TRANSLATIONAL GENETICS

Mining electronic health records: towards better research applications and clinical care

Peter B. Jensen[1], Lars J. Jensen[1] and Søren Brunak[1,2]

Abstract | Clinical data describing the phenotypes and treatment of patients represents an underused data source that has much greater research potential than is currently realized. Mining of electronic health records (EHRs) has the potential for establishing new patient-stratification principles and for revealing unknown disease correlations. Integrating EHR data with genetic data will also give a finer understanding of genotype–phenotype relationships. However, a broad range of ethical, legal and technical reasons currently hinder the systematic deposition of these data in EHRs and their mining. Here, we consider the potential for furthering medical research and clinical care using EHR data and the challenges that must be overcome before this is a reality.

Clinical decision support (CDS). Software systems providing support for decision making to physicians through the application of health knowledge and logical rules to patient data.

Biobanks
Central repositories of biological material that are mainly used for research. They facilitate the re-use of collected samples in different research projects.

[1]*NNF Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark.*
[2]*Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark.*
Correspondence to S.B.
e-mail: brunak@cbs.dtu.dk
doi:10.1038/nrg3208
Published online 2 May 2012

Information technology has transformed the way health care is carried out and documented. Presently, the practice of health care generates, exchanges and stores huge amounts of patient-specific information. In addition to the traditional clinical narrative, databases in modern health centres automatically capture structured data relating to all aspects of care, including diagnosis, medication, laboratory test results and radiological imaging data.

This transformation holds great promise for the individual patient as richer information, coupled with clinical decision support (CDS) systems, becomes readily available at the bedside to support informed decision making and to improve patient safety[1,2]. From a research perspective, integrated patient data constitute a computable collection of fine-grained longitudinal phenotypic profiles, facilitating cohort-wide investigations and knowledge discovery on an unprecedented scale[3]. Biomedical research increasingly uses methods from data mining, machine learning and text mining to investigate, for example, disease comorbidities, patient stratification, drug interactions and clinical outcome.

The ability to derive fine-grained patient phenotypes from health record data complements the increasingly detailed characterization of genetic variation and thus allows fine mapping of genotype–phenotype correlations. Detailed phenotyping is also expected to advance and partly automate the process of recruiting patients for clinical trials and case–control studies. The prospect of patient record data driving genomic research becomes especially interesting when traditional health-care-sector data is linked with biobanks and genetic data[4].

Despite the great potential, researchers who wish to analyse large amounts of patient data are still faced with technical challenges of integrating scattered, heterogeneous data, in addition to ethical and legal obstacles that limit access to the data[5,6]. It is hoped that large-scale adoption of health information technology (HIT) infrastructure in the form of electronic health records (EHRs) and agreed standards for interoperability and schemes for privacy and consent, will improve this situation (TABLE 1). With incentives for improved public health and the expected health budget savings[7,8], these matters are receiving much political attention worldwide. This is all part of a growing realization that secondary usage of patient data for population-wide research is key to bridging the translational gap between bench and bedside and moving closer to a realization of personalized and stratified medicine[9].

In this Review we first introduce the typical content of a generic EHR system. We then focus on how data-driven knowledge discovery on cohort-wide health data can fill knowledge gaps and assist informed clinical decision making. Next we describe how the integration of EHR and genetic data, together with systems biology approaches, can facilitate genotype–phenotype association studies. Finally we discuss some of the structural and political challenges that are facing EHR adoption and we comment on the perspectives and visions for the future.

Table 1 | **Relevant resources and initiatives**

Resource type	Resource	Website	Description
Electronic health records (EHRs) and HIT			
Standards development	OpenEHR	http://www.openehr.org	Open-source EHR standards initiative
	ISO (TC 215)	http://www.iso.org/iso/iso_technical_committee.html?commid=54960	International Standards Organization
	CEN (TC 251)	http://www.cen.eu/cen/Sectors/TechnicalCommitteesWorkshops/CENTechnicalCommittees/Pages/default.aspx	European Committee for Standardization
	HL7	http://www.hl7.org/implement/standards	Health Level 7 HIT standards
	CDISC	http://www.cdisc.org/standards	Data standards for clinical research data
Implementation and coordination	EuroRec	http://www.eurorec.org	European EHR adoption, interoperability and certification
	Integrating the healthcare enterprise	http://www.ihe.net	HIT standards, interoperability and certification
	HITSP	http://www.hitsp.org	Standards harmonization and interoperability
	epSOS	http://www.epsos.eu	Cross-border EHR data access in Europe
	Office of the National Coordinator for HIT	http://healthit.hhs.gov/portal/server.pt/community/healthit_hhs_gov__home/1204	Coordination of HITECH act programmes
	CEN–ISO 13606 Association	http://www.en13606.org	Promotion of the CEN–ISO 13606 standard
	NHS Connecting for Health	http://www.connectingforhealth.nhs.uk	UK National Health Service HIT strategy
Adoption monitoring	HIMSS analytics	http://www.himssanalytics.org; http://www.himssanalytics.eu	International EHR adoption monitoring
Health research			
Integrated DNA–EHR research databases	i2b2	https://www.i2b2.org	Partners HealthCare (opt-in participation)
	eMERGE Network	https://www.mc.vanderbilt.edu/victr/dcc/projects/acc	Cross-institutional EHR–DNA research network
	Vanderbilt BioVU	http://dbmi.mc.vanderbilt.edu/research/dnadatabank.html	Vanderbilt University (opt-out participation)
	Geisinger MYCODE	http://www.geisinger.org/research/centers_departments/genomics/mycode/mycode.html	Geisinger Research (opt-in participation)
	Kaiser RPGEH program	http://www.dor.kaiser.org/external/DORExternal/rpgeh	Kaiser Permanente (opt-in participation)
	Million Veteran Program	http://www.research.va.gov/mvp	US Department of Veteran Affairs (opt-in participation)
EHR research databases	Stanford STRIDE project	https://clinicalinformatics.stanford.edu/research/stride.html	Stanford University EHR research platform
Other	Medicare and Medicaid datasets	www.resdac.org	Insurance reimbursement database
	FDA Sentinel Initiative	http://www.fda.gov/Safety/FDAsSentinelInitiative	FDA product safety monitoring initiative
International initiatives	EHR4CR	http://www.ehr4cr.eu	European EHR research framework initiative
	Innovative Medicines Initiative	http://www.imi.europa.eu	European public–private funding initiative
	EU-ADR project	http://www.alert-project.org	European Union adverse drug reactions research
	I4Health network	http://www.i4health.eu	Bridging the gap between research and medicine
	European Medical Information Framework	Under construction	

Table 1 (cont.) | **Relevant resources and initiatives**

Resource type	Resource	Website	Description
Additional Ressources			
Terminologies and ontologies	UMLS	http://www.nlm.nih.gov/research/umls	Unified Medical Language System
	SNOMED CT (from the IHTSDO)	http://www.ihtsdo.org	International clinical terminology
	ICD (from the WHO)	http://www.who.int/classifications/icd/en	International Classification of Disease
Patient-focused initiatives	PatientsLikeMe	http://www.patientslikeme.com	Patient disease monitoring community
	23andMe	https://www.23andme.com	Personal genotyping
	Microsoft Healthvault	http://www.microsoft.com/en-us/healthvault	Personal health data management service

There are numerous web resources and coordination actions in the area of EHR research and development. This table lists some resources that are driven by authorities and public agencies, major network-oriented research projects and patient-focused initiatives. CDISC, Clinical Data Interchange Standards Consortium; EHR4CR, Electronic Health Records for Clinical Research; eMERGE, electronic Medical Records and Genomics; epSOS, European Patients – Smart open Services; FDA, US Food and Drug Administration; HIMSS, Healthcare Information and Management Systems Society; HIT, health information technology; HITECH, HIT for Economic and Clinical Health; HITSP, HIT Standards Panel; i2b2, Informatics for Integrating Biology and the Bedside; IHTSDO, International Health Terminology Standards Development Organization; RPGEH, Research Project on Genes, Environment and Health; SNOMED CT, Systematized Nomenclature of Medicine — Clinical Terms; STRIDE, Stanford Translational Research Integrated Database Environment; WHO, World Health Organization.

