Applied & Translational Genomics 6 (2015) 18–25

Contents lists available at ScienceDirect

Applied & Translational Genomics

journal homepage: www.elsevier.com/locate/atg

OncDRS: An integrative clinical and genomic data platform for enabling translational research and precision medicine

John Orechia [1], Ameet Pathak [1], Yunling Shi, Aniket Nawani, Andrey Belozerov, Caitlin Fontes, Camille Lakhiani, Chetan Jawale, Chetansharan Patel, Daniel Quinn, Dmitry Botvinnik, Eddie Mei, Elizabeth Cotter, James Byleckie, Mollie Ullman-Cullere, Padam Chhetri, Poornima Chalasani, Purushotham Karnam, Ronald Beaudoin, Sandeep Sahu, Yelena Belozerova, Jomol P. Mathew *

Dana-Faber Cancer Institute, 450 Brookline Ave., Boston, MA-02215, United States

ARTICLE INFO

Article history:
Received 29 July 2015
Accepted 5 August 2015

Keywords:
Genomic profile
Clinical & genomic data integration
Next generation sequencing data
Precision medicine
Clinical and translational informatics

ABSTRACT

We live in the genomic era of medicine, where a patient's genomic/molecular data is becoming increasingly important for disease diagnosis, identification of targeted therapy, and risk assessment for adverse reactions. However, decoding the genomic test results and integrating it with clinical data for retrospective studies and cohort identification for prospective clinical trials is still a challenging task. In order to overcome these barriers, we developed an overarching enterprise informatics framework for translational research and personalized medicine called Synergistic Patient and Research Knowledge Systems (SPARKS) and a suite of tools called Oncology Data Retrieval Systems (OncDRS). OncDRS enables seamless data integration, secure and self-navigated query and extraction of clinical and genomic data from heterogeneous sources. Within a year of release, the system has facilitated more than 1500 research queries and has delivered data for more than 50 research studies.

1. Introduction

In the past two decades, release of the human genome sequence (Venter, et al. 2001) and development of cost effective genomic sequencing technologies (Wheeler et al. 2008; Bonetta 2010) have revolutionized our ability to understand genomic underpinning of diseases and tailor treatment strategies to target genomic alterations. The current trend is to move away from single-gene based tests. Instead, next generation sequencing (NGS) based genomic profiling is performed to detect multiple genetic alterations simultaneously, including, both common and rare sequence variants (MacConaill et al. 2009; Frampton et al. 2013; Dias-Santagata et al. 2010). Genomic profiling is transitioning from research-based tests to mainstream medical care where an individual patient's genetic profile is used to guide patient care/management (Yap and Popat 2014; Olsen and Jorgensen 2014; Larson and Wilke 2015). Detection of targetable alterations enables clinical researchers to design more efficient clinical trials where tailored therapies can be tested on selected patient cohorts that have specific clinical and genetic/molecular features. This will allow matching of patients with drugs that are better suited to their molecular profile, thereby reducing attrition rate of candidate drugs (Roper et al. 2015; Dienstmann et al. 2015). Additionally, retrospective analysis of exceptional responders (both positive and negative) helps in understanding why a particular treatment strategy worked, or did not work, for specific patients (Chau and Lorch 2015; Iyer et al. 2012; Printz 2015).

Informatics challenges for next generation clinical and translational research and precision medicine include ability to: a) aggregate, harmonize, integrate, and analyze the clinical and genomic/molecular data on a patient over a longitudinal continuum and b) access and visualize actionable findings in a timely manner for treatment decisions (Sulakhe et al. 2014); (Mate et al. 2011; Schriml and Mitraka 2015; Louie et al. 2007; Mathew et al. 2007). Clinical data are heterogeneous in general and are usually stored in multiple clinical and operational systems. The same clinical information (e.g. diagnosis) may be captured in different clinical, pathology, radiology and cancer registry systems. These systems may not follow the same standard terminologies or ontology, which may make deciphering the clinical phenotype a daunting task. For example, cancer registries normally use ICD-O codes for site, histology, and behavior, as well as other data, to describe a cancer diagnosis. On the other hand, the hospital-billing systems often use ICD-9 codes to record patient's diagnostic data. The fact that patient clinical information is captured in different systems can lead to problems with base patient population because all patients may not be present in all systems. The extent and granularity of information in each of the systems could

* Corresponding author currently at: University of Massachusetts Medical School, 55 Lake Avenue North, Worcester, MA 01655, United States.
E-mail address: jomol.mathew@umassmed.edu (J.P. Mathew).
[1] Equal Contribution.

http://dx.doi.org/10.1016/j.atg.2015.08.005

also be different resulting in incomplete data for some patients. This can lead to erroneous conclusions if the data analysis does not take into account such underlying differences between the systems.

In this paper, we introduce an informatics framework and a suite of tools that we developed at Dana-Farber Cancer Institute (DFCI) for effective integration of clinical and genomic data and providing relevant genomic results to clinical providers.

2. Methods

In 2011, DFCI and Brigham Woman's Hospital (BWH) launched Profile, one of the nation's most comprehensive personalized cancer initiatives (http://www.precisioncancermedicine.org/research-treatment/what-is-profile/). The project expanded to Boston Children's Hospital (BCH) in 2012. The Profile initiative is expected to create one of the world's largest databases of cancer-driven abnormalities and help advance personalized precision cancer care. In the last few years, more than 37,000 patients have consented to have tumor tissue analyzed for the presence of mutations and other cancer-related DNA abnormalities. More than 11,000 genetic profiles of patients' tumors were tested and more than 300 Profile tests are expected to be performed each month. The genetic test platforms used so far for Profile include mass spectrometric genotyping (OncoMap) (MacConaill et al. 2009) and targeted massively parallel sequencing (OncoPanel) (Wagle et al. 2012). The OncoMap test probed 471 mutations in 41 cancer-related genes. The OncoPanel test includes genomic data on sequence variants, copy number alterations, and selected chromosome rearrangements of 305 genomic regions for each patient who has consented to participate in the Profile study.

