OXFORD

Human Molecular Genetics, 2018, Vol. 27, No. R1 R56–R62

doi: 10.1093/hmg/ddy114
Advance Access Publication Date: 12 April 2018
Invited Review

INVITED REVIEW

The next generation of precision medicine: observational studies, electronic health records, biobanks and continuous monitoring

Benjamin S. Glicksberg[1,2,†], Kipp W. Johnson[1,†] and Joel T. Dudley[1,*]

[1]Institute for Next Generation Healthcare, Department of Genetics and Genomic Sciences, Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, Mount Sinai Health System, New York City, NY 10029, USA and [2]Institute for Computational Health Sciences, University of California San Francisco, San Francisco, CA 94158, USA

*To whom correspondence should be addressed. Tel: 212-731-7065; Fax: 212-731-7099; Email: joel.dudley@mssm.edu

Abstract

Precision medicine can utilize new techniques in order to more effectively translate research findings into clinical practice. In this article, we first explore the limitations of traditional study designs, which stem from (to name a few): massive cost for the assembly of large patient cohorts; non-representative patient data; and the astounding complexity of human biology. Second, we propose that harnessing electronic health records and mobile device biometrics coupled to longitudinal data may prove to be a solution to many of these problems by capturing a 'real world' phenotype. We envision that future biomedical research utilizing more precise approaches to patient care will utilize continuous and longitudinal data sources.

Introduction

Precision medicine consists of medical care more precisely targeted to a patient's phenotypic landscape. Treating to an individual patient's physiological characteristics or symptoms is of course not a new concept—measuring and using variables like physical examination findings, vital signs and laboratory test results have been used by physicians for decades. However, the incorporation of genetic profiling has vastly expanded the number of dimensions through which we can understand an individual's characteristics. Since the early 2000s, genome-wide association studies (GWAS) have been one of the primary tools for the discovery of gene function and genetic mechanisms of disease. GWAS most often involve the creation of case–control cohorts through recruitment of individuals possessing a clinical disease or phenotypic trait alongside control individuals. Statistical requirements generally dictate that these studies enroll large number of patients at great expense; consider the sample sizes for recent complex disease GWAS cohorts in hypertension (475 000) (1); major depressive disorder (460 000) (2); obesity (300 000) (3); type 2 diabetes mellitus (150 000) (4); schizophrenia (110 00) (5); coronary artery disease (100 000) (6); etc. Constructing these massive disease cohorts is an admirable but enormously expensive task—requiring coordination across many academic medical centers, funding agencies and sometimes even different governments. Additionally, even when these cohorts are successfully assembled, there remain important scientific considerations. For example, patients who consent to participate in clinical studies are sometimes systematically different from other patients (7). For example, even simply volunteering to participate in clinical research has been shown to be a powerful predictor of 5-year survival in

[†]Contributed equally to this work.

Received: March 13, 2018. **Revised:** March 23, 2018. **Accepted:** March 27, 2018

patients with heart failure (hazard ratio = 0.3) (8). In contrast, although they are still imperfect, observational studies tend to include patients who are more representative of the broader population than clinical trials (9).

Observational Biobanks Coupled to EHR Data: A Natural Alternative

In this section, we expound upon the limitations of traditional study designs in the post-GWAS era (10) and how they can be addressed through the use of biobanks linked to electronic health records (EHR) and smart devices with observational study designs (Fig. 1). Briefly, here we use the term 'biobanks' to refer to collections or repositories of patient samples (generally whole blood) which have been genotyped by either microarray, exome (WES)-, or whole-genome sequencing (WGS). These results are then deposited to a database for retrospective analysis by researchers. There are many impressive initiatives to collect health and genomic information from population-scale cohorts such as the NIH's All of Us Research Program (https://allofus.nih.gov/) and U.S. Department of Veteran Affair's Million Veteran Program (MVP; https://www.research.va.gov/mvp/), as well as others that we list in Table 1.

First, many traditional prospective study designs are typically limited to the use of trait measurements collected during a trial period. In the case of a mostly static phenotype such as adult height, this is likely sufficient. However, many complex human traits and diseases are dynamic in nature and have strong temporal components that should not be ignored in the analysis of genetic associations. To continue with our example case of hypertension: since blood pressure (BP) measurements exhibit trends at timescales of hours, days, years, etc., it follows that incorporating this dynamism into analysis can improve discovery. One natural and convenient data source for longitudinal trends are those found in EHR, which often contain years of data for routine clinical measurements like BP. In fact, Hoffman *et al.* employed this strategy when they compared using single measurements versus incorporating multiple measurements from EHR in a recent GWAS (11). This study found that the use of longitudinal measurements doubled the variance explained by SNPs for systolic and diastolic blood pressure, and tripled the variance explained in pulse pressure (SBP minus DBP). They also discovered up to (in the case of pulse pressure) 23 new SNPs through the incorporation of multiple measurements.

