Citation: CPT Pharmacometrics Syst. Pharmacol. (2017) 6, 188–196; doi:10.1002/psp4.12174

ORIGINAL ARTICLE

CATTLE (CAncer Treatment Treasury with Linked Evidence): An Integrated Knowledge Base for Personalized Oncology Research and Practice

E Soysal, H-J Lee, Y Zhang, L-C Huang, X Chen, Q Wei, W Zheng, JT Chang, T Cohen, J Sun and H Xu*

Despite the existence of various databases cataloging cancer drugs, there is an emerging need to support the development and application of personalized therapies, where an integrated understanding of the clinical factors and drug mechanism of action and its gene targets is necessary. We have developed CATTLE (CAncer Treatment Treasury with Linked Evidence), a comprehensive cancer drug knowledge base providing information across the complete spectrum of the drug life cycle. The CATTLE system collects relevant data from 22 heterogeneous databases, integrates them into a unified model centralized on drugs, and presents comprehensive drug information via an interactive web portal with a download function. A total of 2,323 unique cancer drugs are currently linked to rich information from these databases in CATTLE. Through two use cases, we demonstrate that CATTLE can be used in supporting both research and practice in personalized oncology.
CPT Pharmacometrics Syst. Pharmacol. (2017) **6**, 188–196; doi:10.1002/psp4.12174; published online 0 Month 2017.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

☑ Cancer drug knowledge bases exist; but few of them cover information across the full spectrum of drug development, especially textual data.

WHAT QUESTION DID THIS STUDY ADDRESS?

☑ This study addresses the issue of lack of comprehensive cancer drug knowledge bases that can support both research and practice of personalized cancer therapy.

WHAT THIS STUDY ADDS TO OUR KNOWLEDGE

☑ This study developed a cancer drug knowledge base that provides linked evidence across the full spectrum of drug development and the heterogeneous data sources, to support both research and practice of personalized cancer therapy.

HOW THIS MIGHT CHANGE DRUG DISCOVERY, DEVELOPMENT, AND/OR THERAPEUTICS

☑ The CATTLE system collects, integrates, and visualizes a broad range of cancer drug-related information, thus enabling further computation analysis for drug discovery/development, as well as addressing the information needs during practice.

Cancer is a leading cause of death worldwide. Tremendous efforts have been devoted to develop effective cancer therapies, resulting in thousands of drugs that have been approved or are under investigation in clinical trials. A recent trend in cancer drug development is to provide personalized cancer therapy (also known as personalized oncology), which involves using patients' tumor genomic information to optimize therapies.[1] For example, many new cancer clinical trials are now based on targeted therapies, where drugs interfere with specific molecules related to cancer growth and survival. As cancer is a complex disease affected by genetic, environmental, and clinical attributes, the development of a personalized cancer therapy is a complicated and time-consuming process, involving a broad range of biomedical research from basic biological through clinical sciences.[2] During this process, various types of data (housed in repositories) are generated, ranging from databases containing raw experimental results (e.g., PubChem bioassays[3] and Connectivity Map[4–6]), knowledge bases about genetic variations of cancer cells and cancer drugs (e.g., canSAR[7] and PharmGKB[8,9]), to textual data collections such as published literature, patents, and clinical trials documents. Thus, it is very challenging for biological or clinical researchers to keep up-to-date with such broad and diverse information about personalized cancer therapy.[10]

Fortunately, the biomedical research community has been aware of this challenge and a substantial amount of effort has been devoted to building integrated drug resources from heterogeneous data sources.[6,8,9,11,12] Such efforts include general drug knowledge bases such as DrugBank, which provides summarized information on small molecule drugs, including chemical structures, pharmacokinetics and pharmacodynamics, drug–drug interactions, side effects, etc., as well as genes involved in the drugs' pharmacological actions.[13] The SIDER is comprehensive database for drug side effects.[14] The Kyoto Encyclopedia of Genes and Genomes (KEGG) DRUG provides drug target, metabolizing enzyme, and other molecular interaction network information for approved drugs in Japan, the USA, and Europe.[15,16] The Chemical European Molecular Biology Laboratory (ChEMBL) is a database of bioactive drug-like small molecules with abstracted bioactivities (e.g., binding

University of Texas Health Science Center at Houston, Houston, Texas, USA. *Correspondence to: H Xu (hua.xu@uth.tmc.edu)
Received 9 December 2016; accepted 17 January 2017; published online on 0 Month 2017. doi:10.1002/psp4.12174

constants and pharmacology data).[17,18] The Therapeutic Target Database (TTD) focuses on drugs and their protein and nucleic acid targets.[19,20] PubChem Compound provides information of pharmacology, biochemistry, toxicity, biomolecular interactions, and pathways for chemicals.[3] In the cancer domain, canSAR is an extensive data source with interconnected information on cancers, drugs, and genes.[7,21] The Genomics of Drug Sensitivity in Cancer (GDSC) database concentrates on specific aspects of cancer drugs, namely, drug sensitivity in cancer cells and molecular markers of drug responses.[22,23] CTdatabase[24] and CancerPPD[25] emphasize cancer-testis antigens and anticancer peptides/proteins, respectively. The Personalized Cancer Therapy Knowledge Base for Precision Oncology focuses on the associations between genomic alterations with tumor development, growth, response to therapy, and available therapies.[26] However, despite the extensive contribution of current cancer drug resources, there are several limitations. First, they are often limited to a few related data sources, and do not cover the full spectrum of drug development (e.g., from preclinical studies to postmarket surveillance). Furthermore, existing resources often have limited links to rich textual evidential data such as biomedical literature and clinical trial documents. Therefore, there is an urgent need for an integrated, comprehensive cancer drug resource that biomedical scientists and clinicians can use to search across existing, scattered data for research and practice in personalized cancer therapy.