To support the Profile initiative, the Synergistic Patient and Research Knowledge System (SPARKS) was established as an institution-wide informatics framework, to accelerate scientific discoveries and their translation into personalized medicine and clinical practice. Three major systems were developed under the SPARKS framework:

1. OncoTracker: enables tracking of high-level status of Profile testing related processes including requisition, consent, specimen request, DNA extraction, and completion of genomic profiling and bioinformatics pipeline. OncoTracker allows providers and research teams to check on the status of profile test for their patients.
2. Profile Results Viewer: enables easy visualization of relevant genetic results by care providers based on informed consent status. Results are provided as a pathology report in 'PDF' format. The report highlights the clinical relevance of the variants detected in the patient's biospecimen, and may include recommendations for treatment and clinical trials when available and appropriate.
3. The Oncology Data Retrial System (OncDRS): implements policies, standards, systems, and tools, that addressed the challenges in integration of clinical and genomic data, and data governance, across disease areas. OncDRS, which is the focus of this paper, is a self-service application for investigators, and one-stop shop for data query, data request, data access approval, and data extraction.

2.1. Data governance and related policies

Development of OncDRS was guided by a set of processes and governance practices that are outlined in detail elsewhere and mentioned only briefly in here.

2.1.1. Logical separation of routine clinical data and consented research data collection & IRB approved master protocols governing data repositories and data access

The IRB policy stipulates that identified data, obtained during routine care of a patient, can be used for research if a waiver of consent and authorization is obtained from the IRB. Explicit patient consent is not required for research use of data obtained during routine patient care. However, research use of biospecimens and specimen-derived data and linkage to clinical data requires explicit consent from patients. In order to facilitate diversified use of both types of data, two separate repositories: Clinical Operational and Research Information System (CORIS) and Consented Research Data Repository (CRDR) were designed for warehousing clinical and biospecimen derived research data respectively. Two separate protocols were written to govern the collection, archival, and access of 1) clinical data in CORIS and 2) banking of biospecimens, storing specimen derived research data in CRDR and linkage to clinical data (Rollins and Kantoff, 2011, personal communication). The unified biospecimen and specimen derived data protocol simplified consenting and linking of genomic data to clinical data.

At the time of writing this manuscript, CORIS included clinical data captured in six different clinical and operational systems (Fig. 1). Clinical data including scheduling, registration, and billing information, are fetched from the clinical system (currently, GE-IDX). Laboratory test results were obtained from Sunquest Lab System. Medications dispensed in infusion clinics are sourced from Outpatient Pharmacy System. Cancer diagnosis information on all patients initially diagnosed or treated for cancer at the institute is recorded in the Cancer Registry System (Metriq). Details of chemotherapy drugs, and non-chemo drugs for patients seen at DFCI are obtained from DFCI Chemotherapy Order Entry System. Additionally, patient entered data from patient surveys and clinical data abstracted from medical charts by disease programs through the Clinical Research Information Systems (CRIS) is also stored in this repository. Patient enrollment on therapeutic, and ancillary therapeutic cancer-related clinical trials, for patients across the DF/HCC, are received from the DF/HCC Protocol Enrollment System.

Genomic data generated from Profile tests are fed into the Clinical Research Data Repository (CRDR). Genetic abnormalities detected with OncoPanel test include Single Nucleotide Variations (SNV), small indels, Copy Number Variations (CNV), and chromosomal translocations. Genomic coordinates from the human genome build hg19, reference allele and alternative allele of the SNVs are fed into the web service of OncoTator (http://www.broadinstitute.org/oncotator/) to obtain extensive annotation of the variants and transcripts. OncoMap used a three-tier schema while the newer OncoPanel uses a five-tier schema (MacConaill et al. 2014) to describe clinical actionability.

Table 1 presents a snapshot of all data in OncDRS based on the data refresh on June 2015.

2.1.2. Generalized and uniform informed consent

Profile was designed as a research test that necessitates patient's written consent for conducting the test and release of the data. A generalized and uniform consent across all cancer types was designed and approved as part of the banking and research data protocol with three questions for informed consent from the patient:

Q 1. Permission to analyze leftover clinically acquired specimens, link results to medical information, bank specimens, and derivatives for possible future research use.

Q 2. Permission to take an extra tube of blood, buccal swab and urine for genetic analyses and store materials for possible future research, and share the results of the analyses after removing personal identifying information.

Q 3. Permission to return relevant actionable results to the medical care team and to re-contact the patient about research studies that might be relevant for them in the future.

Currently, a profile test requisition is entered for every new patient who comes to the institute. However, sample retrieval and profile test will not start until a written consent is received from the patient. The written consent is entered into DF/HCC Clinical Trials Management System (CTMS) where all cancer protocol registrations are done.