Electronic health data

Ideally, EHRs capture and integrate data on all aspects of care over time, with the data being represented according to relevant controlled vocabularies (FIG. 1). EHR adoption is growing thanks to initiatives like the US$19 billion HITECH act[10] in the United States and the €2 billion public–private partnership Innovative Medicines Initiative (IMI)[11] in the European Union. Many national strategies also exist for the development and adoption of nationwide interoperable HIT architectures and EHRs[12,13].

Not surprisingly, standardized, nationwide EHR systems are harder to implement in large countries than in smaller ones. According to the Healthcare Information and Management Systems Society (HIMSS) Analytics organization, larger countries (such as the United States, Canada, Germany, France, Italy and Spain) are behind several smaller European countries (such as Denmark, Holland and Sweden) in reaching the highest level of paperless data sharing, storage and decision support (Uwe Buddrus, HIMSS Analytics Europe, personal communication). Even though relatively few hospitals have adopted top-level EHR systems covering all aspects of hospital operations, a substantial fraction do have IT infrastructure in place that captures valuable research data from different auxiliary clinical and administrative systems[14]. Such data represent the majority of available retrospective data for years to come, even if advanced EHR systems were to be universally adopted instantaneously.

EHR data comprise various data types, from structured information such as drug prescription data consisting of dates and dosages that are captured through a standardized ePrescription system, to unstructured data such as clinical narratives that describe the medical reasoning behind the prescription (FIG. 1). This range of different data types highlights the challenge in EHR integration.

Administrative data. Much of the data that is captured in EHR systems serve administrative purposes, such as monitoring hospital activity and performance, and government or insurance reimbursement. Even simple EHR systems will typically capture demographic patient information such as age, gender, ethnicity and address, as well as structured information about a given encounter in the form of dates and ICD-encoded diagnoses (often referred to as billing codes). The almost ubiquitous use of these data types makes them a favoured source of phenotype information in population-based health research. However, because the amount of financial reimbursements depend on which codes are assigned to a given encounter there are biases in coding practice that need to be considered[15].

Often researchers do not obtain health data directly from the EHR systems, but instead from derived central health registries or insurance databases (such as Medicare), which gather data for reimbursement, quality assurance or health statistics purposes. Merged national registries on health and socio-economic status[16] are a valuable source of standardized, longitudinal, population-wide data for traditional epidemiological studies[17,18].

Ancillary clinical data. Regarding the generation of clinical data, the most important ancillary hospital functions are those that are provided by laboratories, pharmacies, and radiological and medical imaging departments. Physicians increasingly place requests for these services electronically through computerized physician order entry (CPOE) systems that automatically log detailed information about the transaction. The data that are recorded from these transactions include information on drugs (dosages and time periods for prescriptions), and the test types and results from laboratory requisitions. Radiological imaging requests and results are exchanged the same way, but the extraction of information traditionally relies on human

Electronic health records (EHRs). In this review we do not distinguish between EHRs, Electronic Patient Records (EPRs) and Electronic Medical Records (EMRs).

HITECH act
Part of the American Recovery and Reinvestment act from 2009. The Health Information Technology for Economic and Clinical Health (HITECH) act allocates funding and attention to HIT infrastructure and electronic health record adoption and research in the United States.

ICD
The International Classification of Diseases (ICD) published by the World Health Organization. It has been translated into numerous languages.

Medicare
A US government health insurance programme primarily covering people aged 65 years and older.

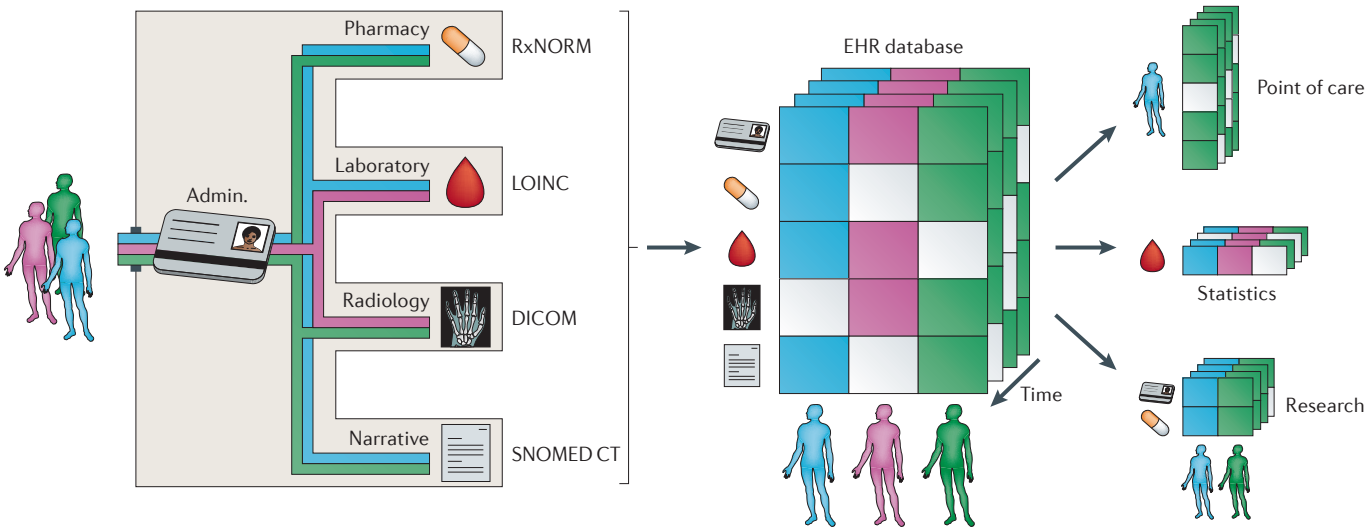


Figure 1 | **Electronic health record content.** The electronic health record (EHR) of a patient can be viewed as a repository of information regarding his or her health status in a computer-readable form. An encounter with the health-care system generates various types of patient-linked data. In the example shown, medication, laboratory, imaging and narrative data are all generated. Each data type is ideally captured according to standards or classifications, such as RxNorm[111] for prescription data, Logical Observation Identifiers Names and Codes (LOINC) for laboratory data and Digital Imaging and Communication in Medicine (DICOM) for imaging files. Clinical narratives are inherently free text, but often contain clinical terms that are coded according to International Classification of Disease-9 (ICD-9) or ICD-10 (REF. 112) or Systematized Nomenclature of Medicine — Clinical Terms (SNOMED CT)[25]. Integrated auto-coding systems may in some cases map free text to clinical terms. Patient data are stored in a database and can be viewed in formats matching the needs and authorities of specific user groups. For example, a clinician might request EHR data for a particular patient, a statistical summary of all laboratory procedures and a specified cohort extraction for drug research.

interpretation. However, image-analysis techniques can automate this data extraction (for example, by detecting pulmonary embolisms in computed tomography angiographies[19]), thus adding a structured component that is usable in CDS systems.

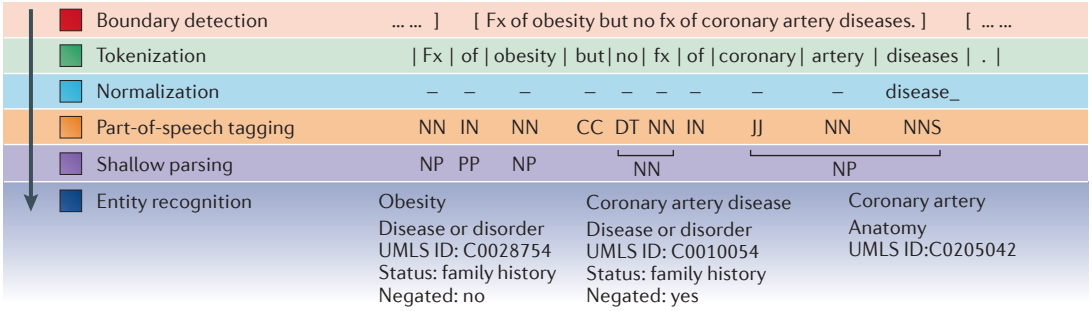
Another ancillary source of potentially structured data is genotype and sequence data. Genotyping is still relatively novel in standard clinical practice, but common examples of its use include testing for primary lactose intolerance, cystic fibrosis, or BRCA variants in breast cancer therapy. Representation of the actual genetic sequence information is currently not well supported in EHR systems but will probably be included as genotyping becomes more widespread in clinical practice[20,21].