This article introduces CATTLE (CAncer Treatment Treasury with Linked Evidence), an integrated knowledge base for cancer drugs, with the goal to support personalized oncology research and practice. The scope of the CATTLE project includes 1) collection of diverse types of data generated through the complete spectrum of drug development (e.g., from basic biological and preclinical research to clinical trials and postmarketing surveillance); 2) integration into a drug-centered, unified, comprehensive knowledge base; and 3) public availability for use in research and clinical practice via a user-friendly web interface. In the current release, CATTLE contains 2,323 normalized cancer drugs, with linked data to 22 major data resources, including raw experimental databases (e.g., PubChem[27]), general and cancer-specific knowledge bases (e.g., DrugBank[13] and GDSC[22,23]), and textual data collections (e.g., PubMed and ClinicalTrial.gov[28]), that range from biological and preclinical research data to postmarket surveillance data (e.g., the US Food and Drug Administration's (FDA) AERS[29]). To the best of our knowledge, CATTLE is the first integrated, comprehensive knowledge base covering data sources from the full spectrum of cancer drug development, and we believe that such a knowledge base will greatly benefit research and practice of personalized cancer therapy, by complementing existing databases.

METHODS

Figure 1 shows an overview of the workflow for building CATTLE. It consists of three steps: 1) Data Collection: we carefully decided on data sources that are highly relevant to cancer drug research and collected data from each source; 2) Data Integration: we built a normalized drug list and linked data from all the sources about each drug in the list, using various technologies such as record linkage algorithms, natural language processing (NLP), and manual review; and 3) Data Visualization: we developed a publicly available web portal (www.drugkb.org) that provides not only user-friendly exploring functions, but also batch downloading functions for further computational analysis. The three steps are described in the following sections.

Data collection

Many resources that contain information relevant to cancer drugs have been developed. It is not possible to include all such relevant resources at the beginning. In the first version, our research team manually identified 22 highly relevant data sources based on several inclusion criteria such as having a high impact and containing general or cancer-specific drug information. **Table 1** lists the 22 selected resources with a brief description about the data in each resource. For each resource, we developed ETL (Extract, Transform, and Load) tools to extract raw data into a local database. One benefit of the ETL tools is that it makes the maintenance easier, allowing us to efficiently keep the data up-to-date. For some data sources that provide APIs for real-time access, we do not keep another copy of the data locally. For example, we query PubMed abstracts on the fly using the Entrez Programming Utilities (eUtils), without keeping a local copy of MEDLINE.

Data integration

Different data sources may use different drug naming systems and have different drug IDs, which is the main challenge for data integration. Our data integration approach consisted of two steps: 1) we compiled a list of normalized cancer drugs of interest; and 2) we retrieved additional information for each drug by linking it to records in different data sources.

To compile a list of cancer drugs of interest, we limited the scope to small molecule drugs that are either approved by the FDA or reported in clinical trials for cancer treatment. We retrieved the FDA-approved cancer drugs from the FDA online label repository (http://labels.fda.gov/). Cancer drugs in clinical trials were collected from ClinicalTrials.gov (http://clinicalTrials.gov), by a semiautomated method that was developed to retrieve cancer treatment trials.[30] In total, we collected 146 cancer drugs from the FDA and 2,916 drugs from clinical trials. As different drug names (e.g., Tretinoin and Panretin) may refer to the same molecule, we further normalized the drug names in the list by mapping them to external drug terminologies. We first mapped each drug name in the list to IDs in various drug terminologies, including DrugBank,[13] National Cancer Institute Thesaurus (NCIT),[31] PubChem Compound/Substance,[3] Chemical Entities of Biological Interest (ChEBI),[32] ChEMBL, TTD, RxNorm,[33] Unified Medical Language System (UMLS),[34] Medical Subject Headings (MeSH),[35] and Anatomical Therapeutic Chemical classification (ATC).[36] If two drug names shared any IDs in these terminologies, we paired them and asked domain experts to manually review them to determine if these two drugs should be merged into a single entry in CATTLE. After normalization, there were 2,323 unique drugs (146 FDA-approved, 2,177