Institutional policy requires that real-time checking of consent must be performed before a physician caring for the patient can view the Profile test results and before an investigator can obtain detailed patient level data for a research project from OncDRS (Fig. 2). The existence of

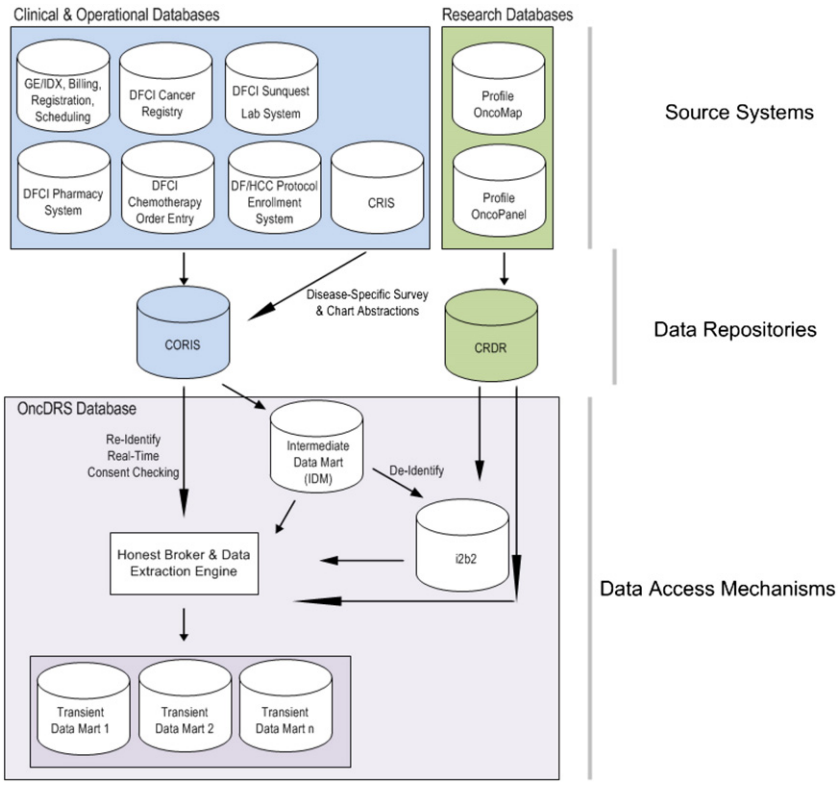


Fig. 1. OncDRS data sources and data integration pipeline. DFCI, Dana-Farber Cancer Institute; CRIS, Clinical Research Information System; CORIS, Clinical Operational and Research Information System; CRDR, Consented Research Data Repository.

the uniform informed consent process and real time consent checking before data release through OncDRS, not only maximizes patient participation in research, but also ensures compliance with patient directives on data use.

2.1.3. Data governance and access rules

Well-defined and clearly articulated data governance and access rules made the authentication and authorization process simple from a system design perspective. These rules governing who can access the system and what kind of permissions they have are outlined in Table 2. In addition, the levels of permissions and approvals needed from data governance groups and IRB to obtain aggregate, de-identified, limited and identified data sets for research are shown in Table 3.

Following these rules enables the OncDRS system to ensure that data is only provided to users based on the "need to know" principle. The principle of the "need to know" is enforced by two levels of review process: disease program specific user committee review and Institutional Review Board (IRB) review. A designated user committee is selected by each disease program and is responsible for the scientific review of data requests submitted by an investigator. A cross-disease committee oversees requests that span multiple diseases. The user committees also help in promoting collaborative research, by identifying similar requests, and encouraging requesters to collaborate. The user committees oversee data requests for de-identified, limited, and identified data. An additional IRB human subject protection review is required for requests for limited and identifiable data.

Designated disease user committee, and IRB, review significantly reduces time required for data approval, therefore, accelerating operational efficiency.

Table 1
A snapshot of all data in OncDRS, based on the data refresh on June 2015.

Data type	No. of records	System type
Demographics	276,039	Clinical operations
Medical billing diagnosis	8,270,751	Clinical operations
Outpatient clinic appointments	8,442,045	Clinical operations
Cancer registry	69,297	Clinical operations
Laboratory results	64,218,515	Clinical operations
Outpatient pharmacy dispenses	6,383,140	Clinical operations
Chemotherapy order entry	4,649,146	Clinical operations
Protocol enrollments	169,968	Clinical operations
Profile OncoMap results	5148	Research
Profile OncoPanel results	6378	Research

2.1.4. Repeatability of results

In order to ensure repeatability of results, policies were enforced to carry out systematic quality assurance (QA) and quality control (QC) of data. The QA procedure scans data following any data movement from the source systems to the data repository layer, and from data repository layer into the access layer (Fig. 1). Any discrepancies found in source data during the QA process is reported back to operational departments who will be responsible to fix the problem in the source systems. Currently, clinical data in CORIS is refreshed every month while genomic data in CRDR is refreshed twice a year. Rigorous data consistency checks also help unearth any mismatches/issues in terms of patient information, sample identifier, histology, tumor type, and genetic abnormalities.