Clinical text. Written or dictated clinical narratives such as admission notes, treatment plans and patient summaries continue to be a cornerstone in the everyday process of ensuring informed decision making through the recording and consultation of careful clinical documentation. Clinical text is the most abundant data type, but is also the most difficult to analyse computationally. It is highly heterogeneous, does not always conform to normal grammar and is rich in author- and domain-specific idiosyncrasies, acronyms and abbreviations, as well as spelling and typing errors[22]. The context is complicated by many negations or references to different subjects, and assessments are often tentative or uncertain. Free text allows the flexibility to express case nuances and clinical reasoning; this flexibility is valued by clinicians and is not easily replaced by structured reporting formats that are often considered inflexible and time consuming[23]. Striking the right balance between expressiveness and structure is a complex and domain-dependent matter[24].

Deriving structured information about patient phenotypes from clinical text generally requires named entities or concepts in the text to be recognized and mapped to codes in a relevant controlled vocabulary such as Systematized Nomenclature of Medicine — Clinical Terms (SNOMED CT)[25] or one of the other >100 vocabularies in the Unified Medical Language System[26]. This is usually done using natural language processing (NLP) tools, which combine a range of linguistic, statistical and heuristic methods to analyse free text. BOX 1 describes typical tasks that are involved in using NLP to extract structured information (such as disease, drug, symptom and procedure terms) from clinical text.

The Mayo clinic's clinical Text Analysis and Knowledge Extraction System (cTAKES)[27], the Informatics for Integrating Biology and the Bedside (i2b2) HITEX[28] and the long-running Medical Language Extraction and Encoding (MedLEE)[29] system from Columbia University are all examples of NLP-based clinical phenotyping systems. Although still mostly used in research contexts, MedLEE has demonstrated performance similar to expert human curators[30] and is in clinical operation at several health facilities. Another system, MedEx[31], focuses on extracting detailed medication data from text. For a more detailed account

Box 1 | **Natural language processing**



Typical natural language processing (NLP) steps exemplified by the clinical Text Analysis and Knowledge Extraction System (cTAKES) clinical text-mining pipeline[27]. First, sentence boundary detection splits the text into units of individual sentences. This is followed by tokenization, which splits the text using space and punctuation as a guide to identify individual tokens (typically individual words), with rules for handling special cases such as dates. Tokens are reduced to a base form by normalizing, for example, case, inflection or spelling variants. The next step assigns part-of-speech tags to each token to identify its grammatical category in the context (for example, NN for noun, IN for preposition or JJ for adjective). This is not a trivial task as many words have ambiguous meaning. After the tokens have been tagged, the shallow parsing step identifies syntactic units, most importantly noun phrases (NPs), which are grammatical units, built from a noun with optional modifiers such as adjectives. In the entity recognition step, NPs and various lexical permutations are then mapped to controlled vocabularies using tools such as MetaMap[106]. Importantly, such systems also identify the presence of negating terms, such as 'no' or 'never', near identified entities. The various steps are typically implemented using combinations of logical rules (and their exceptions) and machine-learning methods. For example, a full stop (period) followed by a space and a capital letter indicates a sentence boundary. In the figure, two disorders are identified as well as one anatomical structure. Both disorders are tagged as relating to family history (Fx), and, in the case of coronary artery disease, the preceding word 'no' tags the term as negated. Clinical information extraction systems generally perform best when fine-tuned for specific tasks or clinical domains, such as identifying smoking status or analysing radiology reports. Vocabularies can be customized for a task with domain-specific terms, and the rules and training can be focused. The annual Informatics for Integrating Biology and the Bedside (i2b2) NLP shared tasks[107–110] meeting provides a good demonstration of state-of-the-art practice in clinical NLP applied to increasingly difficult challenges. The 2010 challenge prompted participants to extract concepts, assertions and relations from clinical text[110]. CC, coordinated conjuction; DT, determiner; NNS, plural noun.

of NLP and the text mining of clinical text see REF. 22 and the September 2011 special issue of the Journal of the American Medical Informatics Association (JAMIA)[32].

EHR knowledge discovery informing care

EHR at the point of care. Cohort-wide data mining of EHR databases provides an approach to generate context-specific and clinically actionable knowledge that can inform the tailoring of treatments to the individual. Clinical decision making is a complicated task in which the physician must attempt to bridge what has been referred to as an inferential gap[1] between the information at hand in a given case and the clinical knowledge that is required to decide on the best treatment. EHR systems can narrow this gap by programmatically implementing clinical guidelines in CDS systems that can process all of the EHR data that have been recorded about the patient. A common CDS module in clinical operation is for monitoring conflicts between ordered drug prescriptions and recorded allergies or existing prescriptions. Another example is for the early warning of infection, with suggestions for antibiotic treatment and dosage based on physiological data[33]. For complex rules or when the necessary information is not available in structured form, NLP-based processing of text may supplement structured data points[34].

As described in the next sections, exploring this knowledge-discovery potential from multivariate health data increasingly involves a suite of statistical, machine-learning and computational methods that are often referred to as data-mining methods or knowledge discovery in databases (KDD) methods[35–37].

Correlating clinical features. A clinical topic of great interest is disease co-occurrence (comorbidity). In clinical practice, the cumulated disease burden of a patient is often summarized in comorbidity scores such as the Charlson index, and these scores are also often used to interpret confounding effects of comorbid diseases in cohort studies[38]. Except for clear hypothesis-driven investigations of specific diseases, small cohort sizes have often not allowed a detailed view of the co-occurrence patterns of individual pairs or small sets of diseases. Structured and narrative EHR and registry data allow us to approach comorbidity in a data-driven statistical way by using simple contingency tables to quantify comorbidities for any pairs of diseases in terms of the disproportionality with which they co-occur in large cohorts[39–41] (FIG. 2a). Population-wide comorbidity patterns from Medicare data have, for example, been visualized in comorbidity networks[42]. Combining and validating comorbidity evidence from EHR text mining,

Charlson index
A measure of the accumulated disease burden for a patient. It is calculated as a weighted sum of 22 selected medical conditions that are assigned scores depending on the severity of the condition.