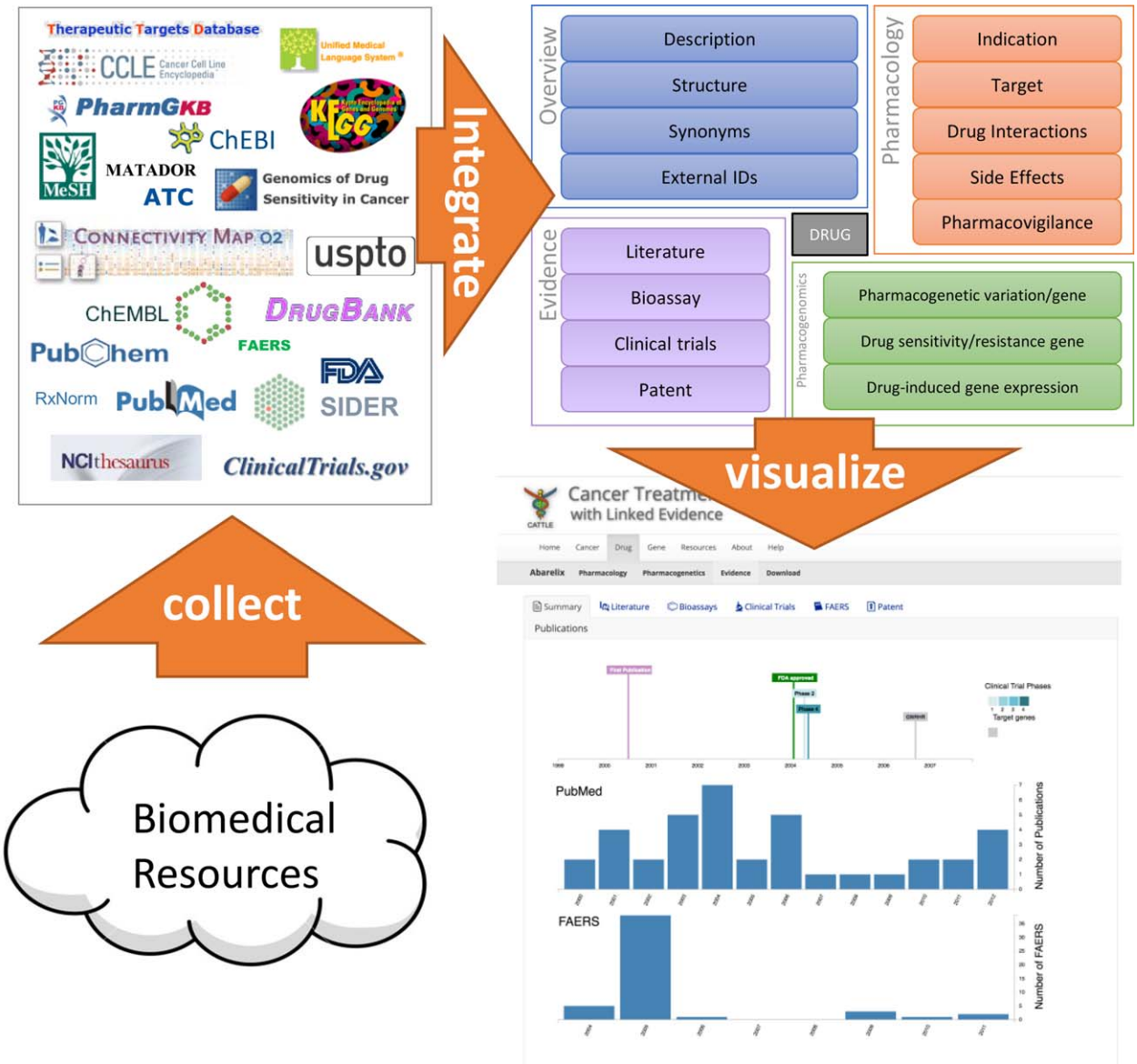

Figure 1 An overview of the workflow for building CATTLE: data collection, integration, and visualization.

from clinical trials), which were each assigned a unique ID. A side benefit of this normalization process was that we also compiled a list of synonyms for each drug in the final list.

After normalizing the drug names to 2,323 unique entries, we started retrieving various types of information about the drugs from each of the data sources described in **Table 1,** to build a unified database. Three different technologies: record linkage algorithms, NLP, and manual review were used to link drugs to data sources. For most of the databases, a record linkage algorithm[37] that compares similarity among key fields (i.e., drug names, synonyms, and IDs) between two records was used to determine the linkage. For textual data sources, e.g., patents, we applied previously developed NLP methods in our lab[38] to recognize chemical/drug names in the documents and then link them to the drugs in the list. Whenever the information was uncertain or the source was not suitable for automated processing, we conducted manual review to ensure the correctness of the linkage. For example, drug description in the overview section of each FDA-approved drug was extracted solely based on the manual process, since FDA drug labels were in PDF format, without any reliable markers for start and end of description. We also generated structure information using third-party tools such as NCI/CADD online tool (http://cactus.nci.nih.gov/gifcreator/) or SWISS-MODEL (http://swissmodel.expasy.org/).

Data visualization

A website was developed (http://www.drugkb.org) to display all the integrated information collected from different knowledge bases summarized in **Table 1.** Users can query the CATTLE database using drug, cancer, or gene names to

Table 1 Data sources used in the CATTLE (CAncer Treatment Treasury with Linked Evidence) database

Resource	Link	Contribution	# of Records
FDA drug label	http://www.accessdata.fda.gov/scripts/cder/drugsatfda/	Description, Indication	6,825 drugs
NCIT	http://ncit.nci.nih.gov/	Description, Synonyms, External IDs	121,794 Records
PubChem	https://pubchem.ncbi.nlm.nih.gov/	Structure, Synonyms, External IDs, Bioassay	1.1 million bioassays, 90 million compounds, 220 million substance
ChEMBL	https://www.ebi.ac.uk/chembl/	Structure, External IDs	1,3 million Chemicals
TTD	http://bidd.nus.edu.sg/group/TTD/ttd.asp	Structure, External IDs, Indication, Drug Targets	2,025 Targets, 17,816 Drugs
MeSH	http://www.ncbi.nlm.nih.gov/mesh	Synonyms, External IDs	822,268 Concepts
DrugBank	http://www.drugbank.ca/	External IDs, Indication, Drug interaction	8,246 Drugs, 4,170 Targets, 15,438 Drug-Target associations
KEGG	http://www.genome.jp/kegg/	Pathway, Indication, Drug interaction, Drug Targets	10 thousand Drugs, 20 million Genes, 501 Pathways
ChEBI	https://www.ebi.ac.uk/chebi/	Synonyms, External IDs	50 thousand Compounds
RxNorm	http://www.nlm.nih.gov/research/umls/rxnorm/	Synonyms, External IDs	630 thousand Records
UMLS	http://www.nlm.nih.gov/research/umls/	Synonyms, External IDs	7.4 million Concept
ATC	http://www.whocc.no/atc_ddd_index/	External IDs, drug class	6,601 Drugs
MATADOR	http://matador.embl.de/	Drug targets	15 thousand Drug-Target Association
SIDER	http://sideeffects.embl.de/	Side effects	5,868 Side Effects of 1,430 Drugs with 139 thousand Associations.
FAERS	http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/	Pharmacovigilance	4,070,067 Reports
PharmGKB	https://www.pharmgkb.org/	Pharmacogenetic variation/genes	24,784,816 Records
CCLE	http://www.broadinstitute.org/ccle/home	Pharmacogenetic variation/genes	23,390 Records
GDSC	http://www.cancerrxgene.org/	Drug sensitivity/resistance genes	Total number of IC50 values 224,510
cMAP	https://www.broadinstitute.org/cmap/	Drug-induced gene expression	1.5M gene expression profiles from ~5,000 small-molecule compounds, and ~3,000 genetic reagents
PubMed	http://www.ncbi.nlm.nih.gov/pubmed	Literature	24.6 million
ClinicalTrials.gov	https://clinicaltrials.gov/	Clinical trials, Indications	231,860 Trials
USPTO	http://www.uspto.gov/	Patents	326,032 Patents

