (for example, a health-care centre) is often termed the data warehouse. For a specified purpose (for example, a disease-specific study), a subset of the data warehouse, called the datamart, is extracted for a group of analysts.

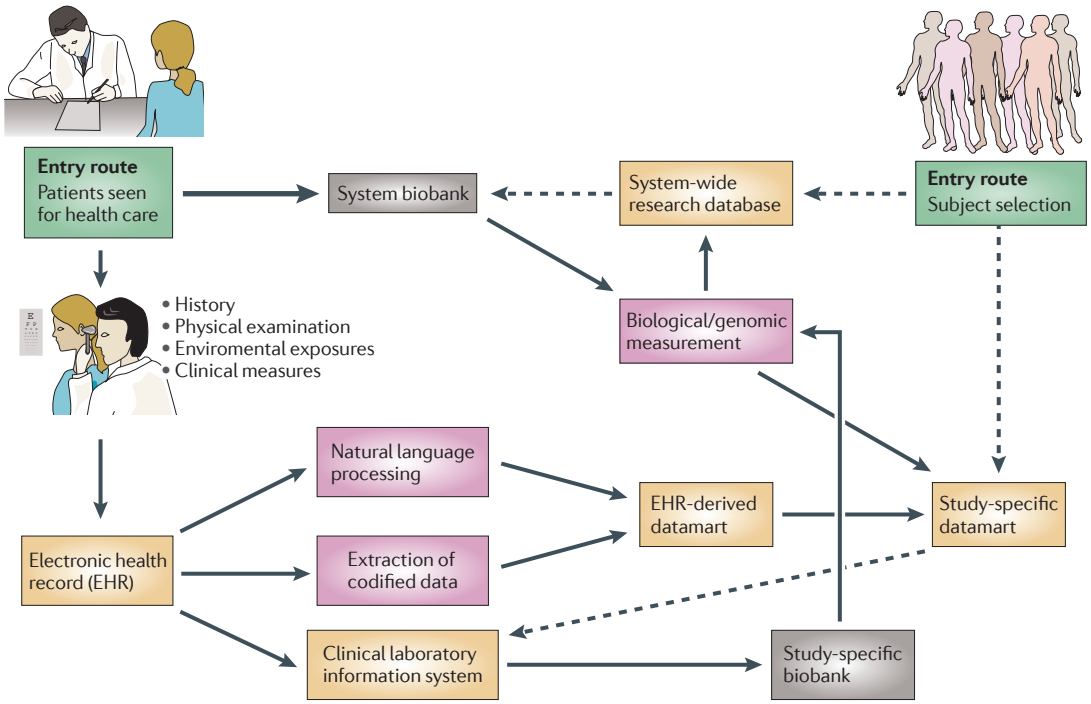


Figure 2 | **Two archetypal workflows in electronic health record-driven genomic research.** Electronic health record-driven genomic research (EDGR) starts with the entry of patients or subjects into the health-care system. In one workflow (green box, left side of the figure), all patients are entered prospectively into a single database and their samples are entered into a single biorepository (for example, BioVU[44]). In another workflow (green box, right side of the figure), it is only when a research question is asked that a study-specific datamart is generated, which then triggers accrual of patient samples discarded by the clinical laboratory information system. Dashed lines represent the selection of subjects for a study.

In an alternative approach, exemplified by the BioVU system[44,58], excess blood from samples drawn during clinical care is banked after the patient registers with the health-care system. For BioVU there is an opt-out option that is exercised by less than 5% of all patients. In this approach, there is a single system biobank, which approaches the size of the total population of the health-care delivery system. Subjects can then be selected by phenotype through the system-wide research database derived from the EHR, or conversely by genotype, as described in the aforementioned phenome-wide study[55].

The types of biological measurements that can be performed on these samples include not only studies of the human genome but also metagenomic studies that can be carried out using blood, urine and faeces specimens that are obtained through the course of clinical care. Epigenetic marks and the more stable RNA species (for example, microRNAs (miRNAs)) can also be measured reliably (L. Bry, personal communication). An interesting feature of EDGR, given the multiple clinical visits across a patient's lifetime, is the multiplicity of opportunities for obtaining additional specimens. That is, EDGR-derived biobanks are at least partially renewable resources. This suggests that genome-scale measurements need no longer be carried out just once, but could evolve into an on-demand process that is repeated with each new epoch of genome-scale measurement technologies.

Metagenomics
The study of metagenomes, which consist of genetic material recovered directly from environmental samples. Increasingly, it is used to describe the shotgun sequencing and analysis of the microbial genomes found in the milieu of the human body and its waste products.