Second, the use of observational data acquired from hospital-based systems has another advantage. Because these data are acquired during the process of provision of care, it is essentially 'free"'for research use after accounting for the administrative and personnel costs required to structure and anonymize EHR data for analysis. Large academic medical centers and health systems routinely may have records for up to millions of patients, while the cost of building such disease-specific cohorts may be impossible for an individual investigator. A related, subtle consideration may be that results in the EHR can be more representative of actual clinical practice than those collected during a clinical trial. For example, the recent SPRINT clinical trial for the effects of intensive antihypertensive therapy (12) was criticized for using a non-standard BP measurement technique that was thought not to be indicative of the BPs measured in the clinic on a day-to-day basis. We were able to use a causal inference method applied to records from the Mount Sinai Hospital to validate conclusions reached in SPRINT against real-word BP measurements (13). SPRINT, as a large and complex multicenter trial, likely cost millions of dollars to perform, while our confirmatory analysis required only a few weeks of labor. While we did not incorporate genetic analysis

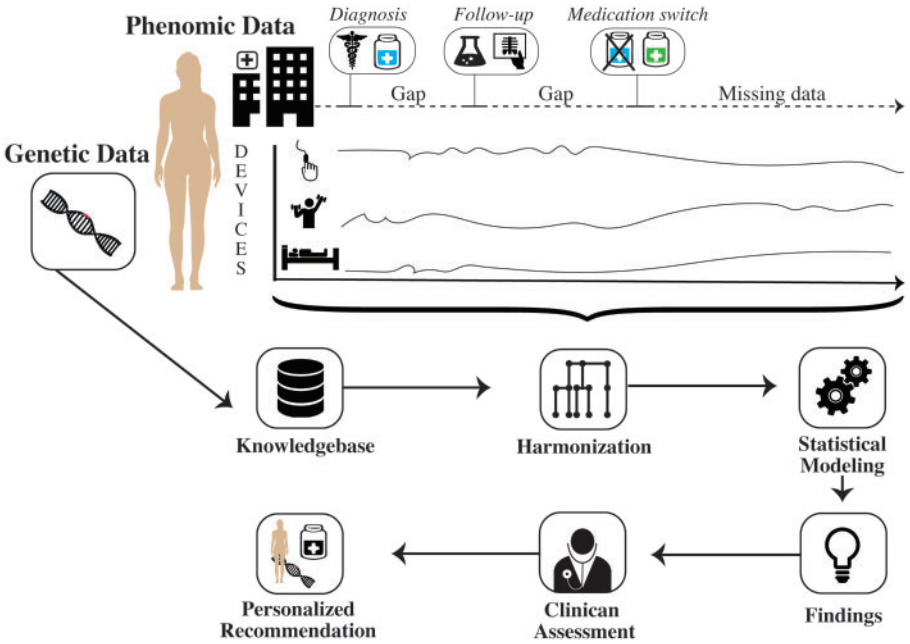

Figure 1. Workflow for harmonizing multi-omic and health data for personalized medicine. These data modalities have limitations in their applicability for healthcare when taken in isolation. For instance, genomic data often are most informative in the context of a phenotype. EHR data are restricted to patient encounter visits and therefore have many gaps in phenotype information. When these data types are harmonized and assessed in relation to one another, a more powerful 'quantified self' is produced which allows for capturing a more comprehensive and overall reflective health state. In this scenario, a patient's phenotype data from their EHR encounters are supplemented by information tracked by various devices, such as blood pressure, activity and sleep monitors, which can help physicians tailor treatment based on her health state.

Table 1. Notable biobanks with linked EHR

Institution	Data types	Approximate sample size	Access	Link
UK Biobank	Genotype	500 000	Application	http://www.ukbiobank.ac.uk/
Genomics England	WGS	100 000[a]	Application[a]	https://www.genomicsengland.co.uk/
UK10K	WGS/WES	10 000	Application	https://www.uk10k.org/
discovEHR	WES	50 000	Private	http://www.discovehrshare.com/
eMERGE	Genotype/WES/WGS	100 000	Private	https://emerge.mc.vanderbilt.edu/
deCODE	WGS	500 000	Private	https://www.decode.com/

WGS, whole genome sequencing; WES, whole exome sequencing.
[a]Planned.

into our trial since SPRINT also did not, the same idea holds true for the analyses of biobanked genetic data. Finally, we also note that EHR-based analyses allow for serendipitous findings. In contrast, for example in a prospective trial, investigators cannot generally assess associations between seemingly unlinked variables if they were not collected from the beginning of the study.

Best Practices for Utilizing Electronic Health Records Data

Because the purposes of EHR are foremost to enable the provision of medical care and for hospital billing, the data contained within these systems are typically not in a format readily suitable for research. In particular, EHR data structures often result from years of haphazard integration across various clinical units, software vendors, informatics systems, etc. and are thus typically quite complex. Even when the data is readily obtainable, it is well known that there are a plethora of issues revolving around usability of EHR for research at both the data and operational level (14). For instance, EHR data are often not quality controlled and may contain substantial errors or high rates of redundancy and missingness (15). It is also important to note that there are also challenges pertaining to use of retrospective cohorts, such as selection and confounding biases (16). Furthermore, the formats these data are recorded in usually cannot be easily linked to other biomedical or bioinformatic databases and resources.