FDA, Food and Drug Administration; NCIT, National Cancer Institute Thesaurus; ChEMBL, Chemical Database of Bioactive Molecules by European Bioinformatics Institute-European Molecular Biology Laboratory; TTD, Theurapeutic Target Database; MeSH, Medical Subject Headings; KEGG, Kyoto Encyclopedia of Genes and Genomes; ChEBI, Chemical Entities of Biological Interest; RxNorm, Normalized naming system for generic and branded drugs; UMLS, Unified Medical Language System; ATC, Anatomical Therapeutic Chemical Classification System; MATADOR, Manually Annotated Targets and Drugs Online Resource; SIDER, Side Effect Resource; CCLE, Cancer Cell Line Encyclopedia; GDCS, Genomics of Drug Sensitivity in Cancer; cMAP, Connectivity Map O2; USPTO, United States Patent and Trademark Office.

retrieve required information. Since the principal focus of the site is to provide information on cancer treatments, the web interface is organized primarily around a drug.

Users can find a drug by searching through the search box or by selecting from a comprehensive list of cancer medications that is listed alphabetically. As drugs at different development stages have different types of evidence, we explicitly marked drugs either with CT (for clinical trials), or FDA (for FDA-approved drugs), to indicate the source of the information. Users can access details of the individual drug by clicking its name. Information on a drug is organized in several pages. The "overview" page provides a general summary of the drug including the description, molecular structure, synonymous names, and other identifiers for the drug in the source knowledge bases with direct web links to the corresponding resource. The "pharmacology" page has drug target information, indications, known drug–drug interactions, as well as side effects of the selected drug. Information related to genetic aspects of the drugs is presented in the "pharmacogenomics" page, which covers pharmacogenetic genes and variations derived data from PharmGKB and CCLE. This page also contains information related to drug sensitivity and resistance—genes as well as drug-induced gene expression data.

Linked textual evidence is organized as multiple subpages. A summary page contains a timeline that visually presents major events across the full spectrum of drug development and research (i.e., important publications, clinical trials at different phases, FDA approvals, etc.). **Figure 2** shows an example of the timeline of the drug Abarelix, which highlights important events such as clinical trials, target genes, and FDA approvals. By clicking the individual events, users can find more details of the linked evidence. In addition, CATTLE also provides bar graphs that summarize the numbers of publications and the FDA adverse event reports over time. To display results in the "literature" subpage, we search PubMed