2.2. OncDRS components

OncDRS is composed of a suite of component tools: administrative components, Aggregate Query Tool (AQT), which leverages the Informatics for Integrative Biology and Bedside (i2b2) open source application (Kohane et al., 2012), Data Request Engine (DRE), and Data Extraction Engine (DEE). At a high level, these components enable an authenticated and authorized faculty user to set up a project, manage team members on their project, perform queries to define the cohorts,

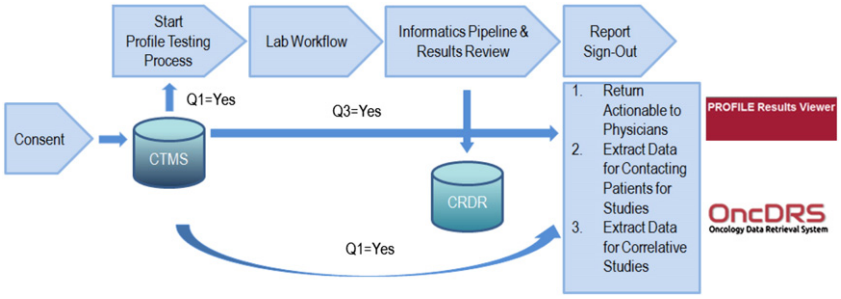


Fig. 2. Real-time consent checking in OncDRS. CRDR, Consented Research Data Repository; CTMS, Clinical Trials Management System.

request data on the defined cohort, and visualize and analyze the data using SAS and R. A screen shot of the SPARKS portal displaying OncDRS capabilities is presented in Fig. 3. Faculty members who are also members of the user committees can review data requests. User committee chair and IRB members can review and approve requests — following similar principles as a journal article review by an editorial board.

2.2.1. Administrative components

These tools are used to set up user accounts, project teams, and to assign appropriate roles to each user. The authentication components include LDAP authentication, to ensure that the user is an active institutional member. Harvard Profiles authentication is done to ensure active Harvard faculty status at the level of instructor or above. The project cell of i2B2 is leveraged to ensure project-based access, and ability to share query artifacts among collaborators.

2.2.2. Aggregate Query Tool (AQT)

The AQT includes hierarchy presentation and the query engine. The i2b2 hierarchy presentation for clinical data was adapted and designed hierarchy presentation for genomic data. AQT allows users to run queries under three settings: 1) *Same Patient level*: meaning the patient has the selected criteria at any point of time. For example, a patient was dispensed carboplatin and had a thrombocytopenia at some point during the care process. 2) *Same date level*: meaning that the patient has the selected criteria on the same date. For example, a patient was dispensed carboplatin and had nausea and vomiting on the same date. 3) *Same primary cancer episode*: meaning that the patient has the selected criteria on the same episode of cancer. For example, a patient was given carboplatin and had an adverse reaction when cancer site = breast, histology = inflammatory carcinoma, clinical group stage = II. All project team members can perform aggregate queries and can choose to share queries with other team members or keep them private.

Table 2
OncDRS user roles and privileges.

Permissions	Faculty member	Non-faculty member
To access the system data dictionary	Yes	Yes
To perform aggregate queries	Yes	Yes, if a faculty member grants access.
To request detailed patient level data	Yes	Yes, if a faculty member grants access. At present submission of the request has to be done by the faculty member.
To access detailed patient data through system	Yes	Not currently, but they will be able to if listed in the protocol as collaborator
To create project teams with multiple collaborators	Yes	No
To function as a disease user committee chair or member	Yes	No

2.2.3. Data Request Engine and Data Extraction Engine (DRE)

The data request and data extraction engine application components collectively include the data request form, approval routing, Transient Data Mart (TDM) generator, email sender, and data viewer. Users can define cohorts using a previously run query or a Medical Record Number (MRN) list, to specify the data categories of interest and specify the disease program from which they are requesting data. All project members can fill in the request form, and the lead faculty member can submit the request. An approval routing cycle is activated automatically once a data request is submitted and the chair of the user committee of the disease (or cross-disease) program from which the data is requested, gets notified through email. The chair can then decide to route the request to other members, request more information from the data requester, or approve or deny requests. Once the disease committee chair approves a request for limited or identifiable data, it gets routed to the IRB. Once necessary approvals (Table 3) are in place, the TDM generator extracts data on requested data categories on the specified cohort of patients identified through the AQT query, or the MRN list, attached to the data request. A TDM database is created with user's identification and password is encapsulated. TDMs are locked from editing, are for view only, and are connected to SAS and R to support easy analysis. TDMs are available to the faculty requester only for the number of days specified in the data request. In the meanwhile, the project identifier, released patient sets, and user's credentials are collected in the database for auditing purposes. A “Detailed Data Viewer” component provides additional search and filtering capabilities.

2.3. Technological architecture

OncDRS is an N-tier application. Fig. 4A displays architecture of the system and Fig. 4B describes the deployment pattern of OncDRS. The technological product stack used for OncDRS development is listed in Table 4. The database tier contains a series of databases (AQT, IDM, OncDRS, and Transient Data Mart (TDM) hosted mainly on Oracle database servers. MySQL database will be used for retired TDMs for cost efficiency. Extract, Transform, Load (ETL) tools from Informatica (2100 Seaport Blvd, Redwood City, CA 94063) are used to make transmission of data seamless, between the various OncDRS data storage layers. The middle tier consists of core applications (Administrative components, AQT DRE and DEE) on two different application servers (JBOSS 4.2.2 GA and Apache Tomcat 6.0.24). JBOSS is the default i2b2 server and is required for AQT. Apache Tomcat server hosts all other application components. The user interface (UI) tier spans across Client Browser and Apache Webserver 2.2 (httpd). All database servers and application servers are located within the institute's global firewall. Users can access OncDRS either within the institute facilities or via VPN service to ensure secure and private access of the system.

2.3.1. Data integration and honest broker

The OncDRS system transforms the “need to know” principle into an honest broker system, which prevents unintentional disclosures. The

Table 3
OncDRS governing body approval requirements for different data types.