Because of these myriad difficulties, much effort has been exhausted on techniques to harmonize these data in a standardized fashion which can be connected to other biomedical resources. Interoperability issues can be partially overcome from a schema structure that is normalized, like the particularly notable efforts from the Observational Health Data Sciences and Informatics (OHDSI) consortium (17). OHDSI was able to use its OMOP data structure to integrate health records for more than 250 million patients from 11 sites to characterize treatment patterns for major depressive disorder, hypertension and type 2 diabetes mellitus.

Even with cleaned and harmonized data, there still remains the issue of precise and accurate disease phenotyping. Specifically, the data that is recorded in EHR are not necessarily reflective of the criteria by which a physician makes a diagnosis. When a patient is seen in a clinical visit, their health state is broken down into structured (e.g. ICD codes) and unstructured (e.g. notes) data that are pre-defined by available fields. While some of the insight of the physician is recorded as impressions in clinical notes, it is not a replacement for this interaction. Accordingly, when EHR are used for research, we must reconstruct a patient's health state from this imperfect representation. With no prospective study criteria or quality control, the question then becomes: 'How do we accurately define a disease using EHR data?' Fortunately, electronic phenotyping algorithms have been introduced as a tool to systematize the process in order to promote better accuracy and reproducibility. These algorithms are developed by experts that indicate the EHR fields and data that should be used for inclusion and exclusion criteria per disease. The performance of these algorithms are typically evaluated through manual chart review. The Phenotype Knowledgebase (PheKB; https://phekb.org/) collects and houses these algorithms for research use. Recently, there have been promising efforts to attempt to automate this process through machine learning. For example, Blecker *et al.* recently compared the utility of five different methods (two clinical heuristic; one logistic regression model from structured clinical variables; and two machine learning approaches using both structured and unstructured EHR-derived data) to electronically phenotype heart failure patients (18). They found that the machine-learning algorithms outperformed more traditional strategies for phenotyping. Similarly, we have implemented a machine-learning approach called word2vec with promising preliminary results for automated phenotyping in five different diseases (19). In this study, we compared cohorts derived from this automated process to those processed from 'gold standard' electronic phenotyping algorithms from PheKB as a metric of performance. Our automated approach had promising performances for some diseases, such as sickle cell disease, but less so for others, like dementia. We attribute this discrepancy to such factors as complexity of criteria and underlying difficulty in representing diseases using structured EHR data. These findings indicate that completely automated procedures may not be ready for replacing manual phenotyping strategies but hopefully will augment them in the near future. Lastly, natural language processing of physician notes in particular may finally be enabled by the advent of deep learning models (20,21).

Recent Advances in Precision Medicine from EHR-coupled Biobanks

Verifying true variant-disease associations

EHR-coupled biobanks can be used to assess the validity of previously reported variant-phenotype associations. Consider the excellent study by Haggerty *et al.* which examined the effect of database-annotated pathogenic and likely pathogenic variants associated with arrhythmogenic right ventricular cardiomyopathy (ARVC) (22). The authors interrogated the sequenced exomes for 30 000 individuals and found 18 individuals carrying a variant supposedly causal for ARVC; however, none of the individuals were diagnosed with ARVC (which presents with a striking phenotype) and can be diagnosed in most cases in a

straightforward fashion with an electocardiogram. Upon further review, most of the patients bearing a so-called pathogenic mutation had no corresponding phenotype. Similarly, another powerful study by Van Driest *et al.* interrogated the presence of arrhythmia-associated variants with phenotypes documented in the EHR (23). Similarly, they found that the individuals bearing many supposedly pathogenic mutations were in fact not statistically enriched for diagnosis of any arrhythmia, nor did they possess notable arrhythmia-related phenotypes documented in the EHR.

Drug target discovery and drug response

EHR-linked biobanks have proven themselves to be an incredible resource for the discovery of novel drug targets (24–27). In an example of successful academic–industry partnership, Dewey *et al.* from Regeneron Pharmaceuticals used the exome sequencing results connected to the EHR from 58 000 patients collected at the Geisinger Health System in rural Pennsylvania. The authors discovered that loss-of-function mutations in *ANGPTL3* were associated with the development of coronary artery disease (28). They used this knowledge to develop both an inactivating monoclonal antibody to the protein which shows promise for disease prevention. Similarly, Ionis pharmaceuticals has since also demonstrated that antisense oligotides to *ANGPTL3* are also protective against cardiovascular disease in humans (29).