Consent regimens. Most EDGR efforts implement one of a small number of possible consent regimens[45]. Increasingly the most popular is an optional opt-out consent[58] obtained during the enrolment for clinical care. In this consent regimen, patients are informed that their de-identified clinical data and specimens are provided to researchers to conduct genomic studies. The stipulation of anonymity thereby precludes the recontact of patients. An alternative model, which appears to be the dominant model in Europe and Asia for traditional cohort and biobank studies, is to provide consent for full identification as part of a research protocol. In this regimen the EHR data (for example, updated medications and morbidities) are subsequently and recurrently appended to the clinical characterizations obtained during the original study recruitment. One important concern regarding EDGR is that the details of the consent regimens of specific health-care systems may limit efforts for broader data sharing efforts, such as the database of Genotypes and Phenotypes (dbGaP)[59]. That is, if patients do not consent for the use of their data beyond the health-care system (which is currently the default), shared genomic–clinical repositories will not be feasible and distributed repositories or meta-analyses of local results will instead be required. As discussed in 'Challenges for EDGR' below, despite the initial success of EDGR, a durable, broadly accepted consent regimen has yet to emerge. Furthermore, there is a lack of consensus as to what the

consent process should include with regard to the investigator's responsibilities for follow-up and interpretation of measurement data obtained in EDGR. Should these responsibilities be similar to those of a clinician[60]?

Results so far from EDGR

EDGR studies were first designed to demonstrate feasibility and comparability: that is, whether EDGR studies could reproduce the results obtained through conventional cohort studies. Ritchie *et al.*[47] studied 9,483 patient-samples accrued over a 4-month period. They genotyped these samples for SNPs previously found to have genome-wide significance for one of the following disorders: atrial-fibrillation, Crohn's disease, multiple sclerosis, rheumatoid arthritis or type 2 diabetes. Each of the 21 SNPs genotyped was found to have odds ratios in the same direction as the original studies. Moreover, when the original reported odds ratio of a SNP was greater than 1.25, the odds ratios from the EDGR study reached statistical significance. In another study, Denny *et al.*[61] studied the variability in the PR interval of electrocardiograms, a measure that reflects atrioventricular conduction, using an EDGR-derived population of 2,334 patients. In this study, PR intervals, clinical characteristics and medications were all automatically extracted. The identification of four SNPs in the sodium channel, voltage-gated, type X, alpha subunit (*SCN10A*) gene, which had been previously implicated by three conventional GWA studies, was replicated and these SNPs were found to contribute to 11% of the observed variance of the PR interval.

More recently, EDGR studies have reported not only replication of prior studies but also novel results. Kurreeman *et al.*[48] used a population of 1,515 EDGR-selected patients with rheumatoid arthritis, together with 1,480 controls matched for both genetic ancestry and disease-specific autoantibodies (anti-citrullinated protein antibodies (ACPA)). These researchers genotyped 29 SNPs that had been implicated with genome-wide significance ($p < 5 \times 10^{-8}$) in at least one study or in a recent meta-analysis. Of the 29 SNPs, 26 had odds ratios in the same direction as the original study, of which 16 achieved statistical significance. The three remaining SNPs had odds ratios close to 1.0 in the original studies. The authors also extended the knowledge gained from the previous studies. They showed that the aggregate genome risk scores for rheumatoid arthritis defined in the prior studies (across the measured SNPs) are significantly different from controls in individuals of European ancestry, and for the first time they demonstrated the applicability of these risk scores to African and Hispanic Americans[48]. Furthermore, there was previously little evidence relating to whether the majority of risk alleles discovered in ACPA-positive disease also contribute to risk in ACPA-negative disease, but the results showed that the genome risk scores in ACPA-negative subjects were significantly different from controls[48]. Individual SNP analyses revealed that the effect sizes and contributions of these SNPs were only partially overlapping those of the ACPA-positive individuals, suggesting related but not identical pathophysiologies in the two subpopulations.

In another recent study, Kullo *et al.*[46] used red-blood-cell characteristics as quantitative traits in a genome-wide study of 3,012 patients. These characteristics were identified from codified and NLP-derived parameters in the Mayo Clinic's EHR. In this study, 11 SNPs in four genomic regions were identified with genome-wide significance ($p < 5 \times 10^{-8}$) as being associated with specific red-blood-cell phenotypes. Three of the loci (corresponding to the *HBS1L–MYB* interval; transmembrane protease, serine 6 (*TMPRSS6*); and haemochromatosis (*HFE*)) had been previously identified in GWA studies for haematological traits and the fourth (solute carrier family 17 (sodium phosphate), member 1 (*SLC17A1*)) was found to be associated with mean corpuscular haemoglobin, although this finding awaits replication.