Data type	Definition	Review needed	Module
Aggregate	Patient count is returned for a given query.	No review	Aggregate Query Tool
De-identified	All identifiable elements[a] are set to null or modified to prevent patient identification.	Disease-based user committees	Data request and data extraction
Limited	Most identifiable elements are set to null or modified to prevent patient identification.	Disease-based user committees & IRB	Data request and data extraction
Identified	Identifiable elements are displayed, as they exist in the source system.	Disease-based user committees & IRB	Data request and data extraction

[a] Identifiable data elements include items such as medical record numbers, names, phone numbers, addresses, age, and encounter dates.

two data warehouses (CORIS and CRDR) contain a patient's real identifiers, and data is always identified in these two repositories (Fig. 1.). Users do not have direct access to either of these repositories. The clinical data is pushed into an Intermediate Data Mart (IDM) where each patient is assigned an *internal research master patient identifier*. OncDRS only moves the research master patient identifier into AQT, and leaves the real patient identifiers in the IDM. Thus, unless a user specifies a need to review patient chart or to obtain "identifiers of patients who may be appropriate for a study, and requests patient information with identifier or contact information in the data request form, the user cannot identify the patient. Furthermore, patient counts returned by AQT are always obfuscated by ± 3, and no patient counts are provided if the count is less than three. Obfuscation and blocking of low patient counts were implemented as additional measures to prevent users from determining patient identity using aggregate queries.

2.3.2. Real-time consent check

Even though a Profile test will not be done until signed informed consent is received from patients, a patient can change his/her mind and withdraw the consent during or after the test is completed. One of the critical features of the OncDRS application is a real-time consent check mechanism to ensure that the answer is "yes" on the primary enrollment question before any data extraction and delivery (Fig. 2.). If the consent status is "No" at the time of data release, the patient's existing profile test data will be excluded from the dataset delivered to the TDM. These checks help ensure that a patient's consent directives are followed even when a patient withdraws consent after initial consent.

2.3.3. System security

OncDRS implements security in two stages: authentication and authorization (Fig. 4A). Authentication is the first stage of security check, where OncDRS ensures that the user is indeed who he/she claims to be. Authentication is implemented by using Partners Healthcare System's (PHS) LDAP and Harvard University Profiles System. The second stage is authorization that grants users permission to perform different functions in OncDRS based on the user's identity in accordance with rules aforementioned, and outlined in Table 2. The project management (PM) cell of i2B2 is used to enforce institutional policies regarding privileges.

3. Discussion

Dana-Farber Cancer Institute has developed and implemented an enterprise translational informatics system to integrate clinical and genomic data with HIPAA compliance. Since its launch in the summer of 2014, the system has successfully handled more than 1500 research queries and has released data for more than 50 research studies. So far the system has facilitated searches and data extractions of the following broad categories:

- Feasibility assessment: are there enough number of patients that match certain criteria for a potential clinical trial?
- Cohort identification and recruitment for clinical trials: to identify the potential list of participants for an open trial
- Correlative studies: extract all clinical and genomic data on the following for detailed analysis:

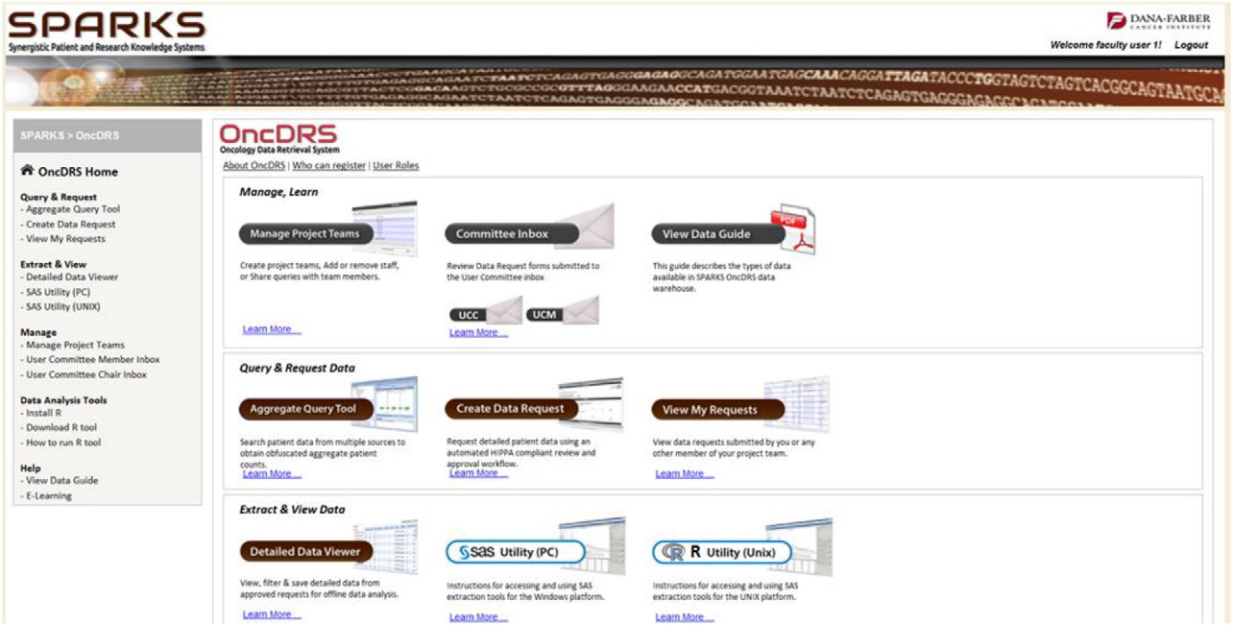
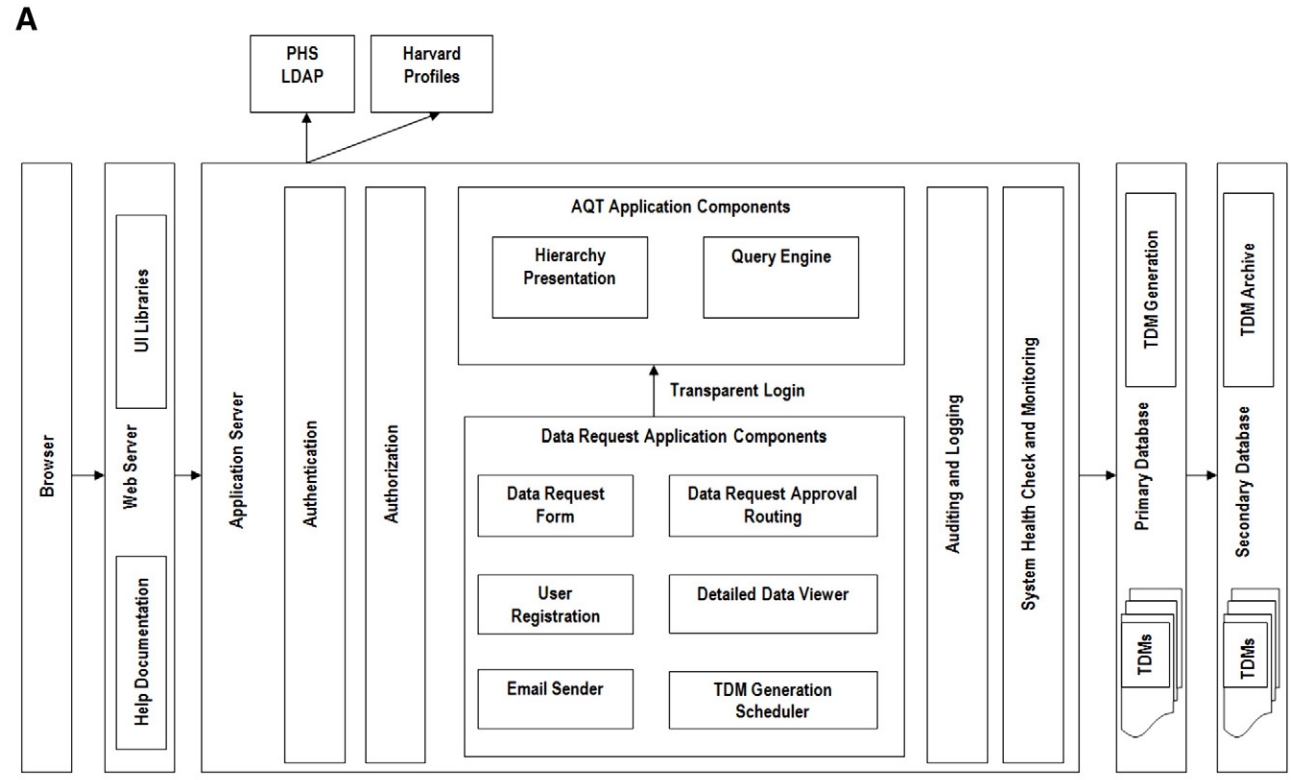


Fig. 3. SPARKS portal displaying OncDRS capabilities. SPARKS, Synergistic Patient and Research Knowledge System; OncDRS, Oncology Data Retrial System.