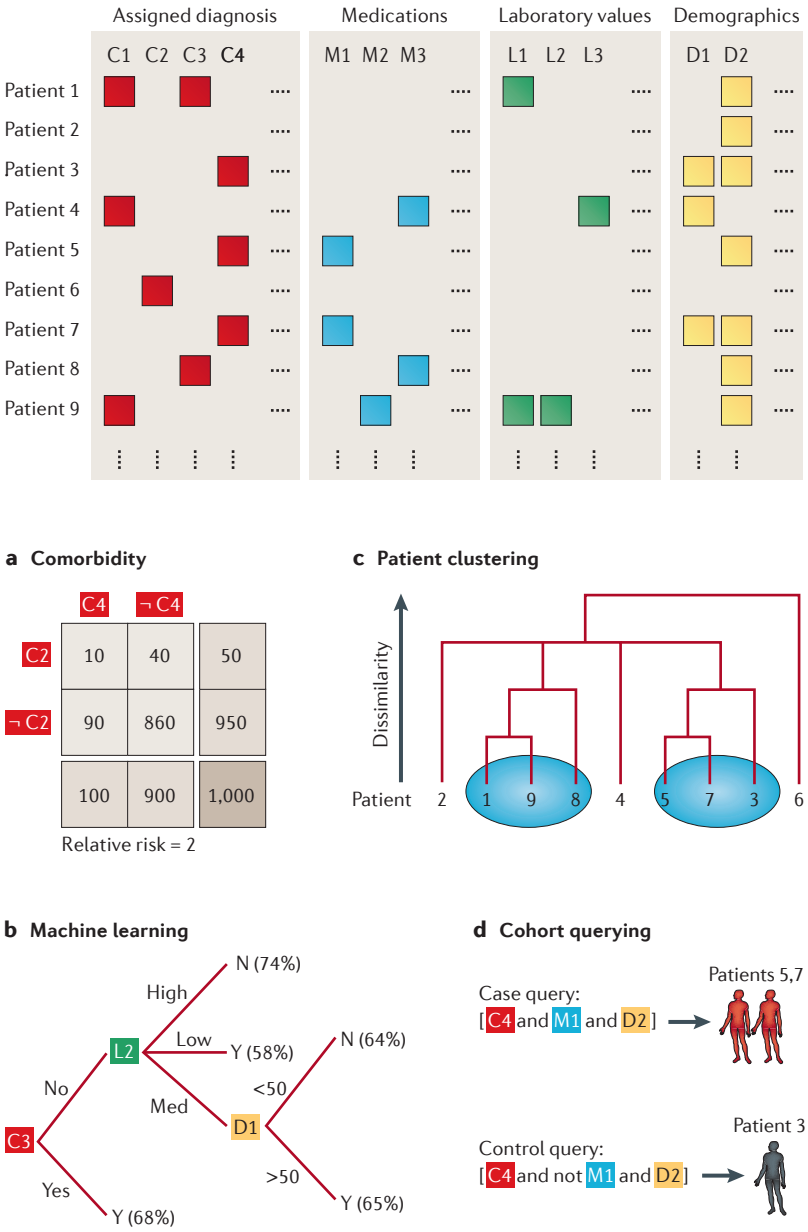


Figure 2 | **Four ways to analyse EHR data.** A simplified illustration of an electronic health record (EHR) research database and some of the data-driven methods that are discussed in the main text. We represent 1,000 patients as a binary association or non-association with clinical features. These features include diagnosis codes (in red), medications (in blue), laboratory data (in green) and demographic data (in yellow), but we consider no temporal aspects. **a** | From the database prevalence of any two diseases (here, C4 = 100 and C2 = 50) it is possible to calculate the expected number of patients associated with both diseases assuming independence ((100/1000)×(50/1000)×1000 = 5) and compare it with the observed value (10 in this example), thus giving a relative risk of 2. **b** | A very simple decision tree classifier illustrating how the likelihoods of two outcomes (Y and N) are classified using three clinical features. Associations are illustrated with diagnosis C3, results from laboratory test L2 and age (whether higher or lower than 50 years). **c** | Using a simple measure for clinical similarity (number of identical associations) hierarchical-clustering methods can be used to group patients according to their clinical profile. **d** | In order to assemble a case cohort with matching controls, a query can be submitted to the database for patients associated with, and not associated with, certain features. In this example, both cases and controls must be associated with diagnosis C4 and the demographic feature D2, but only cases are allowed association with medication M1.

with medical knowledge extracted through literature mining, has also been attempted[41].

In pharmacovigilance, disproportionality methods are used in post-marketing drug surveillance[43], which is mainly based on the voluntary reporting of adverse drug event (ADE) data to spontaneous reporting databases; such events suffer from gross underreporting[44]. As an alternative, mining of medication and ADE data that are captured routinely in EHRs is being explored as an approach to rapidly uncover drug–ADE associations[45]. Such an approach may be more powerful if data from several countries are pooled, such as in the European Union adverse drug reactions (EU-ADR) project[46]; the international nature of the project emphasizes once again the importance of EHR interoperability. Similar approaches have been used to investigate off-label drug prescriptions and adverse drug interactions[47].

Association rule mining[48,49] is a KDD method that is often applied to the challenge of exploring the enormous combinatorial space of clinical features to identify potential correlations. From a frequently co-occurring set of features (for example, [warfarin, aspirin, bleeding]) the method identifies statistical rules (such as [warfarin, aspirin] → [bleeding]). This indicates an increased likelihood of a person being associated with the clinical finding 'bleeding' if that patient is also associated with the drugs warfarin and aspirin[50].

To identify interesting correlations from a background of trivial and spurious findings, any data-driven approach must filter the results based on statistical significance, interest and novelty. Seemingly exciting correlations are often due to transitive relationships; for example, a correlation between insulin use and hypertension is trivially explained by insulin use correlating with diabetes, which in turn correlates with hypertension. However, transitive relationships may themselves be interesting discoveries. In text mining of the biomedical literature this approach has long been used to discover hidden associations[51,52]. Ultimately, manual curation by domain experts remains a necessary layer of analysis[40].

Because most current EHR data sets that cover specific disease areas over long time periods are only of moderate size, findings often need to be validated in independent data sets. These can come either from EHR systems or from population-wide registries that collect data across hospitals and geographical regions. There is thus a high degree of synergy between EHR data mining and conventional epidemiology on registry data, because the latter can be used to replicate discoveries that have been made in smaller EHR cohorts.

Prediction from data. The existence and detection of correlations in data provide the basis for predicting future patient outcomes from a given scenario. Predictive clinical modelling uses machine-learning methods to build multivariate models from clinical data and subsequently to make inferences on unknown data (FIG. 2b; BOX 2). Examples include the prediction of surgery outcome[53], breast cancer survival[54] and coronary heart disease risk[55] from variables such as age, sex, smoking

Box 2 | **Machine learning on EHR data**

Machine-learning techniques are data-driven approaches that are designed to discover statistical patterns in high-dimensional, multivariate data sets, such as those that are frequently found in electronic health record (EHR) systems. The starting point for machine learning is a data set of training examples, which in the context of EHRs typically originate from individual patients. Each example is represented by a feature vector, which may be any combination of data items that are stored in EHR systems. Machine-learning methodology is well suited for handling nonlinear correlations between the feature vector components and metadata, such as patient subcategories.

Supervised and unsupervised learning

Data sets come in two types: labelled and unlabelled. In labelled data sets, each example has a pre-assigned category or value, whereas in unlabelled data sets this is not the case. Machine-learning methods can be grouped into supervised and unsupervised approaches according to the labelling of the examples.

A supervised training method handles a data set of labelled examples from which it derives a model that predicts the labels from the features. Labels may represent, for example, alternative diagnoses to be predicted from laboratory test results. Some of the most commonly used supervised methods are: naive Bayes; artificial neural networks; support vector machines; and random forests. By contrast, unsupervised methods, such as self-organizing maps and clustering algorithms, take an unlabelled data set and attempt to find groups of examples sharing similar features. Patient stratification is a typical medical use for such methods.

Strengths and weaknesses

Data from EHR systems are challenging to analyse for a variety of reasons. The data have many dimensions but are sparse (that is, many features describe patients but most of them are typically absent for any given patient). The features are heterogeneous, encompassing quantitative data, categorical data and text. Furthermore, these data are subject to random errors and systematic biases.

Given sufficiently large data sets, most machine-learning methods are highly robust to random errors, both in the input features and the labelling. This is especially true if they are combined with preprocessing and so-called feature selection algorithms such as principal component analysis and data normalization to put them on a common scale.

The Achilles heel of machine-learning methods — and all other statistical methods — is systematic bias in the data. Being purely data-driven, the methods have no way of distinguishing medically relevant signals from systematic, undesired biases in the data. For example, this could be the systematic erroneous use of disease terminology codes caused by strategic billing. Another risk in machine learning is overfitting, for which the predictive performance is overestimated owing to a lack of data, or because incorrect test or validation procedures are used to create models.

Clinical usage

The safety considerations of clinical work mean that for decision support relating to treatment, machine-learning models generally serve only a supporting role. Whether predictive models outperform the manual inspection of EHR data clearly depends on the specific task. With the growing number of recorded features in EHRs, machine-learning methodology has a huge potential for adding and complementing expertise that is currently held by staff in the health-care sector.

status, hypertension and various biomarkers. Another high-profile example is the $3 million Heritage Health Prize data-mining competition that invites participants to predict future hospital admission based on insurance claims data, prescription data and laboratory test data[56]. Identifying the most explanatory features from the set of all features is a crucial task in this type of supervised learning.

Although some methodology exists[36], temporal data mining of longitudinal health data is still in its early stages. Uncovering patterns in patient trajectories through disease and intervention nodes (such as medication or clinical procedures) in a clinical feature space is a statistically and computationally challenging task that has so far not been thoroughly explored[57]. Establishing patterns of directionality in comorbidity and disease progression is a first step towards using EHRs for predictive purposes, and this has been explored in network analysis studies of Medicare data[42,58]. Health data that are recorded over short timescales (such as time series data of blood glucose levels) are sometimes collected in EHR systems; for these short-term data, various analysis methods exist[59].