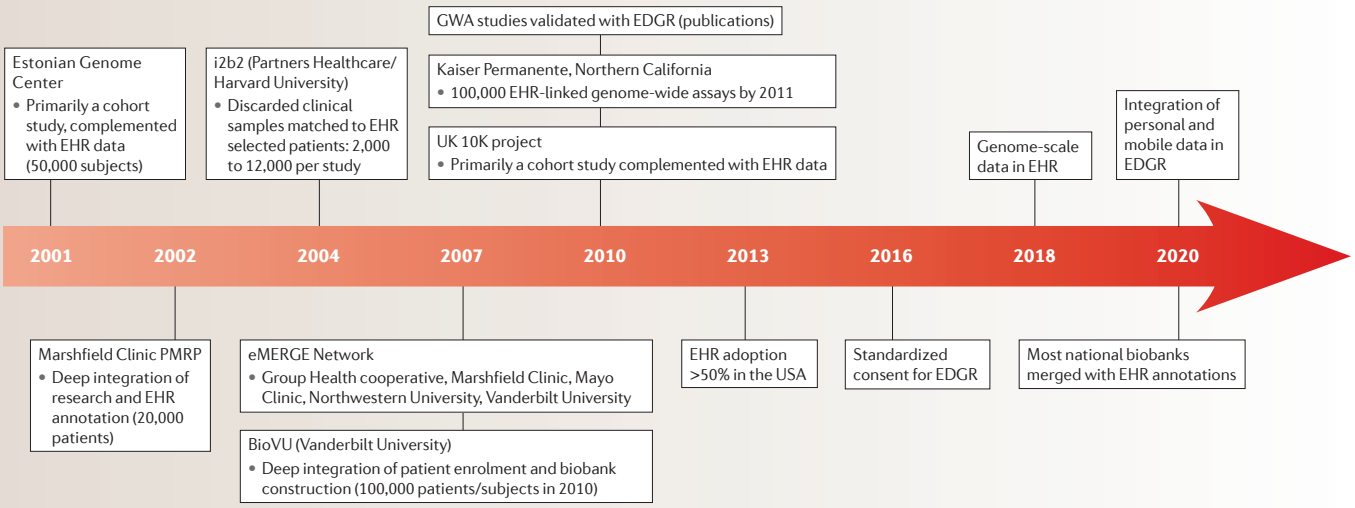
Finally, in the aforementioned phenome-wide scan[55], five SNPs that had been previously associated with clinically relevant traits were studied across the entirety of the diagnostic categories available within the EHR. New associations were identified; for example, rs3135388 was previously implicated in multiple sclerosis and systemic lupus erythematosus but was also found to be significantly associated with pulmonary heart disease, acute renal failure, conduct disorders, benign neoplasms of the gastrointestinal tract, intrathoracic and respiratory organs, and cancer of the rectum and anus. Except for multiple sclerosis, these other associations await additional studies for validation, but they present hypotheses that are readily tested in other genetic studies, particularly EDGR.

Challenges for EDGR

The foremost challenge for successful conduct of EDGR is the relatively sparse implementation of comprehensive EHRs, particularly in the United States[27]. There is a substantial investment required for EHR implementation[62,63]. This investment has been argued for on the basis of expected benefits to the delivery of clinical care and therefore does not enter into the marginal cost of EDGR. Yet without it, there will be insufficient accessible and computable data in the health-care system to drive EDGR. It is therefore not a coincidence that the leaders in EDGR have also led the charge for implementation of EHRs.

Another crucial concern is the lack of high-quality data[64]. Even with a comprehensive EHR, clinical documentation is highly variable and rife with error and imprecision. Moreover, the primary driver of EHR implementation has been clinical reimbursement rather than the potential for reuse of the clinical data for research[65]. This effect is particularly noticeable in the billing codes[8] and creates a set of confounders that are opaque to most researchers. For example, clinicians will often bill the evaluation for a diagnosis (such as rheumatoid arthritis) as if the patient had rheumatoid arthritis, even if the evaluation did not reveal the disease. In the EDGR efforts so far, these limitations have been partially addressed by sub-sampling those populations with the least such error and imprecision — that is, characterization with high specificity

Timeline | **The use of electronic health records in human disease genomics research**



Key achievements so far and ongoing projects are shown, as well as a roadmap for near-term developments in electronic health record-driven genomic research (EDGR). EHR, electronic health record; eMERGE, electronic Medical Records and Genomics; GWA. genome-wide association; PMRP, Personalized Medicine Research Project.

but low sensitivity. Fortunately, these confounders can further be eliminated by the NLP process, whereby clinical experts review the records and tune the NLP algorithms for clinical accuracy rather than reimbursement optimization. For example, in identifying patients with rheumatoid arthritis, a positive predictive value of 94% can be achieved using NLP[52]. Similar or better performance has been shown for attributes such as smoking and medication history[53,66]. However, with the increasing focus on gene–environment interaction, it is of some concern that EHRs provide only the exposures that clinicians have been routinely trained to gather: substance abuse, smoking history and medication history. Other important environmental considerations such as dietary and exercise history, toxin exposure, or workplace milieu are only captured in those cohort studies that make the explicit and considerable effort to obtain these data. Of interest, many practitioners of EDGR have noted that once clinician-researchers start using these systems, they notice that their own research benefits from the quality of clinical annotations and take the lead locally and nationally to increase the accuracy and completeness of the medical record. Indeed it may be an ironic side effect of genomic research that it results in an increase in the quality of EHRs.

Combining data across multiple EHRs or larger population studies is particularly challenging in those countries that lack a national health identification system or have a system that is error prone. To overcome the risks of 'double counting' patients in queries across multiple EHRs, investigators use pseudo-probabilistic matching schemes on common demographic attributes[67,68]. In this context, matching genomic 'fingerprints' across health-care systems promises much greater accuracy in uniquely identifying patients, but raises many data privacy issues[69].

Like existing genetic case–control and cohort studies, EDGR studies are currently observational. Therefore, they too are limited strictly to demonstrating probabilistic dependencies between variants and phenotypes. Causal implications require additional lines of evidence, whether experimental or through alternative study designs, such as the randomized clinical trial. Indeed, EDGR is most usefully understood as an early and cost-effective step in a discovery and validation process that is followed by many more steps using the full armamentarium of genetic investigation.