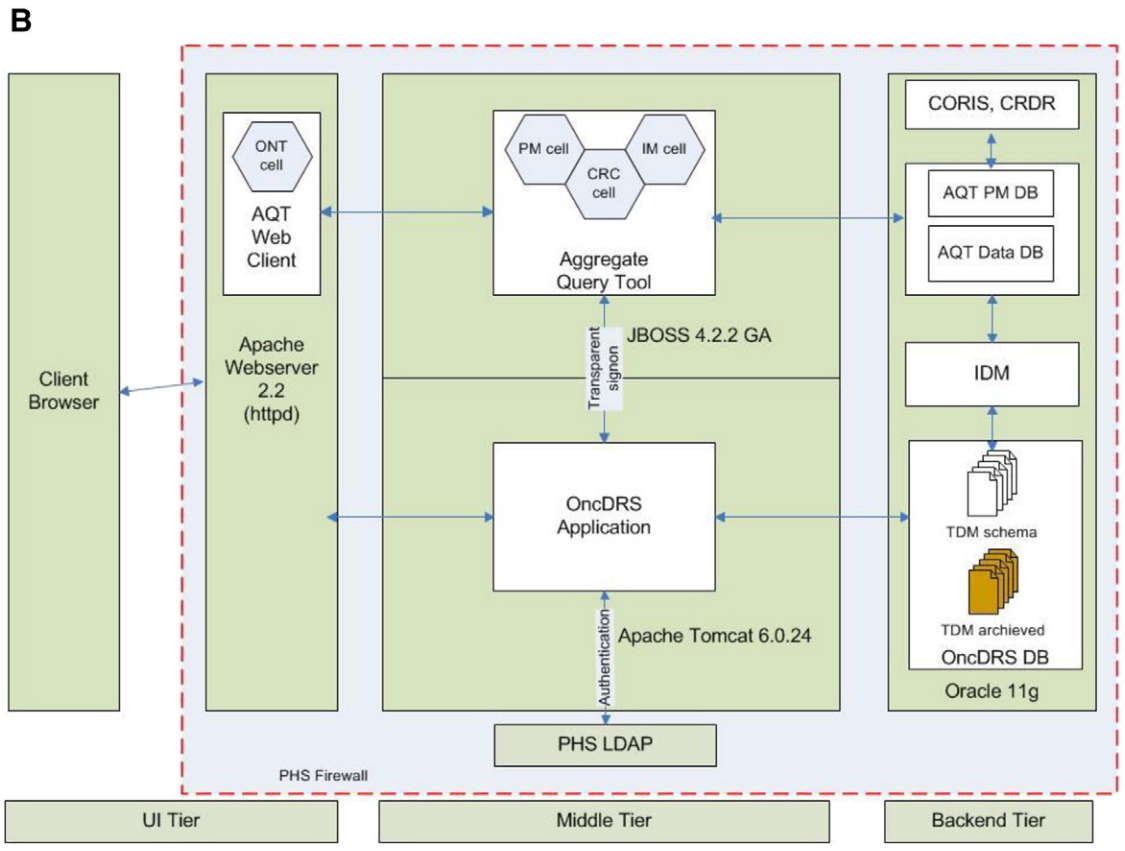


Fig. 4. A. OncDRS System details. PHS, Partners Healthcare System; LDAP, Lightweight Directory Access Protocol; AQT, Aggregate Query Tool; TDM, Transient Data Mart. B. OncDRS deployment diagram. PHS, Partners Healthcare System LDAP, Lightweight Directory Access Protocol; ONT, ontology; PM, project management; CRC, i2b2 Data Repository; IM, identity management; CORIS, Clinical Operational and Research Information System; CRDR, Consented Research Data Repository; AQT, Aggregate Query Tool; TDM, Transient Data Mart; IDM, Intermediate Data Mart.

Table 4
Technological product stack used in OncDRS development.

Product	OncDRS components	Vendor
Java/JEE 1.6	Server side components	Oracle Inc. — open source
Spring 3.0.4	Server side framework	Spring.io — open source
Hibernate 3.4.0.GA	ORM framework	Hibernate.org — open source
Servlet 2.5	Server side components	Oracle Inc. — open source
JSP 2.1	Web pages	Oracle Inc. — open source
Mail 1.4.1	Email API	Oracle Inc. — open source
displaytag 1.2	Web Presentation Library	http://www.displaytag.org/1.2/ — open source
SiteMesh 2.4.2	Web Pages Theme Library	http://wiki.sitemesh.org/ — open source
Jaxb 2.1	Java XML Processing Library	Oracle Inc. — open source
Log4j 1.2.26	Logger Library	Apache Foundation — open source
jQuery 1.4.2	Javascript UI Library	https://jquery.com/ — open source
Oracle database	Oracle 11g	Oracle Inc.
MySQL database	MySQL 5.0.95	Oracle Inc. — open source
Data Integration & ETL Tool	Power Center 9.1	Informatica

- Adult testicular cancer (ICD9: 186.9 and V10.52) patients diagnosed (between 2008 and 2012) and received Cisplatin treatment in the last 2 years.
- Prostate cancer patients (PSA levels 4–10) treated between 2005 and 2012.
- Primary ovarian cancer (with carcinoma and adenocarcinoma) patients with selected variants.

The success of OncDRS in such a short period of time can be attributed to self-service capability and ease of use. One half hour training is usually sufficient for a new user to learn how to develop an aggregate query, complete a data request form, and extract data for analysis. Also, because AQT returns counts of patients in a matter of seconds, users can easily get a good estimate of expected cohort size before investing time in initiating research projects.

The modularity of OncDRS, by design, makes it easy to adapt to the natural and anticipated evolution of software and scientific technology. For example, test results from the Profile OncoMap test, and the Profile OncoPanel test, which utilizes different technologies, have been successfully integrated.

Although OncDRS is very efficient in integrating and enabling access to heterogeneous data, several opportunities to improve the system have been identified. To ensure that the data requests are processed in a timely and efficient manner, user committee chairs and IRB directors need to be provided with the ability to delegate administrative governance tasks to staff members.

Additionally, OncDRS does not efficiently support on-going clinical trial recruitment, as the user currently needs to do a new data request every time they need to get a list of potential candidate patients. Screening and selecting patients with particular genetic alterations for a clinical trial are popular use-cases, and so, OncDRS needs to support automated periodic generation of new recruiting lists based on a single approved data request as long as the trial is open to accrual. Improvements in the frequency of clinical and profile data are also essential because current refresh cycles are not sufficient to serve all the needs for clinical trials recruiting.

AQT enables efficient queries by mapping each concept into a row in the observation_fact table and by encapsulating the concept and hierarchical structure in the metadata table. While this works perfectly with clinical data, it may not lend well for future genomic queries. Currently genetic alterations are mapped to a concept, and added into the observation_fact table. With only 305 genes and limited genetic alterations, the system works well. As we march into the era of the whole exome and whole genome sequencing, the data will grow exponentially. Modeling a vast number of genetic variations in a hierarchy may become unrealistic. A hybrid approach with hierarchy presentation on clinical data side and an integrated new module for genomic data could be a potential solution and needs to be explored.

References

Bonetta, L., 2010. Whole-genome sequencing breaks the cost barrier. Cell 141, 917–919.

Chau, N.G., Lorch, J.H., 2015. 'Exceptional responders inspire change: lessons for drug development from the bedside to the bench and back. Oncologist.

Dias-Santagata, D., Akhavanfard, S., David, S.S., Vernovsky, K., Kuhlmann, G., Boisvert, S.L., Stubbs, H., McDermott, U., Settleman, J., Kwak, E.L., Clark, J.W., Isakoff, S.J., Sequist, L.V., Engelman, J.A., Lynch, T.J., Haber, D.A., Louis, D.N., Ellisen, L.W., Borger, D.R., Iafrate, A.J., 2010. Rapid targeted mutational analysis of human tumours: a clinical platform to guide personalized cancer medicine. EMBO Mol. Med. 2, 146–158.