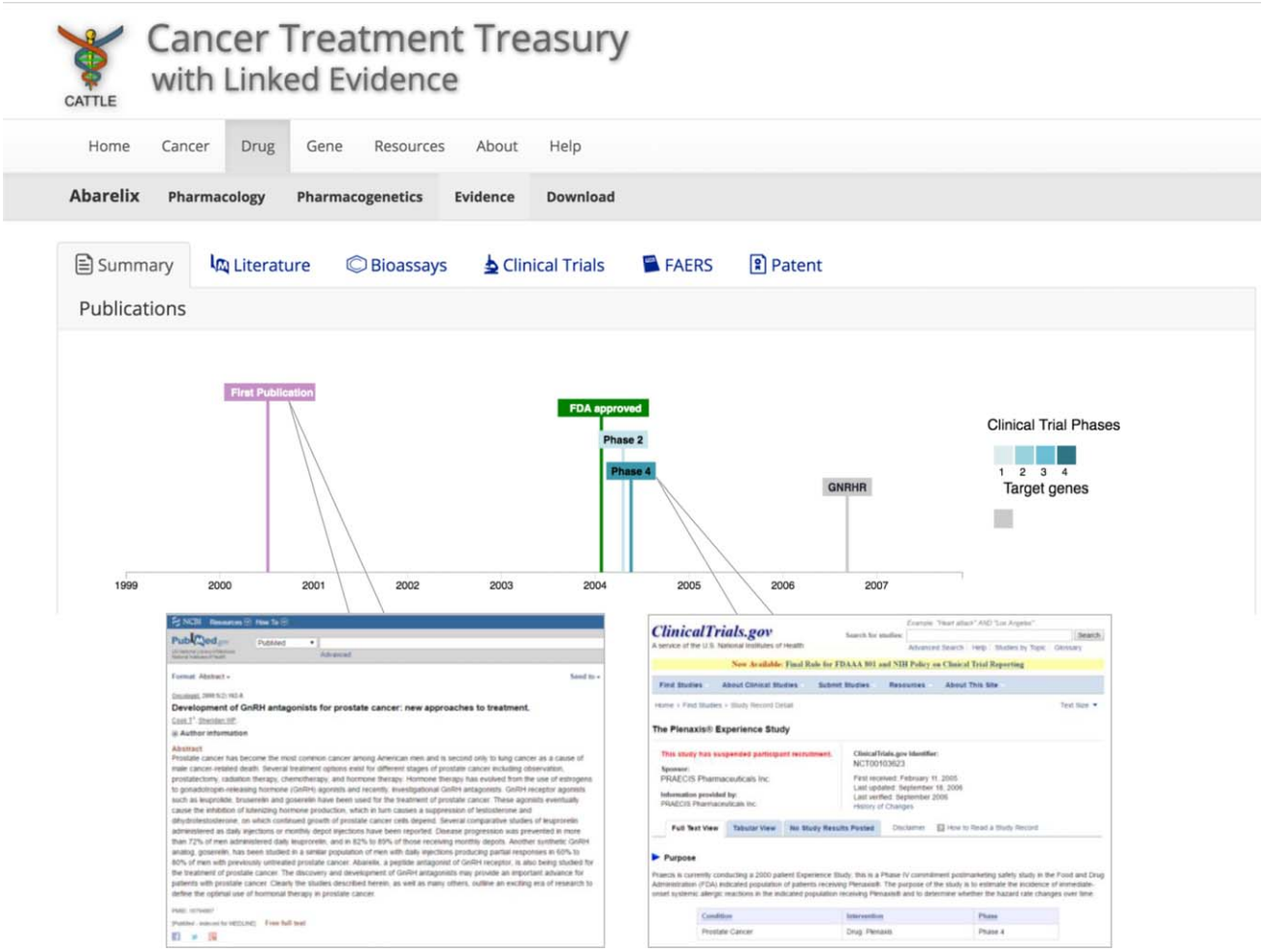

Figure 2 An example of visualization of integrated information across the full spectrum of drug development.

for the drug using its name and synonyms dynamically. This is accomplished using the EUtils tool provided by the NCBI with real-time queries to PubMed APIs. This approach ensures access to up-to-date publications directly from the source. We also use the same services to capture bioactivity screens for the drugs from the PubChem BioAssay to provide access to the most recent data. Related clinical trials and the FDA AERS reports are also provided to support different use case scenarios. Additionally, patent information related to each drug that was collected from the United States Patent and Trademark Office (USPTO) records are provided.

Another important feature of the CATTLE website is to provide downloading functionality for the users. For each drug, a link is provided to download all available information in the CATTLE database in the XML format. In addition, users can select a list of drugs of interest (e.g., all lung cancer drugs) and download all the information of these drugs in a batch both in XML and tab separated values (TSV) formats. This function provides a convenient way for researchers who want to download the data locally for further computational analysis.

In addition to searching drugs, the CATTLE website also allows users to search for genes or cancers. Genes are organized based on their primary function in the cell such as DNA repair or cell–cell communication, and each gene is connected to related drugs and external references such as Uniprot or Genetics Home for additional information. Cancer information is organized using the International Classification of Disease Oncology v. 3 (ICD-O-3), with the topology and morphology axes for clinical relevance. A cancer introduction page also lists the most common cancer topology and morphologies in clinical practice for convenience. Moreover, CATTLE provides a browsing function that allows users to quickly look through information within genes and cancers organized by subcategories such as gene functions and cancer morphology. In addition, it also provides linkages among these different types of information. For example, one can quickly find drug treatment information for a cancer subtype. Such information is collected by linking different knowledge sources, e.g., Drugbank, Clinical Trials, and FDA labels.

RESULTS

In the current version, CATTLE has collected information from 22 large databases with a total over 341 million data

Table 2 Distribution of CATTLE drugs in different phases of clinical trials, and clinical trials in different phases

Drug development stage		# of drugs	%	# of trials	%
	Unknown	31	1.33	1,691	3.42
Exploratory study	Phase 0	6	0.26	129	0.26
Safety and effectiveness study	Phase I	646	27.80	8,898	17.99
	Phase I/phase II	214	9.21	4,926	9.96
	Phase II	682	29.36	23,872	48.27
	Phase II/phase III	39	1.68	717	1.45
	Phase III	446	19.20	8,304	16.79
Postmarket study	Phase IV	258	11.11	917	1.85
	Total	2,323	100.00	49,454	100.00

records (**Table 1**). After integration, the CATTLE database contains information on 2,323 cancer drugs and 8,414 genes (1,984 drug target genes, 412 genes affecting cancer sensitivity to these cancer drugs, and 6,474 variants of 186 genes in tumor cells). The numbers of cancer drugs in different phases of clinical trial are summarized in **Table 2.** Furthermore, **Figure 3** illustrates **(a)** the numbers of drugs and genes per cancer topology for some common cancers, **(b)** the numbers of drugs and genes per cancer morphology for some frequent cancer types, and **(c)** the numbers of drugs that are related to genes in each gene category. In addition, drugs with most linked evidence (top 5) are shown in **Table 3.**

Use cases

To further demonstrate the utility of CATTLE, we describe two use cases of CATTLE.

Use Case 1: Identify potential drug repurposing signals by computing target similarity. As mentioned previously, users can download a list of drugs of interest with all the integrated information from CATTLE, thus enabling further computational analysis for drug research. One such example is to conduct drug repurposing signal analysis, which aims to find new indications for existing drugs. Among many computational drug repurposing approaches, the common target-based approach, which assumes that the drugs sharing common targets could be used to treat the same disease, is one of the "guilt-by-association" strategies commonly used to identify novel drug–disease associations.[39] Here we demonstrate how to utilize data collected in CATTLE to conduct such studies. We first identified all FDA-approved cancer drugs (146) and then downloaded all related information of these drugs, including drug targets, indications, and clinical trials information from