Patient stratification. The general goal of clinical prediction models is the stratification of a patient cohort into different subpopulations so that within each subpopulation, patients have similar characteristics, such as likely outcomes, risk or prognosis. A more direct approach to stratification is to use clustering methods and semantic similarity metrics to group patients[60,61] based on their associated clinical features and temporal patterns[40] (FIG. 2c).

Much clinical and genetic research depends critically on the identification and recruitment of large, phenotypically constricted cohorts of cases and a matched set of controls to ensure statistical power for rare and weakly penetrant alleles. Similarly, clinical trials rely on the homogeneity of the study population. The stratification task has remained costly and time consuming, while the cost of, for example, genotyping has dropped dramatically. One of the biggest research promises of EHR systems is to alleviate this bottleneck in cohort studies[62].

Phenotype querying of structured data and NLP-encoded text in derived, de-identified EHR databases (FIG. 2d) facilitates the inexpensive and timely identification of matching subjects for cohort studies. This is the scenario realized in the i2b2 framework[63] and in the electronic Medical Records and Genomics Network (eMERGE Network)[64]. A high precision[65,66] of the approach has been documented and demonstrated in genome-wide association studies (GWASs) and other types of cohort research[67–74]. In Europe, the IMI-funded Electronic Health Records for Clinical Research (EHR4CR) initiative aims to build a similar, multi-language platform for EHR-based medical research across distributed EHR systems.

Linking to the molecular level

Integrating genetics. Realizing the full potential of EHR-based recruitment of patients for genetic research requires a framework for the easy acquisition of matched DNA samples and for patient consent[4]. Affiliates of i2b2 and eMERGE, such as Vanderbilt University, have facilitated this by linking EHR data to biobanked blood samples that have been accrued during routine clinical care[62,64]. Sanction for the use of EHR data in research in anonymous form is obtained as part of the standard treatment consent form[75]. This type of integrated biobanking not only leverages the entire recruitment process, but also enables the re-use of patient samples and genetic data in later studies that require little or no new genotyping[70]. The addition of genetic data also extends the set of features that are available for predictive modelling and rule-mining approaches.

Pharmacovigilance
Monitoring of adverse drug events during clinical trials and after marketing in order to prevent harm to patients. It is typically based on statistical pattern-finding in databases of reported adverse events.

Adverse drug event
(ADE). Used in pharmacology to describe any unexpected or harmful event associated with a given medication.

Feature vector
The representation of objects (patients) as vectors in the space of all relevant features. Each dimension of the vector specifies the association of a patient with a certain feature.

Clustering
A common task in statistical data exploration using measures of similarity between data points, network topology or other methods to group data points with similar characteristics together in clusters.

Semantic similarity
A measure of the similarity of two concepts in terms of their meaning or semantic content. Often quantified using topological measures of distance in an ontology of concepts, such as WordNet or Systematized Nomenclature of Medicine — Clinical Terms (SNOMED CT).

Electronic Medical Records and Genomics Network
(eMERGE Network). An institutional network that is exploring the potential of electronic health record data in genetic and medical research. Participating institutions are: GroupHealth, Geisinger, Marshfield Clinic, Mayo Clinic, Mount Sinai School of Medicine, Northwestern University and Vanderbilt University.

Pharmacogenomics
The study of how genetic variants influence the effects of drugs on, for example, drug metabolism, efficacy and toxicity, with the goal of improving and personalizing drug therapy.

Million Veteran Program
A research project initiated by the Veterans Affairs Office of Research and Development that is aimed at establishing a database with DNA and health record data from one million people. Participation is opt-in.

Kaiser RPGEH
The Kaiser Permanente Research Project on Genes, Environment and Health (Kaiser RPGEH). This project aims to establish a research database with genetic data, environment data and health record data from 500,000 people. Participation is opt-in.

The general goal of combining detailed EHR-based patient phenotyping and genetic data has also inspired an interesting reversal of the GWAS approach to gene–disease association. A phenome-wide association study (PheWAS)[76], introduced by Vanderbilt BioVU, starts with the individual SNP and checks for statistical association against hundreds of disease phenotypes of patients that have been genotyped for that SNP.

Pharmacogenomics is another research area that is embracing the integration of EHR data with genetic data[77]. It is expected to have large translational impact through basing therapeutic choices and CDS on a patient's genetics[78,79]. Drug efficacy is influenced by genetic variation, as is seen in the example of the dose response to the anticoagulant warfarin being affected by at least three genetic variants[80].

The detailed longitudinal patient profile that can be assembled from EHR data enables drug exposure profiles to be correlated with treatment outcome measures, such as efficacy and toxicity, using structured and unstructured sources. Linked biobank and genetic data then allow the association of such correlations with the underlying genotype. One study based on biobanking and EHR data identified genetic variants that are associated with an increased risk of thromboembolism in patients with breast cancer that were treated with Tamoxifen[81].

Systems biology and gene-network-based decision support. Most phenotypes are influenced by numerous genes. Systems biology approaches go beyond the individual gene to instead analyse protein complexes, pathways and gene networks to bridge the gap between genetic linkage and the underlying molecular biology. Taking a systems-level view of phenotypes can also shed new light on the temporal aspects of phenotypes; for example, in explaining how different mutations in the same genes can lead to disorders that are related to different stages of heart development[82]. Similar types of analyses can be used to interpret differences in the human microbiomes between individuals, as shown in a pathway analysis of gut metagenomics data in the context of obesity and inflammatory bowel disease[83].

Most systems biology approaches for identifying disease genotype–phenotype relationships have started from genetic linkage data rather than traditional EHR data. That being said, it is possible to estimate the degree of genetic overlap between two diseases based on traditional EHR data alone[84]. Also, traditional EHR data from one cohort can be combined with genetics data from other cohorts in an attempt to unravel the molecular basis of comorbidities. An approach to investigate the underlying molecular aetiology of disease correlations that have been uncovered using EHR data is to map the diseases to known associated genes and proteins, and to investigate the resulting protein–protein interaction network for statistical overlaps[40]. This has also been an approach in network medicine, in which diseases are clustered based on shared associated genes, as is seen, for example, in the human disease network[85]. Using comorbidities from Medicare claims data, Park *et al.*[86] used gene–disease association data to document that higher comorbidity was related to increased genetic overlap.

Today, patient-specific information on sequence variation in the genes and proteins included in systems-level models of phenotypes exists mostly for selected research cohorts. However, this will soon change owing to the rapid decline in the cost and time required to sequence a human genome. This will make it possible to include sequence variation in phenotype-specific network models and to use these for decision making in the health-care sector and in the development of personalized treatment regimens.

A key obstacle in the use of genome data for decision making in the clinic is the billions of features that are contained in a single human genome[87]. The huge number of sequence variants represents a general problem of multiple testing and makes it difficult to discriminate between 'causal' variation that has predictive value in the clinic and the substantial amount of 'passenger' variation that travels along in an uncorrelated manner. Systems-level analyses can drastically reduce the combinatorial problem by grouping individual genetic variants that affect the same molecular machinery, thus increasing the feasibility of turning EHR data into valuable clinical markers relative to single-gene approaches[88,89].

Limiting factors — key problems to overcome

Although researchers working at the interface between bioinformatics, systems biology and medical informatics are eager to analyse and integrate medical record data, a wide range of factors are delaying this development. Data-mining analysts are generally accustomed to handling disorganized databases, incompatible formats and missing database interoperability; these problems need to be solved in the long run if the full potential of this type of analysis is to be reached. However, the main impediment today is the problem of making simple data dumps available to researchers.

In a recent public address, the British Prime Minister, David Cameron, announced plans to make the UK National Health Service data available for research. Furthermore, in the United States, large federal grants are aimed at promoting health research through establishing extensive patient databases such as the Million Veterans Program and Kaiser RPGEH. The restriction on access to existing data is the primary hindrance to development.