EHR-linked biobanks have also been used in the field of pharmacogenomics (30,31) to discover and interrogate specific genetic variants that modulate or mediate drug efficacy, response or tolerance, such as those that affect metabolism. The eMERGE and Pharmacogenomics Research Network (eMERGE-PGx) found out that 96.2% of 5000 clinical subjects had at least one actionable variant for 82 pharmacogenes (32).

Probing disease biology through loss-of-function analysis

Inferring gene function from EHR analyses is not an easy task due to complex properties of the association, such as penetrance, expressivity, mode of inheritance, among others. One strategy seeks to bypass some of these issues by focusing on rare, large-effect mutations that one can hypothesize would be easier to detect from its impact on a phenotype. Loss-of-function (LOF) variants, or protein-truncating variants (PTV), within a gene, such as a premature stop gain, are predicted to reduce or abolish its expression. As such, these can serve as proxies for understanding how the system operates without it.

Recently, Dewey *et al.* used a large cohort to demonstrate that individuals carry a median of 21 LoF mutations (33). Harnessing the information contained in linked EHR (with a median of 14 years of follow-up), the authors were able to find a number of new associations between genes and phenotypes such as erythrocytosis. Furthermore, using a validated set of 76 truly clinically actionable genes (as opposed to the thousands of allegedly pathogenic GWAS-originated variants), they found that 3.5% of the population carried a variant and that the associated disease could be validated in 65% of the carrier's medical records. In the same population, Abul-Husn *et al.* further explored the impact of known damaging mutations in genes associated with familial hypercholesterolemia (*LDLR*, *APOB* and *PCSK9*) and found that genetic diagnosis could predict adverse cardiovascular outcomes (34). In a separate study, sequencing of lipoprotein lipase (a gene associated with hypertriglyceridemia) in the same cohort was demonstrated to be of clinical utility (35). In the same cohort it has also been demonstrated that biobank sequencing data is of reliable quality for clinical diagnosis of BRCA1/2 mutations causal for breast cancer, which means that biobanks developed primarily for research can also be clinically useful in a straightforward fashion by enabling increased access to genetic testing (36).

In another fascinating example of the discoveries empowered by EHR-linked biobanks, Belbin *et al.* (37) demonstrated that variants in the gene *COL27A1* are causal for a rare phenotype for short stature, and furthermore that individuals with these variants are actually present at much higher rates than previously assumed (>2%) in individuals of Puerto Rican ancestry. Since minority populations are drastically underrepresented in much of biomedical research, the use of databases where many individuals are ethnic minorities can even be seen as an inclusive policy.

Non-conventional and multimodal analyses enabled by EHR-linked biobanks

Because of the richness of data contained within EHR and the ability to connect EHR to outside data sources, EHR-linked biobanks provide an excellent opportunity for unconventional analyses which can result in surprising findings. For example, we have used deep learning to analyze EHR for over 300 000 patients at our hospital to examine the discrepancy between chronological and physiological aging (38). We found a number of factors related to discrepancies between chronological age and physiological age. For instance, in individuals with a predicted (physiological) age significantly older than their chronological age, we found increased prevalence of hypertension, increased chronic inflammation, poor nutritional status, decreased kidney function and signs of liver damage among others. Alternatively, in individuals with a physiological age significantly younger than their actual age, we found lower risks for hypertension and hyperlipidemia and healthier kidney and liver functions. In an interesting example of an environment-wide association study (39), a group of researchers in England conducted a biobank-enabled study examined the neighborhood prevalence of fast-food establishments and physical recreation facilities that were associated with rates of obesity (40).

The relationship between disease and other biological quantities, or -omics, such as RNA expression levels can also be probed in EHR-linked biobanks (41,42). This approach can further elucidate potential pathways for how risk variants functionally contribute to disease pathogenesis, taking into account regulatory mechanisms unique to different functional units of human biology, for example at the tissue or organ level. Franzén, Ermel, Cohain *et al.* utilized the Stockholm-Tartu Atherosclerosis Reverse Networks Engineering Task study (STARNET) biobank, a collection of 600 patients with RNA data for seven cardiometabolic tissues, to investigate functional pathways by which coronary artery disease variants contribute to disease incidence and progression (43). They identified the gene-regulatory mechanisms for these risk loci, some of which were tissue-specific, through eQTL analyses. Finally, biobanks have been significantly useful for probing biological questions related to race, ethnicity, and genetic ancestry (44,45).