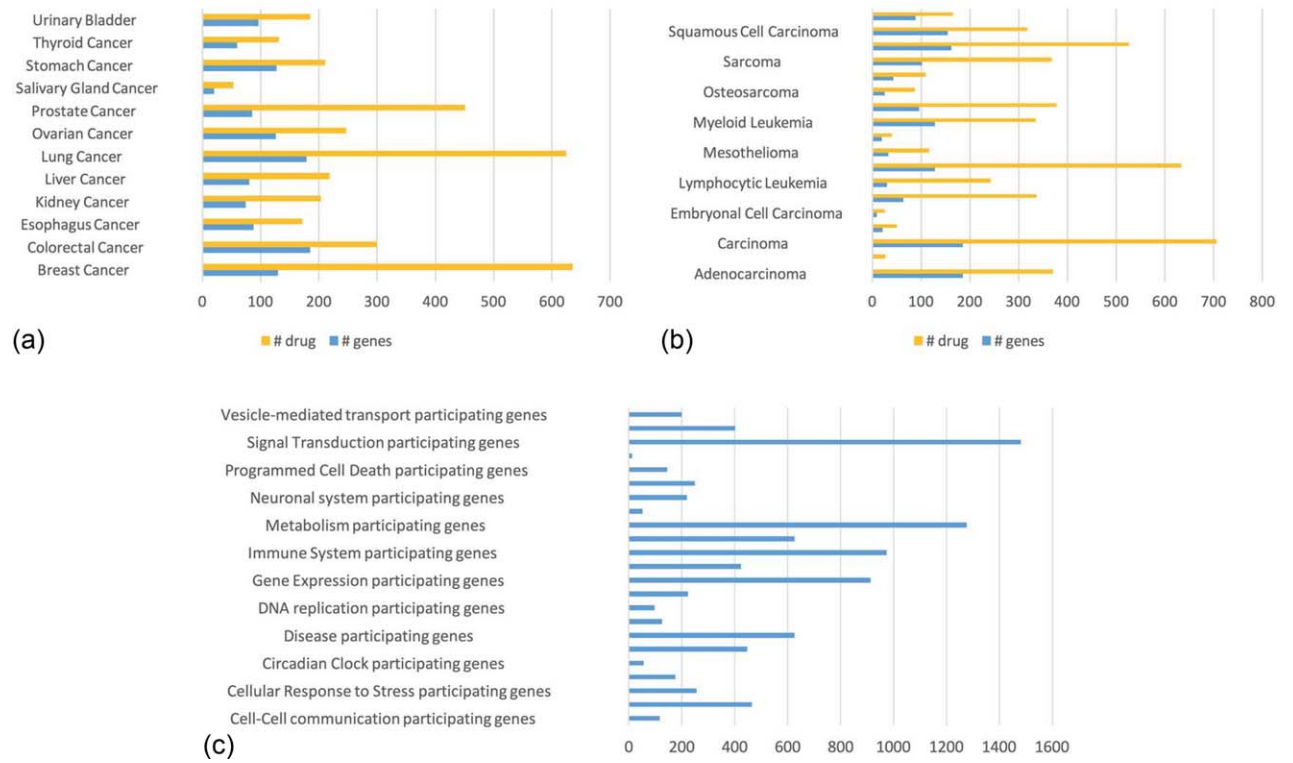

Figure 3 (**a**) number of drugs and genes per cancer type (by topology); (**b**) number of drugs and genes per cancer type (by morphology); (**c**) number of drugs related to each gene category.

Table 3 Top 5 drugs with the most number of evidence documents for each evidence type

Evidence type	Drug name	# of associated docs
Top 5 Literatures	Mycophenolate	107,155
	Estrogen	61,558
	Vascular Endothelial Growth Factor Receptor 2	59,006
	Cisplatin	49,742
	Epinephrine	42,496
Top 5 Bioassays	Ciprofloxacin	12,527
	Doxorubicin	7,377
	Indomethacin	6,189
	Fluconazole	5,386
	Amphotericin B	5,117
Top 5 Clinical Trials	Cyclophosphamide	1,716
	Cisplatin	1,668
	Paclitaxel	1,397
	Gemcitabine	1,335
	Carboplatin	1,325
Top 5 Patents	Dimethicone	20,553
	Phenobarbital	11,720
	Polyethylene Glycol	8,688
	Omeprazole	5,753
	Methylphenidate	4,147

CATTLE. Then we created a drug–target and drug–indication network, which contains 756 nodes (146 drugs, 572 targets, and 36 cancer types) and 1,347 edges (237 drug–cancer associations and 1,110 drug–target associations). By applying the common target-based approach, we predicted 51 novel drugs for treating lung cancer. To evaluate the predicted drug repurposing signals, we utilized the clinical trial information that we downloaded from CATTLE, by checking whether the predicted drug has been investigated in any trials for treating lung cancer. After searching clinical trial information in CATTLE, we found that 31 of the 51 predicted drugs had been investigated in at least one clinical trial for treating lung cancer, indicating the promise of such approaches, as well as the usefulness of CATTLE in such computational studies for drug research.

Use Case 2: The use of CATTLE in personalized oncology. Targeted therapy involves the use of genomic profiling to identify aberrations in genes that drive the progression of an individual's tumor. These aberrations may sensitize tumors to drugs that specifically target that gene, enabling the personalization of therapeutic regimens for precision oncology. Such genomic profiling information is frequently available to oncologists. However, the information required to guide therapy in the context of a particular profile is embedded in the biomedical literature and other sources, such as descriptions of ongoing clinical trials. To render this information clinically actionable, scientist-curators must review a broad array of resources, including the biomedical literature, that describe the characterization of effects of pharmaceutical agents at the molecular level.[40]

With its combination of automated data aggregation and manual curation, the CATTLE database serves as a unique resource for this purpose. Drug–target information present in CATTLE can be used to guide the literature search of curators by providing links to supporting evidence. Closer to the bedside, CATTLE's clinical trial-related information can be used to broaden the range of therapeutic alternatives for a patient to include experimental drugs, based on the patient's genomic profile. At large cancer centers such as MD Anderson (MDA), there are over hundreds of phrase I/II clinical trials ongoing every year, of which many are target-based therapies that rely on patients' genetic characteristics. Oncologists face the problem to determine the trials that will benefit the patient, mostly based on genetic alteration of tumors. We have started working with collaborators at MDA to utilize CATTLE for trial selection. Through a prior study,[30] CATTLE can highlight requirements about genetic alteration in the eligible criteria section of trials, thus to facilitate oncologists quickly finding more appropriate trials for patients with certain genetic alteration.