Privacy, autonomy and consent. Although patients typically have no legally recognized property right to their health records[90], privacy legislation in many countries has traditionally placed great weight on personal autonomy and has required informed consent for accessing personal health data for research. In the case of health databases this has often been translated into opt-in participation models. From a research point of view this amounts to increased time and cost[91] and the risk of biased data resources owing to differing inclinations to participate among demographic groups[92]. It has been

argued that today, when database-driven research can be conducted without interfering with people's lives, legacy-consent models that have been put in place to protect research subjects from harm are no longer relevant and should be updated or waivered in the interest of the common good[6,93]. Whereas this is already largely the case in, for example, Scandinavian countries, such loss of autonomy would probably be met by public outcry in countries with strong privacy advocacy movements, such as the United States.

A legitimate public concern that is related to the use of personal health data for research is the risk of privacy breaches. A technical solution is to de-identify research data according to specifications such as those in the Health Insurance Portability and Accountability Act (HIPAA) privacy rule. De-identification normally allows researchers to circumvent costly consent regimes[94], but the lack of identifiers makes certain types of population-wide research impossible, as other information cannot be linked to data subjects. Moreover, despite such precautions, re-identification has been shown on some occasions to be a genuine risk[95], especially when data on human DNA are involved, as even a relatively small set of markers can enable unique identification[96–98]. Similar concerns apply to other types of molecular-level data, for example, individual gut microbiome profiles[99]. Another public concern that is at odds with loosened consent models is the fear that personal data could be used for research that is in conflict with an individual's ethical, moral or religious convictions[100]. The permanency of data only adds to these concerns, as many of the data types will largely remain static throughout life, whereas data protection should be effective over long periods of time.

Interoperability across institutions, countries and continents. By analogy to the meta-analysis of epidemiological and genetic data, in which single cohorts often provide insufficient statistical power to make inferences, EHR data need to be merged across regional barriers in order to provide the strongest basis for research. Due to the currently limited EHR adoption and interoperability, centralized health registries partially fulfil this role, although their level of detail is lower than what is obtainable from complete EHRs.

True data interoperability requires the development and implementation of standards and clinical-content models[101,102] for the unambiguous representation and exchange of clinical meaning[103]. Various international certification and standards bodies pursue this goal, such as the European Committee for Standardization (CEN), the International Standards Organization (ISO) and Health Level 7 (HL7).

The CEN–ISO 13606 standard[101] adopts the concept of archetypes to formalize all aspects of a concept that a clinician might want to describe and combines this with international ontologies such as SNOMED CT. Archetypes separate the representation of clinical data from the underlying technical implementation and are authored by domain experts themselves without the need for technical expertise[104]. For example, an archetype for the concept 'blood pressure' would contain a representation of both the actual blood pressure measurements and contextual attributes (patient resting or active; sitting or lying; arm or leg; or left or right) that could be important for the correct interpretation of data. An archetype for a broader composite concept such as 'family history' becomes more complex.

Implementations of such standards in current and future EHR systems are underway and will make it possible to establish large cohort studies with harmonized phenotypes and will enable the creation of international health data repositories for research.

Outlook

The molecular-level characterization of human diversity is becoming continually more detailed, not only in terms of basic DNA sequencing but also regarding histone modifications and their functional impact. By contrast, the phenotypic manifestations of all this detail are often insufficiently represented using broad categories of disease, in which categories typically represent end points rather than exhaustive trajectories of disease development. Clearly, if the functional impact of human genetic variation is to be fine-mapped onto the space of phenotypes, data such as those that are included in EHRs will be needed in order to define, by data-driven methods, detailed phenotypes that also cover their inherent comorbidities. In turn, this may make it possible to categorize the functional part of human genetic variation in much higher detail.

Citizens are increasingly becoming engaged and empowered participants in managing their own health. For example, this can be through using tools for extracting data from hospital systems, general-practitioner systems and pharmacies. As a special case of today's data sharing trend, online health resources such as PatientsLikeMe allow patients to share detailed health and treatment information, which again is providing a novel data source for cross-cohort studies[105]. This type of resource may also be seen as a reaction to the fact that the official EHR data are not being made sufficiently available for research. However, although many of these developments focus on access for the individual, what is clearly needed is access for researchers across cohorts and populations. This is still the weak aspect in the research on new data sources such as EHRs.

A recent survey by HIMSS Analytics Europe identified 'funding' as the key barrier to progress in the EHR area, and other factors such as 'staff habits' were also high on the list in many countries. The intimate link to funding also makes it hard to forecast when a major shift in data availability may take place. As EHRs and EHR-related data mining is meant to reduce the overall cost of health care long-term (and increase the quality of life), this is obviously an incongruity. However, it is also evident that there is no strong conflict of interest between the populations, politicians and researchers. All stakeholders have a joint and urgent task to solve EHR challenges, but the solutions hold great promise.