Finding a durable consent regimen that meets societal and scientific goals remains a considerable challenge. When patient data and samples for research are shared (particularly beyond the confines of the institution where a patient has been cared for and without their explicit consent to do so) patients may demand the destruction of the biobank, as recently occurred after a class action lawsuit filed by the parents of neonates screened in Texas[70]. However, consent alone may not address multiple important societal goals, such as justice and accurate biomedical science. As outlined by Taylor[71], the underserved and underrepresented, such as low income, aboriginal or elderly populations, are those most likely to refuse consent. The growing awareness of this challenge has led to pilot projects in EDGR where patients are explicitly educated about the benefits and risks of EDGR during health-care encounters and then can opt to join 'informed cohorts'. In these informed cohort studies, not only are anonymous health-care data and discarded samples studied, but patients control whether they are contacted for additional studies or follow-ups and they can also choose which medical topics they are contacted about[72,73].

A near-term roadmap for EDGR

The TIMELINE illustrates a plausible short-term trajectory for EDGR. It presupposes that the investment in EHRs, already alluded to, continues apace. European and Asian biobanks — which have first focused on comprehensive population sample acquisition, comprehensive consents and initial phenotyping — will take advantage of their national infrastructure to use EHRs[74–76] for phenotypic augmentation of the cohorts. Perhaps because of the substantial effort required for the establishment of national biobanks, patient selection in these cohorts using EDGR is only beginning to occur[77], driven by the need for more timely and cost-effective characterization (for example, adverse events for new medications). Also, NLP efforts in the United States have benefited from the relative uniformity of the language that is spoken across a population of more than 300 million. This has allowed multiple NLP development efforts to achieve important synergies, which are harder to achieve across smaller countries that use different languages.

Consent procedures will have to become standardized, especially in the United States, if populations across more than one health-care centre are to be combined to achieve the scale of millions of patients that will be required to identify sufficient numbers of rare variants or epistasis between common variants of modest effect[4,5]. Among the unresolved issues are: the adoption of opt-in versus opt-out procedures; consent to share genomic and clinical data obtained during clinical care[58] (beyond the immediate boundaries of the local health-care delivery system); and ownership of the derived intellectual property[78]. These challenges will also affect the biobanks as they attempt to aggregate across national boundaries[16]. In the European Union, discussion of standardization across the biobanks and consent procedures appears to be in its early stages[11]. In the United States, it may be that the centralized model that is exemplified by dbGaP[59] may not scale well to the genomic and clinical data that are associated with clinical care for both ethical and organizational reasons, although this too is under discussion. One of the thorniest questions remains whether unanticipated incidental findings should be communicated back to the patient and provider, particularly if the link to the EHR is maintained[60,72,79,80]. If centralized management of data is not achievable because of regulatory or legal challenges, then a form of distributed query system[81,82] that allows for the dynamic queries of the locally aggregated results will be required to drive large meta-analyses.

Genomic researchers will also have to become far more aware of the details and foibles of health-care system data and terminologies, in addition to codification and NLP. They will also have to work on establishing productive and ongoing collaborations with their clinical colleagues, for which they must obtain funding. Regarding funding, clinical reimbursement by payers, whether governmental or private, will have to include incentives to ensure sufficient quality of the data entered into EHRs, both for genomic research and also to improve the overall quality of the health-care enterprise.

Within the next 10 years two large movements will change the nature of EDGR. The first is the aforementioned merging of data from national biorepositories with EHR data, which will provide a quantum leap in the comprehensiveness and richness of information regarding these cohorts. However, the commoditization of genome-scale measurements within clinical care[83] and the adoption of these genomic data within EHRs will result in a much larger fraction of all patients' data in some form of database that can be used for EDGR. Increasingly, these institutional databases will be complemented by valuable data from informal sources such as social networks of patient support groups, mobile device reports and lifestyle instrumentation[84]. Temporally and spatially constrained public health records of potential exposures will provide additional environmental context[85,86].

Finally, the most significant prognostic for EDGR may well be the attitudes of consumers of health care[87]. Multiple studies have documented public willingness to participate in (and obtain personally relevant results) from genomic research[88–90]. Many, if not most, patients expect that the health-care system will return new insights and new medical therapies as one of the secondary benefits of clinical care. In the same breath, they expect privacy and control of access to their health data[91,92]. When these dual expectations are robustly implemented within a fusion of the health-care delivery and research enterprise[72], we are likely to achieve the goal of having every patient visit contribute to furthering our understanding of how the human genome and the environment interact across the spectrum of disease and health.