From a preclinical perspective, it is common practice to evaluate the activity of large libraries of screening agents for their activity against cancer cell lines in high-throughput screening experiments. However, information about the known activities and targets of agents in such libraries is not widely available in electronic form. Consequently, the interpretation of the results of such experiments often necessitates manual curation of drug–target relationships (see, for example, Seashore-Ludlow *et al.*[41]). CATTLE integrates data from multiple pharmacogenomics knowledge resources, permitting rapid retrieval of known targets of drugs in a library for interpretation of observed activities in single agent and combinatorial screens, and selection of agents for such screening experiments.

DISCUSSION

In this study we developed CATTLE, a comprehensive cancer drug knowledge base, to support the research and clinical practice for personalized cancer therapy. Presently, CATTLE has collected and integrated 22 important data sources, which cover the full spectrum of drug development and research life cycle. In addition, the CATTLE website provides not only search and browsing functions, but also downloading functionality, which allows users to further utilize the data for local computational analysis. To the best of our knowledge, CATTLE is the first cancer drug database that provides linked evidence across the full spectrum of drug development and compiles information from heterogeneous data sources including multiple textual data collections.

As demonstrated by the two use cases, we believe CATTLE will greatly benefit both the research and practice of personalized cancer therapy by bridging bench and the bedside. In the current clinical knowledge ecosystem, medical students mainly obtain basic biological knowledge from textbooks, generating a time gap of several years before such knowledge becomes available to clinical practice.[42] On the other hand, biomedical knowledge is constantly and rapidly growing, making follow-up difficult for clinical

practitioners. Additionally, each cancer case is unique because of individual genomic variations,[43] which makes it even more challenging in the clinical setting. As a comprehensive knowledge base, CATTLE provides access to all available cancer treatment information in one place. Furthermore, the latest evidence about each individual drug is also linked to CATTLE using live queries to large databases such as PubMed and PubChem. From the perspective of the biological research domain, CATTLE provides in-depth information not only from basic sciences, but also the clinical aspects of drug use (e.g., clinical trials and postmarketing surveillance) and conveys the clinical data back to basic medical research. This bilateral information flow improves data availability from diverging sources to further improve research. Overall, the full spectrum of cancer treatment knowledge in CATTLE facilitates narrowing of the gap between basic medical science research and clinical science applications and supports the translational aspect of medical sciences by presenting basic medical science research discoveries to clinical practice.

CATTLE has the potential to provide more valuable use cases in addition to the ones described in the Results section, such as helping clinicians to adjust medical treatment of cancer, based on other information such as drug sensitivity or resistance, and inspiring biological researchers to investigate the mechanisms for novel drug repurposing signals derived from the clinical data. Moreover, with increasing active patient involvement in their therapeutic options, they could also gain valuable cancer treatment knowledge and match themselves to potential clinical trials using CATTLE. Although it has the potential to support clinical decision-making in precision oncology, the current version of CATTLE may have limited uses in clinical practice. For example, it lacks detailed gene–drug associations at the specific variant level, which are often needed for making personalized treatment for patients. Therefore, our next step is to add annotations of more detailed gene–drug associations for clinical precision oncology.

As a comprehensive knowledge base with rich data sources, CATTLE faces the challenge to continuously update its content. Maintenance of such a large knowledge base is not a trivial task. Currently, CATTLE employs several different automated approaches to efficiently keep data up-to-date: 1) use the APIs provided by the data sources to retrieval the most up-to-date information, such as the eUtils by the NCBI. 2) Automated or semiautomated methods developed in-house to keep them up-to-date. For example, a named entity recognition system is built to extract chemicals from patents.[38] Moreover, to collect an accurate clinical trial list for cancer treatment, we have developed a semiautomated framework, in which NLP-based tools are utilized to identify candidates with confidence scores and manual review will be applied when the confidence score is low.[30] Our estimation of required manual review efforts during the process of building CATTLE shows that it is feasible to accommodate the growth of data sources, by conducting updates every 6 months.

CATTLE has several limitations that need to be addressed in our future work. First, the data sources integrated in CATTLE are still limited and need further expansion in the future; second, manual review efforts could be further reduced by developing advanced automated tools for text and data processing using more advanced informatics techniques for assisting manual curation; third, considering that different user groups of CATTLE (e.g., biological researchers, clinicians, patients) may have different information needs, we will carry out an in-depth use case study and customize the data visualization and download services for different users accordingly.

In summary, in order to better support personalized oncology research and practice and complement existing databases, CATTLE takes the initiative to build a comprehensive knowledge base that links evidence across the full spectrum of the drug life cycle, by integrating relevant data from heterogeneous databases into a unified model centralized on drugs. An interactive web interface provides both the search and browsing function as well as the downloading function to fulfill the needs of different user groups. Detailed use cases demonstrate the great benefits of CATTLE in supporting both research and practice in personalized oncology.

Acknowledgment. This project was supported in part by grants R1307 from CPRIT and R01GM102282.

Conflict of Interest/Disclosure. The authors have no conflicts of interest to report.

Author Contributions. H.X., E.S., H-J.L., Y.Z., Q.W., and T.C. wrote the article; H.X., W.Z., J.T.C., and T.C. designed the research; E.S., Y.Z., and J.S. performed the research; H-J.L., L-C.H., X.C., and Q.W. analyzed the data.