Digital medicine

While EHR data undoubtedly enables higher fidelity analysis of human disease, the inclusion of other clinically relevant data

facilitates even further benefit. Despite its many benefits compared to other methods for data acquisition, EHR data is limited to information obtained during a clinical visit, and as such, there are almost always temporal gaps (Fig. 1). Advances in devices, computing power, cloud storage and data transmission capabilities have ushered in the age of digital health, where metrics reflective of health state can be continually collected and analyzed with a variety of devices. It should be noted that continuous measurement of physiological traits is not a new concept—data obtained from methods such as 24 h ambulatory BP measurement have been used for better outcome prediction or to generate new indicators for effectiveness of antihypertensive therapy (46). New Internet-of-Things (IoT) devices, which can receive, display and transmit data in real time and can allow 'high-definition medicine' should further enable advances like this (47). While smartphones are the devices of the most obvious utility, there are increasing amounts of other 'smart' devices that collect data that are reflective of, or contribute to, our health state, such as: activity monitors [e.g. FitBit (https://www.fitbit.com/)], digital scales, electronic air quality monitors [e.g. FooBot (https://foobot.io/)] and digital sleep monitors among countless others.

These devices are critically important in pushing forward precision medicine by capturing pieces of information that can give clinicians a more complete view of health state, rather than piecing together snap shots of information from current and past visits. A recent study by Li, Dunn, Salins *et al.* epitomizes the utility of continual tracking for digital health where they were able to distinguish physiological disturbances from personal baselines to identify the onset of Lyme disease and inflammatory responses (48). Advancing this even further, Muse *et al.* envision a 'smart medical home' which consists of many IoT devices that can be used to both capture data and provide healthcare-related reminders, such as a smart mirror that displays which medications to take (49). One clinical area that could particularly benefit from digital medicine is mental health (50), which suffers from difficulties in accurately measuring emotional state that are paramount toward measuring treatment efficacy. The metrics that these IoT devices can measure relevant information, such as sleep quality, mood abstracted from facial recognition, to social media interactions, that are both reflective of everyday life and at a higher frequency than scheduled check-up or therapy appointments.

Ultimately, these capabilities may allow the digitization of clinical trials (51), where continual collection of participant data and remote monitoring can provide higher quality information than is collected by traditional methods (i.e. automated weight measurements versus self-reported logs) as well as data that was unable to be continually tracked. There have been a number of Mobile Health (mHealth) studies that have followed this model and have helped elucidate patterns underlying diseases, such as the Asthma Mobile Health Study (52), which tracks multiple dimensions of data that could contribute to asthma symptoms such as surveys. With the rise of opportunities that digital medicine affords also comes issues of privacy, security and ownership of data (53). While it is beyond the scope of the current article, it is a critical and on-going issue that deserves attention (54,55).

Conclusion

Solving how biomedicine can incorporate genomics into clinical practice is a vitally important problem that can be further enabled by the incorporation of strategies proposed in this paper. As suggested by Greene and Loscalzo in a wonderful perspective piece, medicine needs to 'put the patient back together' (55). If we want to cure human disease, there is ultimately no better avenue—and we believe that we have laid out compelling alternatives in this article which may be employed to enable meaningful research. To state these ideas simply: First, obtain as much relevant data as is possible to reach your conclusions. This will often require the use of linked EHR instead of the limited data collected during prospective trials. Similarly, incorporate longitudinal data when it is available. Second, ensure that data used as inputs is clinically representative of real patients, and not the sometimes unrepresentative world of clinical trials. Again, one way to increase this probability is to use data that is literally collected during the provision of clinical care. Third, we need to continue to look beyond what information is collected in the genome or EHR. Emerging sources of data such as internet-connected devices have enormous promise for future of medicine and should be pursued with vigor.

Conflict of Interest statement. None declared.

References

1. Kraja, A.T., Cook, J.P., Warren, H.R., Surendran, P., Liu, C., Evangelou, E., Manning, A.K., Grarup, N., Drenos, F., Sim, X. *et al.* (2017) New blood pressure-associated loci identified in meta-analyses of 475 000 individuals. *Circ. Cardiovasc. Genet.*, **10**, 1–8.
2. Wray, N.R. and Sullivan, P.F. (2017) Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *bioRxiv*, in press.
3. Locke, A.E., Kahali, B., Berndt, S.I., Justice, A.E., Pers, T.H., Day, F.R., Powell, C., Vedantam, S., Buchkovich, M.L., Yang, J. *et al.* (2015) Genetic studies of body mass index yield new insights for obesity biology. *Nature*, **518**, 197–206.
4. Scott, R.A., Scott, L.J., Magi, R., Marullo, L., Gaulton, K.J., Kaakinen, M., Pervjakova, N., Pers, T.H., Johnson, A.D., Eicher, J.D. *et al.* (2017) An expanded genome-wide association study of type 2 diabetes in Europeans. *Diabetes*, **66**, 2888–2902.
5. Schizophrenia Working Group of the Psychiatric Genomics Consortium. (2014) Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, **511**, 421–427.
6. Myocardial Infarction, G., Investigators, C.A.E.C., Stitziel, N.O., Stirrups, K.E., Masca, N.G., Erdmann, J., Ferrario, P.G., Konig, I.R., Weeke, P.E., Webb, T.R. *et al.* (2016) Coding variation in ANGPTL4, LPL, and SVEP1 and the risk of coronary disease. *N. Engl. J .Med.*, **374**, 1134–1144.
7. Weng, C., Li, Y., Ryan, P., Zhang, Y., Liu, F., Gao, J., Bigger, J.T. and Hripcsak, G. (2014) A distribution-based method for assessing the differences between clinical trial target populations and patient populations in electronic health records. *Appl. Clin. Inform.*, **05**, 463–479.
8. Clark, A.L., Lammiman, M.J., Goode, K. and Cleland, J.G. (2009) Is taking part in clinical trials good for your health? A cohort study. *Eur. J. Heart Fail.*, **11**, 1078–1083.
9. He, Z., Ryan, P., Hoxha, J., Wang, S., Carini, S., Sim, I. and Weng, C. (2016) Multivariate analysis of the population representativeness of related clinical studies. *J. Biomed. Inform.*, **60**, 66–76.
10. Wijmenga, C. and Zhernakova, A. (2018) The importance of cohort studies in the post-GWAS era. *Nat. Genet.*, **50**, 322–328.
11. Hoffmann, T.J., Ehret, G.B., Nandakumar, P., Ranatunga, D., Schaefer, C., Kwok, P.Y., Iribarren, C., Chakravarti, A. and