1. Green, E. D., Guyer, M. S. & National Human Genome Research Institute. Charting a course for genomic medicine from base pairs to bedside. *Nature* **470**, 204–213 (2011).
2. Ioannidis, J. P., Trikalinos, T. A. & Khoury, M. J. Implications of small effect sizes of individual genetic variants on the design and interpretation of genetic association studies of complex diseases. *Am. J. Epidemiol.* **164**, 609–614 (2006).
3. Dina, C. New insights into the genetics of body weight. *Curr. Opin. Clin. Nutr. Metab. Care* **11**, 378–384 (2008).
4. Gauderman, W. J. Sample size requirements for association studies of gene–gene interaction. *Am. J. Epidemiol.* **155**, 478–484 (2002).
5. Hein, R., Beckmann, L. & Chang-Claude, J. Sample size requirements for indirect association studies of gene–environment interactions (G x E). *Genet. Epidemiol.* **32**, 235–245 (2008).
6. Manolio, T. A., Bailey-Wilson, J. E. & Collins, F. S. Genes, environment and the value of prospective cohort studies. *Nature Rev. Genet.* **7**, 812–820 (2006).
7. Wheeler, D. A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–876 (2008).
8. Gismondi, P. M. *et al.* Strategies, time, and costs associated with the recruitment and enrollment of nursing home residents for a micronutrient supplementation clinical trial. *J. Gerontol. A Biol. Sci. Med. Sci.* **60**, 1469–1474 (2005).
9. Noble, S. *et al.* Feasibility and cost of obtaining informed consent for essential review of medical records in large-scale health services research. *J. Health Serv. Res. Policy* **14**, 77 (2009).
10. Schroy, P. C. *et al.* A cost-effectiveness analysis of subject recruitment strategies in the HIPAA era: results from a colorectal cancer screening adherence trial. *Clin. Trials* **6**, 597–609 (2009).
11. Zika, E. *et al.* A European survey on biobanks: trends and issues. *Public Health Genomics* **14**, 96–103 (2010).
12. Tutton, R., Kaye, J. & Hoeyer, K. Governing UK Biobank: the importance of ensuring public trust. *Trends Biotechnol.* **22**, 284–285 (2004).
13. Nakamura, Y. The BioBank Japan Project. *Clin. Adv. Hematol. Oncol.* **5**, 696–697 (2007).
14. Hawkins, A. K. Biobanks: importance, implications and opportunities for genetic counselors. *J. Genet. Couns.* **19**, 423–429 (2010).
15. Hewitt, R. E. Biobanking: the foundation of personalized medicine. *Curr. Opin. Oncol.* **23**, 112–119 (2011).