1. Hayden, E.C. Personalized cancer therapy gets closer. *Nature* **458**, 131 (2009).
2. Burrell, R.A., McGranahan, N., Bartek, J. & Swanton, C. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* **501**, 338–345 (2013).
3. Bolton, E.E., Wang, Y., Thiessen, P.A. & Bryant, S.H. PubChem: integrated platform of small molecules and biological activities. *Annu. Rep. Comput. Chem.* **4**, 217–241 (2008).
4. Lamb, J. The Connectivity Map: a new tool for biomedical research. *Nat. Rev. Cancer* **7**, 54–60 (2007).
5. Lamb, J. *et al.* The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929–1935 (2006).
6. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
7. Halling-Brown, M.D., Bulusu, K.C., Patel, M., Tym, J.E. & Al-Lazikani, B. canSAR: an integrated cancer public translational research and drug discovery resource. *Nucleic Acids Res.* **40**, D947–D56 (2012).
8. Hewett, M. *et al.* PharmGKB: the pharmacogenetics knowledge base. *Nucleic Acids Res.* **30**, 163–165 (2002).
9. Klein, T. *et al.* Integrating genotype and phenotype information: an overview of the PharmGKB project. *Pharmacogenomics J.* **1**, 167–170 (2001).
10. Searls, D.B. Data integration: challenges for drug discovery. *Nat. Rev. Drug Disc.* **4**, 45–58 (2005).
11. Wishart, D.S. *et al.* DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* **36**(suppl. 1), D901–D906 (2008).
12. Günther, S. *et al.* SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic Acids Res.* **36**(suppl. 1), D919–D922 (2008).
13. Wishart, D.S. *et al.* DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **34**(suppl. 1), D668–D672 (2006).
14. Kuhn, M., Letunic, I., Jensen, L.J. Bork, P. The SIDER database of drugs and side effects. *Nucleic Acids Res.* **gkv1075,** (2015).
15. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
16. Kanehisa, M. *et al.* From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* **34**(suppl. 1), D354–D357 (2006).

17. Bento, A.P. *et al.* The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* **42**, D1083–D1090 (2014).
18. Gaulton, A. *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, D1100–D1107 (2012).
19. Chen, X., Ji, Z.L. & Chen, Y.Z. TTD: therapeutic target database. *Nucleic Acids Res.* **30**, 412–415 (2002).
20. Zhu, F. *et al.* Update of TTD: therapeutic target database. *Nucleic Acids Res.* **38**(suppl. 1), D787–D791 (2010).
21. Bulusu, K.C., Tym, J.E., Coker, E.A., Schierz, A.C. & Al-Lazikani, B. canSAR: updated cancer research and drug discovery knowledgebase. *Nucleic Acids Res.* **42**, D1040–D1047 (2014).
22. Garnett, M.J. *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483**, 570–575 (2012).
23. Yang, W. *et al.* Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.* **41**, D955–D961 (2013).
24. Almeida, L.G. *et al.* CTdatabase: a knowledge-base of high-throughput and curated data on cancer-testis antigens. *Nucleic Acids Res.* **37**(suppl. 1), D816–D819 (2009).
25. Tyagi, A. *et al.* CancerPPD: a database of anticancer peptides and proteins. *Nucleic Acids Res.* **gku892,** (2014).
26. Knowledge Base for Precision Oncology 2016. <https://pct.mdanderson.org>.
27. Chen, B. & Wild, D.J. PubChem BioAssays as a data source for predictive models. *J. Mol. Graph. Model.* **28**, 420–426 (2010).
28. Zarin, D.A., Tse, T., Williams, R.J., Califf, R.M. & Ide, N.C. The ClinicalTrials. gov results database—update and key issues. *N. Engl. J. Med.* **364**, 852–860 (2011).
29. Baum C, Kweder S, Anello C. *The Spontaneous Reporting System in the United States.* Pharmacoepidemiology 125–137 (John Wiley & Sons, New York, 1994).
30. Xu, J. *et al.* Extracting genetic alteration information for personalized cancer therapy from ClinicalTrials. gov. *J. Am. Med. Inform. Assoc.* **ocw009** (2016).
31. Sioutos, N. *et al.* NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. *J. Biomed. Inform.* **40**, 30–43 (2007).
32. Degtyarenko, K. *et al.* ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.* **36**(suppl. 1), D344–D350 (2008).
33. Liu, S., Ma, W., Moore, R., Ganesan, V. & Nelson, S. RxNorm: prescription for electronic drug information exchange. *IT Profess.* **7**, 17–23 (2005).
34. Bodenreider, O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* **32**(suppl. 1), D267–D270 (2004).
35. Lipscomb, C.E. Medical subject headings (MeSH). *Bull. Med. Library Assoc.* **88**, 265 (2000).
36. World Health Organization. *The Anatomical Therapeutic Chemical Classification System With Defined Daily Doses (ATC/DDD)* (WHO, Oslo, 2006).
37. Amir, A., Farach, M. & Matias, Y. Efficient randomized dictionary matching algorithms. In: Apostolico, A., Crochemore, M., Galil, Z., Manber, U., editors. Combinatorial Pattern Matching: Third Annual Symposium Tucson, Arizona, April 29-May 1, 1992. Proceedings, 262–275 (Springer, Berlin, Heidelberg, 1992).
38. Zhang, Y. *et al.* Chemical named entity recognition in patents by domain knowledge and unsupervised feature learning. *Database* **2016**, baw049 (2016).
39. Hodos, R.A., Kidd, B.A., Shameer, K., Readhead, B.P. & Dudley, J.T. In silico methods for drug repurposing and pharmacology. *Wiley Interdisc. Rev. Syst. Biol. Med.* **8**, 186–210 (2016).
40. Meric-Bernstam, F. *et al.* A decision support framework for genomically informed investigational cancer therapy. *J. Natl. Cancer Inst.* **107**, djv098 (2015).
41. Seashore-Ludlow, B. *et al.* Harnessing connectivity in a large-scale small-molecule sensitivity dataset. *Cancer Disc.* **5**, 1210–1223 (2015).
42. Trump, B.D., Linkov, F., Edwards, R.P. & Linkov, I. Not a humbug: the evolution of patient-centred medical decision-making. *Evid. Based Med.* **20**, 193–197 (2015).
43. Evans, W.E. & Relling, M.V. Pharmacogenomics: translating functional genomics into rational therapeutics. *Science* **286**, 487–491 (1999).