Risch, N. (2017) Genome-wide association analyses using electronic health records identify new loci influencing blood pressure variation. *Nat. Genet.*, **49**, 54–64.
12. Group, S.R., Wright, J.T., Jr., Williamson, J.D., Whelton, P.K., Snyder, J.K., Sink, K.M., Rocco, M.V., Reboussin, D.M., Rahman, M., Oparil, S. *et al.* (2015) A randomized trial of intensive versus standard blood-pressure control. *N. Engl. J. Med.*, **373**, 2103–2116.
13. Johnson, K.W., Glicksberg, B.S., Hodos, R.A., Shameer, K. and Dudley, J.T. (2018) Causal inference on electronic health records to assess blood pressure treatment targets: an application of the parametric g formula. *Pac. Symp. Biocomput.*, **23**, 180–191.
14. Tang, P.C., Ash, J.S., Bates, D.W., Overhage, J.M. and Sands, D.Z. (2006) Personal health records: definitions, benefits, and strategies for overcoming barriers to adoption. *J. Am. Med. Inform. Assoc.*, **13**, 121–126.
15. Botsis, T., Hartvigsen, G., Chen, F. and Weng, C. (2010) Secondary use of EHR: data quality issues and informatics opportunities. *AMIA Jt. Summits Transl. Sci. Proc.*, **2010**, 1–5.
16. Haneuse, S. and Daniels, M. (2016) A general framework for considering selection bias in EHR-based studies: what data are observed and why? *EGEMS (Wash. DC)*, **4**, 16.
17. Hripcsak, G., Duke, J.D., Shah, N.H., Reich, C.G., Huser, V., Schuemie, M.J., Suchard, M.A., Park, R.W., Wong, I.C., Rijnbeek, P.R. *et al.* (2015) Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Stud .Health Technol. Inform.*, **216**, 574–578.
18. Blecker, S., Katz, S.D., Horwitz, L.I., Kuperman, G., Park, H., Gold, A. and Sontag, D. (2016) Comparison of approaches for heart failure case identification from electronic health record data. *JAMA Cardiol.*, **1**, 1014–1020.
19. Glicksberg, B.S., Miotto, R., Johnson, K.W., Shameer, K., Li, L., Chen, R. and Dudley, J.T. (2018) Automated disease cohort selection using word embeddings from electronic health records. *Pac. Symp. Biocomput.*, **23**, 145–156.
20. Miotto, R., Wang, F., Wang, S., Jiang, X. and Dudley, J.T. (2017) Deep learning for healthcare: review, opportunities and challenges. *Brief Bioinform.*, doi:10.1093/bib/bbx044.
21. Liao, K.P., Cai, T., Savova, G.K., Murphy, S.N., Karlson, E.W., Ananthakrishnan, A.N., Gainer, V.S., Shaw, S.Y., Xia, Z., Szolovits, P. *et al.* (2015) Development of phenotype algorithms using electronic medical records and incorporating natural language processing. *BMJ*, **350**, h1885.
22. Haggerty, C.M., James, C.A., Calkins, H., Tichnell, C., Leader, J.B., Hartzel, D.N., Nevius, C.D., Pendergrass, S.A., Person, T.N., Schwartz, M. *et al.* (2017) Electronic health record phenotype in subjects with genetic variants associated with arrhythmogenic right ventricular cardiomyopathy: a study of 30,716 subjects with exome sequencing. *Genet. Med.*, **19**, 1245–1252.
23. Van Driest, S.L., Wells, Q.S., Stallings, S., Bush, W.S., Gordon, A., Nickerson, D.A., Kim, J.H., Crosslin, D.R., Jarvik, G.P., Carrell, D.S. *et al.* (2016) Association of arrhythmia-related genetic variants with phenotypes documented in electronic medical records. *JAMA*, **315**, 47–57.
24. Yao, L., Zhang, Y., Li, Y., Sanseau, P. and Agarwal, P. (2011) Electronic health records: implications for drug discovery. *Drug Discov. Today*, **16**, 594–599.
25. Shameer, K., Glicksberg, B.S., Hodos, R., Johnson, K.W., Badgeley, M.A., Readhead, B., Tomlinson, M.S., O'Connor, T., Miotto, R., Kidd, B.A. *et al.* (2017) Systematic analyses of drugs and disease indications in RepurposeDB reveal pharmacological, biological and epidemiological factors influencing drug repositioning. *Brief Bioinform.*, doi:10.1093/bib/bbw136.