16. Ballantyne, C. Report urges Europe to combine wealth of biobank data. *Nature Med.* **14**, 701 (2008).
17. Founti, P. *et al.* Biobanks and the importance of detailed phenotyping: a case study — the European Glaucoma Society GlaucoGENE project. *Br. J. Ophthalmol.* **93**, 577–581 (2009).
18. Tunis, S. R., Stryer, D. B. & Clancy, C. M. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. *JAMA* **290**, 1624–1632 (2003).
19. Charlson, M. E. & Horwitz, R. I. Applying results of randomised trials to clinical practice: impact of losses before randomisation. *BMJ (Clin. Res. Ed.)* **289**, 1281–1284 (1984).
20. Pablos-Méndez, A., Barr, R. G. & Shea, S. Run-in periods in randomized trials: implications for the application of results in clinical practice. *JAMA* **279**, 222–225 (1998).
21. August, J. Market watch: emerging companion diagnostics for cancer drugs. *Nature Rev. Drug Discov.* **9**, 351 (2010).
22. Brownstein, J. S., Freifeld, C. C., Reis, B. Y. & Mandl, K. D. Surveillance Sans Frontieres: Internet-based emerging infectious disease intelligence and the HealthMap project. *PLoS Med.* **5**, e151 (2008).
23. Kielbasa, A. M., Pomerantz, A. M., Krohn, E. J. & Sullivan, B. F. How does clients' method of payment influence psychologists' diagnostic decisions? *Ethics Behav.* **14**, 187–195 (2004).
24. Tuckson, R. V. *et al.* Policy issues associated with undertaking a new large, U. S. population cohort study of genes, environment, and disease. *Department of Health and Human Services, Washington DC* [online], http://oba.od.nih.gov/oba/sacghs/reports/SACGHS_LPS_report.pdf (2007).
A Landmark report by the US Department of Health and Human Services on the value of large cohort genetic studies of one million or more subjects and the attendant costs and challenges.
25. Murphy, S. *et al.* Instrumenting the health care enterprise for discovery research in the genomic era. *Genome Res.* **19**, 1675–1681 (2009).
A summary of the i2b2 approach to EDGR along with detailed estimates of the financial costs of conducting EDGR.
26. Beasley, D. Remembering recruitment: the impact of proactive subject recruitment planning. *Applied Clinical Trials Online* [online], http://appliedclinicaltrialsonline.findpharma.com/appliedclinicaltrials/Closing+Thought/Remembering-Recruitment/ArticleStandard/Article/detail/527739 (2008).
27. Jha, A. K. *et al.* Use of electronic health records in U. S. hospitals. *N. Engl. J. Med.* **360**, 1628–1638 (2009).
A cautionary survey of the lack of implementation of comprehensive EHRs in the United States.
28. Collins, F. S., Green, E. D., Guttmacher, A. E., Guyer, M. S. & US National Human Genome Research Institute. A vision for the future of genomics research. *Nature* **422**, 835–847 (2003).
29. Ranganathan, M. & Bhopal, R. Exclusion and inclusion of nonwhite ethnic minority groups in 72 North American and European cardiovascular cohort studies. *PLoS Med.* **3**, e44 (2006).
30. Stone, V. E., Mauch, M. Y., Steger, K., Janas, S. F. & Craven, D. E. Race, gender, drug use, and participation in AIDS clinical trials. Lessons from a municipal hospital cohort. *J. Gen. Intern. Med.* **12**, 150–157 (1997).
31. Larson, E. Exclusion of certain groups from clinical research. *Image J. Nurs. Sch.* **26**, 185–190 (1994).
32. Michelen, W., Martinez, J., Lee, A. & Wheeler, D. P. Reducing frequent flyer emergency department visits. *J. Health Care Poor Underserved* **17**, 59–69 (2006).
33. Roby, D. H., Nicholson, G. L. & Kominski, G. F. African Americans in commercial HMOs more likely to delay prescription drugs and use the emergency room. *UCLA Center for Health and Policy Research* [online], http://www.healthpolicy.ucla.edu/pubs/Publication.aspx?pubID=371 (2009).
34. Jones, R., Lin, S., Munsie, J. P., Radigan, M. & Hwang, S. A. Racial/ethnic differences in asthma-related emergency department visits and hospitalizations among children with wheeze in Buffalo, New York. *J. Asthma* **45**, 916–922 (2008).
35. Wolff, J. L., Starfield, B. & Anderson, G. Prevalence, expenditures, and complications of multiple chronic conditions in the elderly. *Arch. Intern. Med.* **162**, 2269–2276 (2002).
36. Brownstein, J. S. *et al.* Rapid identification of myocardial infarction risk associated with diabetes medications using electronic medical records. *Diabetes Care* **33**, 526–531 (2010).
A demonstration of the use of EHR data for timely identification of medically relevant trends; in this case the increased cardiovascular-related mortality associated with a specific oral hypoglycaemic agent.
37. Brownstein, J. S., Sordo, M., Kohane, I. S. & Mandl, K. D. The tell-tale heart: population-based surveillance reveals an association of rofecoxib and celecoxib with myocardial infarction. *PLoS ONE* **2**, e840 (2007).
38. McCarty, C. A. & Wilke, R. A. Biobanking and pharmacogenomics. *Pharmacogenomics* **11**, 637–641 (2010).
39. Kosoy, R. *et al.* Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Hum. Mutat.* **30**, 69–78 (2009).
40. Dumitrescu, L. *et al.* Assessing the accuracy of observer-reported ancestry in a biorepository linked to electronic medical records. *Genet. Med.* **12**, 648–650 (2010).
41. Ioannidis, J. P. A. Non-replication and inconsistency in the genome-wide association setting. *Hum. Hered.* **64**, 203–213 (2007).
42. Gulcher, J. & Stefansson, K. deCODE: A genealogical approach to human genetics in Iceland. *Wiley Online Library* [online], http://onlinelibrary.wiley.com/doi/10.1002/9780470015902.a0006270/abstract (2006).
43. Murphy, S. N., Mendis, M. E., Berkowicz, D. A., Kohane, I. S. & Chueh, H. C. Integration of clinical and genetic data in the i2b2 architecture. *AMIA Annu. Symp. Proc.*, 1040 (2006).
44. Roden, D. M. *et al.* Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin. Pharmacol. Ther.* **84**, 362–369 (2008).
A detailed description of the implementation of EDGR in an institution that is one of the leaders in this domain.
45. Clayton, E. *et al.* Confronting real time ethical, legal, and social issues in the Electronic Medical Records and Genomics (eMERGE) Consortium. *Genet. Med.* **12**, 616–620 (2010).
A useful summary of the various ethical and legal controversies that are entailed by EDGR.
46. Kullo, I. J., Ding, K., Jouni, H., Smith, C. Y. & Chute, C. G. A genome-wide association study of red blood cell traits using the electronic medical record. *PLoS ONE* **5**, e13011 (2010).
47. Ritchie, M. *et al.* Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *Am. J. Hum. Genet.* **86**, 560–572 (2010).
One of the earliest examples of conventional cohort GWA study results being reproduced using EDGR.
48. Kurreeman, F. *et al.* Genetic basis of autoantibody positive and negative rheumatoid arthritis risk in a multi-ethnic cohort derived from electronic health records. *Am. J. Hum. Genet.* **88**, 57–69 (2011).
Early example of extending a GWA study result to other populations using EDGR.
49. Melton, G. B. *et al.* Evaluation of family history information within clinical documents and adequacy of HL7 clinical statement and clinical genomics family history models for its representation: a case report. *J. Am. Med. Inform. Assoc.* **17**, 337–340 (2010).
50. Sager, N., Lyman, M., Bucknall, C., Nhan, N. & Tick, L. J. Natural language processing and the representation of clinical data. *J. Am. Med. Inform. Assoc.* **1**, 142–160 (1994).
51. Lindberg, D. A., Humphreys, B. L. & McCray, A. T. The unified medical language system. *Methods Inf. Med.* **32**, 281–291 (1993).
52. Liao, K. P. *et al.* Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis Care Res. (Hoboken)* **62**, 1120–1127 (2010).
A detailed description of the application of NLP in EDGR and estimates of its accuracy.
53. Uzuner, O., Goldstein, I., Luo, Y. & Kohane, I. Identifying patient smoking status from medical discharge records. *J. Am. Med. Inform. Assoc.* **15**, 14–24 (2008).
54. Jones, R., Pembrey, M., Golding, J. & Herrick, D. The search for genenotype/phenotype associations and the phenome scan. *Paediatr. Perinat. Epidemiol.* **19**, 264–275 (2005).
55. Denny, J. *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics* **26**, 1205–1210 (2010).
An impressive demonstration of the particular capability of EDGR to evaluate one or more SNPs for effect size not only in one phenotype but across all phenotypes available in the EHR.
56. Loscalzo, J., Kohane, I. & Barabasi, A. L. Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. *Mol. Syst. Biol.* **3**, 124 (2007).
57. Murphy, S. N. *et al.* Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J. Am. Med. Inform. Assoc.* **17**, 124–130 (2010).
58. Pulley, J., Clayton, E., Bernard, G., Roden, D. & Masys, D. Principles of human subjects protections applied in an opt-out, de-identified biobank. *Clin. Transl. Sci.* **3**, 42–48 (2010).
59. Mailman, M. D. *et al.* The NCBI dbGaP database of genotypes and phenotypes. *Nature Genet.* **39**, 1181–1186 (2007).
60. McGuire, A. L., Caulfield, T. & Cho, M. K. Research ethics and the challenge of whole-genome sequencing. *Nature Rev. Genet.* **9**, 152–156 (2008).
61. Denny, J. *et al.* Identification of genomic predictors of atrioventricular conduction: using electronic medical records as a tool for genome science. *Circulation* **122**, 2016–2021 (2010).
62. Hagen, S., Richmond, P., Vavrichek, B. & Baumgardner, J. Evidence on the costs and benefits of health information technology. *Congressional Budget Office Washington DC* [online], http://www.cbo.gov/ftpdocs/91xx/doc9168/HealthITTOC.2.1.htm (2008).
A sobering accounting of the costs of the implementation of EHR for clinical care.
63. DiLaura, R. P. Clinical and translational science sustainability: overcoming integration issues between electronic health records (EHR) and clinical research data management systems "separate but equal". *Stud. Health Technol. Inform.* **129**, 137–141 (2007).
64. Scheuner, M. *et al.* Are electronic health records ready for genomic medicine? *Genet. Med.* **11**, 510–517 (2009).
65. Sung, N. *et al.* Central challenges facing the national clinical research enterprise. *JAMA* **289**, 1278–1287 (2003).
66. Uzuner, O., Solti, I. & Cadag, E. Extracting medication information from clinical text. *J. Am. Med. Inform. Assoc.* **17**, 514–518 (2010).
67. Grannis, S. J., Overhage, J. M. & McDonald, C. J. Analysis of identifier performance using a deterministic linkage algorithm. *Proc. AMIA Symp.* **2002**, 305–309 (2002).
68. Finney, J. M., Walker, A. S., Peto, T. E. & Wyllie, D. H. An efficient record linkage scheme using graphical analysis for identifier error detection. *BMC Med. Inform. Decis. Mak.* **11**, 7 (2011).
69. Malin, B. A. An evaluation of the current state of genomic data privacy protection technology and a roadmap for the future. *J. Am. Med. Inform. Assoc.* **12**, 28–34 (2005).
A constructive approach to evaluating data privacy risks once genetic and EHR data become co-mingled.
70. Barnes, D. Texas DNA Showdown. *Mayborn, University of North Texas, Frank W. & Sue Mayborn School of Journalism* [online], http://www.themayborn.com/TexasDNAShowdown.html (2010).
71. Taylor, P. Personal genomes: when consent gets in the way. *Nature* **456**, 32–33 (2008).
72. Kohane, I. S. *et al.* Medicine. Reestablishing the researcher-patient compact. *Science* **316**, 836–837 (2007).
A presentation of an alternative EDGR model, now in its pilot phase, in which patients are also subjects and can control if, when and with what information they are recontacted.
73. Kohane, I. S. & Taylor, P. L. Multidimensional results reporting to participants in genomic studies: getting it right. *Sci. Transl. Med.* **2**, 37cm19 (2010).
74. van der Lei, J. *et al.* The introduction of computer-based patient records in The Netherlands. *Ann. Intern. Med.* **119**, 1036–1041 (1993).
75. Greenhalgh, T. *et al.* Adoption and non-adoption of a shared electronic summary record in England: a mixed-method case study. *BMJ* **340**, c3111 (2010).