1. Stewart, W. F., Shah, N. R., Selna, M. J., Paulus, R. A. & Walker, J. M. Bridging the inferential gap: the electronic health record and clinical evidence. *Health Aff.* **26**, w181–w191 (2007).
2. Hillestad, R. *et al.* Can electronic medical record systems transform health care? Potential health benefits, savings, and costs. *Health Aff.* **24**, 1103–1117 (2005).
3. Prokosch, H.-U. & Ganslandt, T. Perspectives for medical informatics. Reusing the electronic medical record for clinical research. *Methods Inf. Med.* **1**, 38–44 (2009).
4. Kohane, I. S. Using electronic health records to drive discovery in disease genomics. *Nature Rev. Genet.* **12**, 417–428 (2011).
5. Kush, R. D., Helton, E., Rockhold, F. W. & Hardison, C. D. Electronic health records, medical research, and the Tower of Babel. *N. Eng. J. Med.* **358**, 1738–1740 (2008).
6. Taylor, P. When consent gets in the way. *Nature* **456**, 32–33 (2008).
7. Himmelstein, D. U., Wright, A. & Woolhandler, S. Hospital computing and the costs and quality of care: a national study. *Am. J. Med.* **123**, 40–46 (2010).
8. Buntin, M. B., Burke, M. F., Hoaglin, M. C. & Blumenthal, D. The benefits of health information technology: a review of the recent literature shows predominantly positive results. *Health Aff.* **30**, 464–471 (2011).
9. Sarkar, I. N. Biomedical informatics and translational medicine. *J. Transl. Med.* **8**, 22 (2010).
10. Blumenthal, D. Launching HITECH. *N. Eng. J. Med.* **362**, 382–385 (2010).
11. Hunter, J. The Innovative Medicines Initiative: a pre-competitive initiative to enhance the biomedical science base of Europe to expedite the development of new medicines for patients. *Drug Discov. Today* **13**, 371–373 (2008).
12. Coiera, E. Building a National Health IT System from the middle out. *J. Am. Med. Inform. Assoc.* **16**, 271–273 (2009).
13. Morrison, Z., Robertson, A., Cresswell, K., Crowe, S. & Sheikh, A. Understanding contrasting approaches to nationwide implementations of electronic health record systems: England, the USA and Australia. *J. Healthc. Engin.* **2**, 25–41 (2010).
14. Jha, A. K., DesRoches, C. M., Kralovec, P. D. & Joshi, M. S. A progress report on electronic health records in US hospitals. *Health Aff.* **29**, 1951–1957 (2010).
15. Serdén, L., Lindqvist, R. & Rosén, M. Have DRG-based prospective payment systems influenced the number of secondary diagnoses in health care administrative data? *Health Policy* **65**, 101–107 (2003).
16. Thygesen, L. C., Daasnes, C., Thaulow, I. & Bronnum-Hansen, H. Introduction to Danish (nationwide) registers on health and social issues: structure, access, legislation, and archiving. *Scand. J. Public Health* **39**, 12–16 (2011).
An overview of Danish health and socio-economic registries and research possibilities as an example of extensive population-wide registration.
17. Frank, L. When an entire country is a cohort. *Science* **287**, 2398–2399 (2000).
18. Øyen, N. *et al.* Recurrence of congenital heart defects in families. *Circulation* **120**, 295–301 (2009).
19. Masutani, Y., MacMahon, H. & Doi, K. Computerized detection of pulmonary embolism in spiral CT angiography based on volumetric image analysis. *IEEE Trans. Med. Imaging.* **21**, 1517–1523 (2002).
20. Hoffman, M. The genome-enabled electronic medical record. *J. Biomed. Inform.* **40**, 44–46 (2007).
21. Sax, U. & Schmidt, S. Integration of genomic data in Electronic Health Records — opportunities and dilemmas. *Methods Inform. Med.* **44**, 546–550 (2005).
22. Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C. & Hurdle, J. F. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb. Med. Inform.* **2008**, 128–144 (2008).
An introduction to NLP and information extraction in the challenging clinical context, which also reviews the relevant research in the field.
23. Rosenbloom, S. T. *et al.* Data from clinical notes: a perspective on the tension between structure and flexible documentation. *J. Am. Med. Inform. Assoc.* **8**, 181–186 (2011).
A summary of the conflicting views on structured and narrative health data in the context of how to produce valuable and reusable data.
24. Johnson, S. B. *et al.* An electronic health record based on structured narrative. *J. Am. Med. Inform. Assoc.* **15**, 54–65 (2008).
25. The International Health Terminology Standards Development Organisation. Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT). [online], http://www.ihtsdo.org/snomed-ct/snomed-present/introducing-snomed-ct
26. Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* **32**, D267–D270 (2004).
27. Savova, G. K. *et al.* Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J. Am. Med. Inform. Assoc.* **17**, 507–513 (2010).
28. Zeng, Q. T. *et al.* Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC Med. Inform. Decis. Mak.* **6**, 30 (2006).
29. Friedman, C., Alderson, P. O., Austin, J. H. M., Cimino, J. J. & Johnson, S. B. A general natural-language text processor for clinical radiology. *J. Am. Med. Inform. Assoc.* **1**, 161–174 (1994).
30. Friedman, C., Shagina, L., Lussier, Y. & Hripcsak, G. Automated encoding of clinical documents based on natural language processing. *J. Am. Med. Inform. Assoc.* **11**, 392–402 (2004).
31. Xu, H. *et al.* MedEx: a medication information extraction system for clinical narratives. *J. Am. Med. Inform. Assoc.* **17**, 19–24 (2010).
32. Ohno-Machado, L. Realizing the full potential of electronic health records: the role of natural language processing. *J. Am. Med. Inform. Assoc.* **18**, 539 (2011).
33. Evans, R. S. *et al.* A computer-assisted management program for antibiotics and other antiinfective agents. *N. Eng. J. Med.* **338**, 232–238 (1998).
34. Demner-Fushman, D., Chapman, W. W. & McDonald, C. J. What can natural language processing do for clinical decision support? *J. Biomed. Inform.* **42**, 760–772 (2009).
35. Bellazzi, R. & Zupan, B. Predictive data mining in clinical medicine: current issues and guidelines. *Int. J. Med. Inform.* **77**, 81–97 (2008).
A review of the use of predictive methods in medicine with a special focus on temporal data.
36. Bellazzi, R., Ferrazzi, F. & Sacchi, L. Predictive data mining in clinical medicine: a focus on selected methods and applications. *WIREs Data Mining Knowl. Discov.* **1**, 416–430 (2011).
37. Lavrac, N. Selected techniques for data mining in medicine. *Artif. Intell. Med.* **16**, 3–23 (1999).
38. Degroot, V., Beckerman, H., Lankhorst, G. & Bouter, L. How to measure comorbidity. A critical review of available methods. *J. Clin. Epidemiol.* **56**, 221–229 (2003).
39. Hanauer, D., Rhodes, D. R. & Chinnaiyan, A. M. Exploring clinical associations using "-omics" based enrichment analyses. *PLoS ONE* **4**, e5203 (2009).
40. Roque, F. S. *et al.* Using electronic patient records to discover disease correlations and stratify patient cohorts. *PLoS Comput. Biol.* **7**, e1002141 (2011).
Patient stratification and discovery of disease comorbidities and their causes at the molecular level using structured data and text mining on a psychiatric cohort.
41. Holmes, A. B. *et al.* Discovering disease associations by integrating electronic clinical data and medical literature. *PLoS ONE* **6**, e21132 (2011).
42. Hidalgo, C., Blumm, N., Barabási, A.-L. & Christakis, N. A dynamic network approach for the study of human phenotypes. *PLoS Comput. Biol.* **5**, e1000353 (2009).
43. Gibbons, R. D. *et al.* Post-approval drug safety surveillance. *Annu. Rev. Public Health* **2010**, 419–437 (2010).
44. Lopez-Gonzalez, E., Herdeiro, M. T. & Figueiras, A. Determinants of under-reporting of adverse drug reactions: a systematic review. *Drug Saf.* **32**, 19–31 (2009).
45. Wang, X., Hripcsak, G., Markatou, M. & Friedman, C. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *J. Am. Med. Inform. Assoc.* **16**, 328–337 (2009).
An example of how text mining of bulk EHR data can be used to uncover statistical correlations between clinical concepts, specifically between medications and ADEs.
46. Gini, R., Herings, R., Coloma, P. M., Schuemie, M. J. & Trifiro, G. Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring : the EU-ADR Project. *Pharmacoepidemiol. Drug Saf.* **20**, 1–11 (2011).
47. Yao, L., Zhang, Y., Li, Y., Sanseau, P. & Agarwal, P. Electronic health records: implications for drug discovery. *Drug Discov. Today* **16**, 594–599 (2011).
48. Mullins, I. M. *et al.* Data mining and clinical data repositories: Insights from a 667,000 patient data set. *Comput. Biol. Med.* **36**, 1351–1377 (2006).
49. Wright, A., Chen, E. S. & Maloney, F. L. An automated technique for identifying associations between medications, laboratory results and problems. *J. Biomed. Inform.* **43**, 891–901 (2010).
50. Harpaz, R., Chase, H. S. & Friedman, C. Mining multi-item drug adverse effect associations in spontaneous reporting systems. *BMC Bioinformatics* **11** (Suppl. 9), S7 (2010).
51. Swanson, D. R. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect. Biol. Med.* **30**, 7–18 (1986).
52. Tsuruoka, Y., Miwa, M., Hamamoto, K., Tsujii, J. & Ananiadou, S. Discovering and visualizing indirect associations between biomedical concepts. *Bioinformatics* **27**, 111–119 (2011).
53. Oztekin, A., Delen, D. & Kong, Z. J. Predicting the graft survival for heart-lung transplantation patients: an integrated data mining methodology. *Int. J. Med. Inform.* **78**, e84–e96 (2009).
54. Delen, D., Walker, G. & Kadam, A. Predicting breast cancer survivability: a comparison of three data mining methods. *Artif. Intell. Med.* **34**, 113–127 (2005).
55. Kurt, I., Ture, M. & Kurum, A. T. Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert Syst. Appl.* **34**, 366–374 (2008).
56. Valentino-Devries, J. May the best algorithm win. *The Wall Street Journal* [online], http://online.wsj.com/article/SB10001424052748704662604576202392747278936.html#articleTabs%3Darticle (2011).
57. Ohlsson, M., Peterson, C. & Dictor, M. Using hidden Markov models to characterize disease trajectories. *Proc. Neural Networks and Expert Systems in Medicine and Healthcare Conference* **2001**, 324–326 (2001).
58. Chen, L. L., Blumm, N., Christakis, N. A, Barabási, A.-L. & Deisboeck, T. S. Cancer metastasis networks and the prediction of progression patterns. *Br. J. Cancer* **101**, 749–758 (2009).
59. Fu, T.-C. A review on time series data mining. *Eng. Appl. Artif. Intell.* **24**, 164–181 (2011).
60. Cao, H., Melton, G. B., Markatou, M. & Hripcsak, G. Use abstracted patient-specific features to assist an information-theoretic measurement to assess similarity between medical cases. *J. Biomed. Inform.* **41**, 882–888 (2008).
61. Melton, G. B. *et al.* Inter-patient distance metrics using SNOMED CT defining relationships. *J. Biomed. Inform.* **39**, 697–705 (2006).
62. Murphy, S. *et al.* Instrumenting the health care enterprise for discovery research in the genomic era. *Genome Res.* **19**, 1675–1681 (2009).
63. 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).
A thorough description of the architecture and capabilities of the i2b2 research platform for biomedical research based on EHR data.
64. McCarty, C. A. *et al.* The eMERGE Network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med. Genomics* **4**, 13 (2011).
65. Kho, A. N. *et al.* Electronic medical records for genetic research: results of the eMERGE consortium. *Science Transl. Med.* **3**, 79re1 (2011).
66. Schildcrout, J. S. *et al.* An analytical approach to characterize morbidity profile dissimilarity between distinct cohorts using electronic medical records. *J. Biomed. Inform.* **43**, 914–923 (2010).
67. 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).
The i2b2 platform put to use for case–control generation and study design based on EHR and DNA data in a rheumatoid arthritis project.