26. Shameer, K., Johnson, K.W., Glicksberg, B.S., Dudley, J.T. and Sengupta, P.P. (2018) Machine learning in cardiovascular medicine: are we there yet? *Heart*, doi:10.1136/heartjnl-2017-311198.
27. Johnson, K.W., Shameer, K., Glicksberg, B.S., Readhead, B., Sengupta, P.P., Björkegren, J.L.M., Kovacic, J.C. and Dudley, J.T. (2017) Enabling precision cardiology through multiscale biology and systems medicine. *JACC Basic Transl. Sci.*, **2**, 311–327.
28. Dewey, F.E., Gusarova, V., Dunbar, R.L., O'Dushlaine, C., Schurmann, C., Gottesman, O., McCarthy, S., Van Hout, C.V., Bruse, S., Dansky, H.M. *et al.* (2017) Genetic and pharmacologic inactivation of ANGPTL3 and cardiovascular disease. *N. Engl. J. Med.*, **377**, 211–221.
29. Graham, M.J., Lee, R.G., Brandt, T.A., Tai, L.J., Fu, W., Peralta, R., Yu, R., Hurh, E., Paz, E., McEvoy, B.W. *et al.* (2017) Cardiovascular and metabolic effects of ANGPTL3 antisense oligonucleotides. *N. Engl. J. Med.*, **377**, 222–232.
30. Scott, S.A., Owusu Obeng, A., Botton, M.R., Yang, Y., Scott, E.R., Ellis, S.B., Wallsten, R., Kaszemacher, T., Zhou, X., Chen, R. *et al.* (2017) Institutional profile: translational pharmacogenomics at the Icahn School of Medicine at Mount Sinai. *Pharmacogenomics*, **18**, 1381–1386.
31. Wilke, R.A., Xu, H., Denny, J.C., Roden, D.M., Krauss, R.M., McCarty, C.A., Davis, R.L., Skaar, T., Lamba, J. and Savova, G. (2011) The emerging role of electronic medical records in pharmacogenomics. *Clin. Pharmacol. Ther.*, **89**, 379–386.
32. Bush, W.S., Crosslin, D.R., Owusu-Obeng, A., Wallace, J., Almoguera, B., Basford, M.A., Bielinski, S.J., Carrell, D.S., Connolly, J.J., Crawford, D. *et al.* (2016) Genetic variation among 82 pharmacogenes: the PGRNseq data from the eMERGE network. *Clin. Pharmacol. Ther.*, **100**, 160–169.
33. Dewey, F.E., Murray, M.F., Overton, J.D., Habegger, L., Leader, J.B., Fetterolf, S.N., O'Dushlaine, C., Van Hout, C.V., Staples, J., Gonzaga-Jauregui, C. *et al.* (2016) Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science*, **354**, aaf6814.
34. Abul-Husn, N.S., Manickam, K., Jones, L.K., Wright, E.A., Hartzel, D.N., Gonzaga-Jauregui, C., O'Dushlaine, C., Leader, J.B., Lester Kirchner, H., Lindbuchler, D.M. *et al.* (2016) Genetic identification of familial hypercholesterolemia within a single U.S. health care system. *Science*, **354**, aaf7000.
35. Khera, A.V., Won, H.-H., Peloso, G.M., O'Dushlaine, C., Liu, D., Stitziel, N.O., Natarajan, P., Nomura, A., Emdin, C.A., Gupta, N. *et al.* (2017) Association of rare and common variation in the lipoprotein lipase gene with coronary artery disease. *JAMA*, **317**, 937–946.
36. Buchanan, A.H., Manickam, K., Meyer, M.N., Wagner, J.K., Hallquist, M.L.G., Williams, J.L., Rahm, A.K., Williams, M.S., Chen, Z.E., Shah, C.K. *et al.* (2017) Early cancer diagnoses through BRCA1/2 screening of unselected adult biobank participants. *Genet. Med.*, doi:10.1038/gim.2017.145.
37. Belbin, G.M., Odgis, J., Sorokin, E.P., Yee, M.C., Kohli, S., Glicksberg, B.S., Gignoux, C.R., Wojcik, G.L., Van Vleck, T., Jeff, J.M. *et al.* (2017) Genetic identification of a common collagen disease in puerto ricans via identity-by-descent mapping in a health system. *eLife*, doi:10.7554/eLife.25060.
38. Wang, Z., Li, L., Glicksberg, B.S., Israel, A., Dudley, J.T. and Ma'ayan, A. (2017) Predicting age by mining electronic medical records with deep learning characterizes differences