76. Jha, A. K., Doolan, D., Grandt, D., Scott, T. & Bates, D. W. The use of health information technology in seven nations. *Int. J. Med. Inform.* **77**, 848–854 (2008).
77. de Lusignan, S., Metsemakers, J. F., Houwink, P., Gunnarsdottir, V. & van der Lei, J. Routinely collected general practice data: goldmines for research? A report of the European Federation for Medical Informatics Primary *Care* Informatics Working Group (EFMI PCIWG) from MIE2006, Maastricht, The Netherlands. *Inform. Prim. Care* **14**, 203–209 (2006).
78. O'Brien, S. Stewardship of human biospecimens, DNA, genotype, and clinical data in the GWAS era. *Annu. Reb. Genomics Hum. Genet.* **10**, 193–209 (2009).
79. Wolf, S. M. *et al.* Managing incidental findings in human subjects research: analysis and recommendations. *J. Law Med. Ethics* **36**, 219–248 (2008).
80. Kohane, I. S., Masys, D. R. & Altman, R. B. The incidentalome: a threat to genomic medicine. *JAMA* **296**, 212–215 (2006).
81. Thorisson, G. A., Muilu, J. & Brookes, A. J. Genotype–phenotype databases: challenges and solutions for the post-genomic era. *Nature Rev. Genet.* **10**, 9–18 (2009).
82. Weber, G. M. *et al.* The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. *J. Am. Med. Inform. Assoc.* **16**, 624–630 (2009).
83. Haspel, R. L. *et al.* A call to action: training pathology residents in genomics and personalized medicine. *Am. J. Clin. Pathol.* **133**, 832–834 (2010).
84. Freifeld, C. C. *et al.* Participatory epidemiology: use of mobile phones for community-based health reporting. *PLoS Med.* **7**, e1000376 (2010).
Going beyond EDGR, an exciting perspective of the use of non-institutional and informal sources of health-related data for population science.
85. Patel, C., Bhattacharya, J., Butte, A. J. & Zhang, B. An environment-wide association study (EWAS) on type 2 diabetes mellitus. *PLoS ONE* **5**, e10746 (2010).
86. Pearson, J. F., Bachireddy, C., Shyamprasad, S., Goldfine, A. B. & Brownstein, J. S. Association between fine particulate matter and diabetes prevalence in the U. S. *Diabetes Care* **33**, 2196–2201 (2010).
87. Pulley, J. M., Brace, M. M., Bernard, G. R. & Masys, D. R. Attitudes and perceptions of patients towards methods of establishing a DNA biobank. *Cell Tissue Bank* **9**, 55–65 (2008).
88. Kohane, I. S. & Altman, R. B. Health-information altruists — a potentially critical resource. *N. Engl. J. Med.* **353**, 2074–2077 (2005).
89. Murphy, J. *et al.* Public expectations for return of results from large-cohort genetic research. *Am. J. Bioeth.* **8**, 36–43 (2008).
90. Kaufman, D., Murphy, J., Scott, J. & Hudson, K. Subjects matter: a survey of public opinions about a large genetic cohort study. *Genet. Med.* **10**, 831–839 (2008).
91. Taylor, P. L. Rules of engagement. *Nature* **450**, 163–164 (2007).
92. Taylor, P. L. Research sharing, ethics and public benefit. *Nature Biotech.* **25**, 398–401 (2007).
93. Fung, K. W., McDonald, C. & Bray, B. E. RxTerms — a drug interface terminology derived from RxNorm. *AMIA Annu. Symp. Proc.* **2008**, 227–231 (2008).
94. Harding, A. & Stuart-Buttle, C. The development and role of the Read Codes. *J. AHIMA* **69**, 34–38 (1998).
95. International statistical classification of diseases and related health problems: 10th revision. *World Health Organization* [online], http://apps.who.int/classifications/apps/icd/icd10online (2007).
96. McCray, A. T. The Unified Medical Language System: The UMLS Semantic Network. *Proc. Annu. Symp. Comput. Appl. Med. Care.* **1989**, 503–507 (1989).
97. Cote, R. A. & Robboy, S. Progress in medical information management: systematized nomenclature of medicine (SNOMED). *JAMA* **243**, 756–762 (1980).