68. Kullo, I. J. *et al.* Complement receptor 1 gene variants are associated with erythrocyte sedimentation rate. *Am. J. Hum. Genet.* **89**, 131–138 (2011).
69. 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**, 9 (2010).
70. Denny, J. C. *et al.* Variants near *FOXE1* are associated with hypothyroidism and other thyroid conditions: using electronic medical records for genome- and phenome-wide studies. *Am. J. Hum. Genet.* **89**, 529–542 (2011).
71. Ritchie, M. D. *et al.* Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *Am. J. Hum. Genet.* **86**, 560–572 (2010).
72. Perlis, R. H. *et al.* Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. *Psychol. Med.* **42**, 41–50 (2012).
73. Kho, A. N. *et al.* Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. *J. Am. Med. Inform. Assoc.* **19**, 212–218 (2011).
74. Himes, B. E., Dai, Y., Kohane, I. S., Weiss, S. T. & Ramoni, M. F. Prediction of chronic obstructive pulmonary disease (COPD) in asthma patients using electronic medical records. *J. Am. Med. Inform. Assoc.* **16**, 371–379 (2009).
75. 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 description of the technical, scientific and legal aspects of the development of an EHR–DNA linked research database with an opt-out consent model.**
76. Denny, J. C. *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* **26**, 1205–1210 (2010). **A demonstration of how EHR data linked with DNA data can be used in a reversal of the normal GWAS approach to search for disease phenotypes associated with SNPs.**
77. Wilke, R. *et al.* The emerging role of electronic medical records in pharmacogenomics. *Clin. Pharmacol. Ther.* **89**, 379–386 (2011).
78. Al Mallah, A., Guelpa, P., Marsh, S. & van Rooij, T. Integrating genomic-based clinical decision support into electronic health records. *Personalized Med.* **7**, 163–170 (2010).
79. McCarty, C. A. & Wilke, R. A. Biobanking and pharmacogenomics. *Pharmacogenomics* **11**, 637–641 (2010).
80. Schwarz, U. I. *et al.* Genetic determinants of response to warfarin during initial anticoagulation. *N. Eng. J. Med.* **358**, 999–1008 (2008).
81. Onitilo, A. *et al.* Estrogen receptor genotype is associated with risk of venous thromboembolism during tamoxifen therapy. *Breast Cancer Res. Treat.* **115**, 643–650 (2009).
82. Lage, K. *et al.* Dissecting spatio-temporal protein networks driving human heart development and related disorders. *Mol. Syst. Biol.* **6**, 1–9 (2010).
83. Greenblum, S., Turnbaugh, P. J. & Borenstein, E. Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proc. Natl Acad. Sci. USA.* **109**, 594–599 (2012).
84. Rzhetsky, A., Wajngurt, D., Park, N. & Zheng, T. Probing genetic overlap among complex human phenotypes. *Proc. Natl Acad. Sci. USA.* **104**, 11694–11699 (2007).
85. Goh, K.-I. et al. The human disease network. *Proc. Natl Acad. Sci. USA.* **104**, 8685–8690 (2007).
86. Park, J., Lee, D.-S., Christakis, N. A. & Barabási, A.-L. The impact of cellular networks on disease comorbidity. *Mol. Syst. Biol.* **5**, 262 (2009).
87. Ashley, E. A. *et al.* Clinical assessment incorporating a personal genome. *Lancet* **375**, 1525–1535 (2010).
88. Hood, L., Heath, J. R., Phelps, M. E. & Lin, B. Systems biology and new technologies enable predictive and preventative medicine. *Science* **306**, 640–643 (2004).
89. Galas, D. J. & Hood, L. Systems biology and emerging technologies will catalyze the transition from reactive medicine to predictive, personalized, preventive and participatory (P4) medicine. *Interdisciplinary Bio Central* **1**, 6 (2009).
90. Hall, M. A. Property, privacy, and the pursuit of interconnected electronic medical records. *Iowa Law Review* **2010**, 631–663 (2010).
91. 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**, 77–81 (2009).
92. Kho, M. E., Duffett, M., Willison, D. J., Cook, D. J. & Brouwers, M. C. Written informed consent and selection bias in observational studies using medical records: systematic review. *BMJ* **338**, 1–8 (2009).
93. Hoffman, S. Balancing privacy, autonomy, and scientific needs in electronic health records research. *Case Research Paper Series in Legal Studies* [online], http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1923187 (2011). **An extensive summary of legal and ethical issues encountered in health research and their potential consequences for conducting scientific research.**
94. Meystre, S. M., Friedlin, F. J., South, B. R., Shen, S. & Samore, M. H. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Med. Res. Methodol.* **10**, 1–16 (2010).
95. Benitez, K. & Malin, B. Evaluating re-identification risks with respect to the HIPAA privacy rule. *J. Am. Med. Inform. Assoc.* **17**, 169–177 (2010).
96. Heeney, C., Hawkins, N., de Vries, J., Boddington, P. & Kaye, J. Assessing the privacy risks of data sharing in genomics. *Public Health Genomics* **14**, 17–25 (2011).
97. Homer, N. *et al.* Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet.* **4**, e1000167 (2008).
98. Malin, B. & Sweeney, L. Re-identification of DNA through an automated linkage process. *Proc. AMIA Symp.* **2001**, 423–427 (2001).
99. Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473**, 174–180 (2011).
100. Rothstein, M. A. Is deidentification sufficient to protect health privacy in research? *Am. J. Bioeth.* **10**, 3–11 (2010).
101. Begoyan, A. An overview of interoperability standards for electronic health records. In *Integrated Design and Process Technology (IDPT-2007)* (2007).
102. Goossen, W., Goossen-Baremans, A. & van der Zel, M. Detailed clinical models: a review. *Healthc. Inform. Res.* **16**, 201–214 (2010). **An introduction to modelling and representation of clinical concepts and meaning, which is important for data interoperability.**
103. Knaup, P., Bott, O., Kohl, C., Lovis, C. & Garde, S. Electronic patient records: moving from islands and bridges towards electronic health records for continuity of care. *Yearb. Med. Inform.* **2007**, 34–46 (2007).
104. Garde, S., Knaup, P., Hovenga, E. & Heard, S. Towards semantic interoperability for electronic health records. *Methods Inf. Med.* **46**, 332–343 (2007).
105. Wicks, P., Vaughan, T. E., Massagli, M. P. & Heywood, J. Accelerated clinical discovery using self-reported patient data collected online and a patient-matching algorithm. *Nature Biotech.* **29**, 411–414 (2011).
106. Aronson, R. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc. AMIA Symp.* **2001**, 17–21 (2001).
107. 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).
108. Uzuner, O. Recognizing obesity and comorbidities in sparse data. *J. Am. Med. Inform. Assoc.* **16**, 561–570 (2009).
109. Uzuner, O., Solti, I. & Cadag, E. Extracting medication information from clinical text. *J. Am. Med. Inform. Assoc.* **17**, 514–518 (2010).
110. Uzuner, O., South, B. R., Shen, S. & Duvall, S. L. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J. Am. Med. Inform. Assoc.* **18**, 552–557 (2011).
111. Fung, K. W., McDonald, C. & Bray, B. E. RxTerms - a drug interface terminology derived from RxNorm. *Proc. AMIA Symp.* **2008**, 227–231 (2008).
112. Steindel, S. J. International classification of diseases, 10th edition, clinical modification and procedure coding system: descriptive overview of the next generation HIPAA code sets. *J. Am. Med. Inform. Assoc.* **17**, 274–282 (2010).

Acknowledgements
We thank U. Buddrus for kindly providing unpublished data on the adoption of EHR systems in Europe. Any errors in communicating these insights are the sole responsibility of the authors. The authors were supported in part by the Villum Kann Rasmussen Foundation, the Novo Nordisk Foundation and the Danish Research Council for Strategic Research.

Competing interests statement
The authors declare no competing financial interests.