between chronological and physiological age. *J. Biomed. Inform.*, **76**, 59–68.
39. Patel, C.J., Chen, R., Kodama, K., Ioannidis, J.P. and Butte, A.J. (2013) Systematic identification of interaction effects between genome- and environment-wide associations in type 2 diabetes mellitus. *Hum. Genet.*, **132**, 495–508.
40. Mason, K.E., Pearce, N. and Cummins, S. (2018) Associations between fast food and physical activity environments and adiposity in mid-life: cross-sectional, observational evidence from UK Biobank. *Lancet Public Health*, **3**, e24–e33.
41. Karczewski, K.J. and Snyder, M.P. (2018) Integrative omics for health and disease. *Nat. Rev. Genet.*, doi:10.1038/nrg.2018.4.
42. Butte, A.J. (2017) Big data opens a window onto wellness. *Nat. Biotechnol.*, **35**, 720–721.
43. Franzen, O., Ermel, R., Cohain, A., Akers, N.K., Di Narzo, A., Talukdar, H.A., Foroughi-Asl, H., Giambartolomei, C., Fullard, J.F., Sukhavasi, K. *et al.* (2016) Cardiometabolic risk loci share downstream cis- and trans-gene regulation across tissues and diseases. *Science*, **353**, 827–830.
44. Glicksberg, B.S., Li, L., Badgeley, M.A., Shameer, K., Kosoy, R., Beckmann, N.D., Pho, N., Hakenberg, J., Ma, M., Ayers, K.L. *et al.* (2016) Comparative analyses of population-scale phenomic data in electronic medical records reveal race-specific disease networks. *Bioinformatics*, **32**, i101–i110.
45. Nadkarni, G.N., Galarneau, G., Ellis, S.B., Nadukuru, R., Zhang, J., Scott, S.A., Schurmann, C., Li, R., Rasmussen-Torvik, L.J., Kho, A.N. *et al.* (2017) Apolipoprotein L1 variants and blood pressure traits in African Americans. *J. Am .Coll. Cardiol.*, **69**, 1564–1574.
46. Redon, J. (2013) The importance of 24-hour ambulatory blood pressure monitoring in patients at risk of cardiovascular events. *High Blood Press Cardiovasc. Prev.*, **20**, 13–18.
47. Torkamani, A., Andersen, K.G., Steinhubl, S.R. and Topol, E.J. (2017) High-definition medicine. *Cell*, **170**, 828–843.
48. Li, X., Dunn, J., Salins, D., Zhou, G., Zhou, W., Schussler-Fiorenza Rose, S.M., Perelman, D., Colbert, E., Runge, R., Rego, S. *et al.* (2017) Digital health: tracking physiomes and activity using wearable biosensors reveals useful health-related information. *PLoS Biol.*, **15**, e2001402.
49. Muse, E.D., Barrett, P.M., Steinhubl, S.R. and Topol, E.J. (2017) Towards a smart medical home. *Lancet*, **389**, 358.
50. Barrett, P.M., Steinhubl, S.R., Muse, E.D. and Topol, E.J. (2017) Digitising the mind. *Lancet*, **389**, 1877.
51. Steinhubl, S.R., McGovern, P., Dylan, J. and Topol, E.J. (2017) The digitised clinical trial. *Lancet*, **390**, 2135.
52. Chan, Y.Y., Wang, P., Rogers, L., Tignor, N., Zweig, M., Hershman, S.G., Genes, N., Scott, E.R., Krock, E., Badgeley, M. *et al.* (2017) The Asthma Mobile Health Study, a large-scale clinical observational study using ResearchKit. *Nat. Biotechnol.*, **35**, 354–362.
53. Mikk, K.A., Sleeper, H.A. and Topol, E.J. (2017) The pathway to patient data ownership and better health. *JAMA*, **318**, 1433–1434.
54. Islam, S.M.R., Kwak, D., Kabir, M.H., Hossain, M. and Kwak, K.S. (2015) The Internet of Things for health care: a comprehensive survey. *IEEE Access.*, **3**, 678–708.
55. Tarouco, L.M.R., Bertholdo, L.M., Granville, L.Z., Arbiza, L.M.R., Carbone, F., Marotta, M., and Santanna, JJCD. (2012) Internet of Things in healthcare: Interoperability and security issues. *2012 IEEE International Conference on Communications (ICC)*, Ottawa, ON, Canada, pp. 6121–6125.
56. Greene, J.A. and Loscalzo, J. (2017) Putting the patient back together - social medicine, network medicine, and the limits of reductionism. *N. Engl. J. Med.*, **377**, 2493–2499.