Acknowledgements
The following were kind enough to share insights into their respective EDGR-related efforts: S. Brunak, J. Kim, J. Terdiman, L. Walter, J. Starren, J. Vilo, D. Masys, D. Roden, N. Stimson, L. Bry and S. Churchill. Any errors in communicating these insights are the sole responsibility of the author. The author was supported in part by US National Institutes of Health funding for the US National Centers for Biomedical Computing, U54 LM008748.

Competing interests statement
The author declares no competing financial interests.

FURTHER INFORMATION

Isaac S. Kohane's homepage: http://www.childrenshospital.org/cfapps/research/data_admin/Site113/mainpageS113P0.html
Biobank Japan Project: http://www.src.riken.jp/english/project/person
Danish National Biobank: http://www.ssi.dk/English/RandD/The%20Danish%20National%20Biobank.aspx
Database of Genotypes and Phenotypes (dbGaP): http://www.ncbi.nlm.nih.gov/gap
eMERGE Network: https://www.mc.vanderbilt.edu/victr/dcc/projects/acc/index.php/Main_Page
Estonian Genome Center: http://www.geenivaramu.ee
Genome.gov DNA sequencing costs: http://genome.gov/sequencingcosts
i2b2: https://www.i2b2.org
Kaiser Permanente Research Program on Genes, Environment, and Health (RPGEH): http://www.dor.kaiser.org/external/DORExternal/rpgeh/index.aspx
Marshfield Clinic Personalized Medicine Research Project (PMRP): http://www.marshfieldclinic.org/pmrp
UK Biobank: http://www.ukbiobank.ac.uk
Vanderbilt BioVU: http://dbmi.mc.vanderbilt.edu/research/dnadatabank.html

ALL LINKS ARE ACTIVE IN THE ONLINE PDF