Dienstmann, R., Rodon, J., Tabernero, J., 2015. Optimal design of trials to demonstrate the utility of genomically-guided therapy: putting precision cancer medicine to the test. Mol. Oncol. 9, 940–950.

Frampton, G.M., Fichtenholtz, A., Otto, G.A., Wang, K., Downing, S.R., He, J., Schnall-Levin, M., White, J., Sanford, E.M., An, P., Sun, J., Juhn, F., Brennan, K., Iwanik, K., Maillet, A., Buell, J., White, E., Zhao, M., Balasubramanian, S., Terzic, S., Richards, T., Banning, V., Garcia, L., Mahoney, K., Zwirko, Z., Donahue, A., Beltran, H., Mosquera, J.M., Rubin, M.A., Dogan, S., Hedvat, C.V., Berger, M.F., Pusztai, L., Lechner, M., Boshoff, C., Jarosz, M., Vietz, C., Parker, A., Miller, V.A., Ross, J.S., Curran, J., Cronin, M.T., Stephens, P.J., Lipson, D., Yelensky, R., 2013. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. Nat. Biotechnol. 31, 1023–1031.

Iyer, G., Hanrahan, A.J., Milowsky, M.I., Al-Ahmadie, H., Scott, S.N., Janakiraman, M., Pirun, M., Sander, C., Socci, N.D., Ostrovnaya, I., Viale, A., Heguy, A., Peng, L., Chan, T.A., Bochner, B., Bajorin, D.F., Berger, M.F., Taylor, B.S., Solit, D.B., 2012. Genome sequencing identifies a basis for everolimus sensitivity. Science 338, 221.

Kohane, I.S., Churchill, S.E., Murphy, S.N., 2012. A translational engine at the national scale: informatics for integrating biology and the bedside. J. Am. Med. Inform. Assoc. 19, 181–185.

Larson, E.A., Wilke, R.A., 2015. Integration of genomics in primary Care. Am J Med.

Louie, B., Mork, P., Martin-Sanchez, F., Halevy, A., Tarczy-Hornoch, P., 2007. Data integration and genomic medicine. J. Biomed. Inform. 40, 5–16.

MacConaill, L.E., Campbell, C.D., Kehoe, S.M., Bass, A.J., Hatton, C., Niu, L., Davis, M., Yao, K., Hanna, M., Mondal, C., Luongo, L., Emery, C.M., Baker, A.C., Philips, J., Goff, D.J., Fiorentino, M., Rubin, M.A., Polyak, K., Chan, J., Wang, Y., Fletcher, J.A., Santagata, S., Corso, G., Roviello, F., Shivdasani, R., Kieran, M.W., Ligon, K.L., Stiles, C.D., Hahn, W.C., Meyerson, M.L., Garraway, L.A., 2009. Profiling critical cancer gene mutations in clinical tumor samples. PLoS One 4, e7887.

MacConaill, L.E., Garcia, E., Shivdasani, P., Ducar, M., Adusumilli, R., Breneiser, M., Byrne, M., Chung, L., Conneely, J., Crosby, L., Garraway, L.A., Gong, X., Hahn, W.C., Hatton, C., Kantoff, P.W., Kluk, M., Kuo, F., Jia, Y., Joshi, R., Longtine, J., Manning, A., Palescandolo, E., Sharaf, N., Sholl, L., van Hummelen, P., Wade, J., Wollinson, B.M., Zepf, D., Rollins, B.J., Lindeman, N.I., 2014. Prospective enterprise-level molecular genotyping of a cohort of cancer patients. J. Mol. Diagn. 16, 660–672.

Mate, S., Burkle, T., Kopcke, F., Breil, B., Wullich, B., Dugas, M., Prokosch, H.U., Ganslandt, T., 2011. Populating the i2b2 database with heterogeneous EMR data: a semantic network approach. Stud. Health Technol. Inform. 169, 502–506.

Mathew, J.P., Taylor, B.S., Bader, G.D., Pyarajan, S., Antoniotti, M., Chinnaiyan, A.M., Sander, C., Burakoff, S.J., Mishra, B., 2007. From bytes to bedside: data integration and computational biology for translational cancer research. PLoS Comput. Biol. 3, e12.

Olsen, D., Jorgensen, J.T., 2014. Companion diagnostics for targeted cancer drugs — clinical and regulatory aspects. Front Oncol. 4, 105.

Printz, C., 2015. NCI launches exceptional responders initiative: researchers will attempt to identify why some patients respond to treatment so much better than others. Cancer 121, 803–804.

Roper, N., Stensland, K.D., Hendricks, R., Galsky, M.D., 2015. The landscape of precision cancer medicine clinical trials in the United States. Cancer Treat. Rev. 41, 385–390.

Schriml, L.M., Mitraka, E., 2015. The Disease Ontology: fostering interoperability between biological and clinical human disease-related data. Mamm. Genome.

Sulakhe, D., Balasubramanian, S., Xie, B., Berrocal, E., Feng, B., Taylor, A., Chitturi, B., Dave, U., Agam, G., Xu, J., Bornigen, D., Dubchak, I., Gilliam, T.C., Maltsev, N., 2014. High-throughput translational medicine: challenges and solutions. Adv. Exp. Med. Biol. 799, 39–67.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H., Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J., Gabor Miklos, G.L., Nelson, C., Broder, S., Clark, A.G., Nadeau, J., McKusick, V.A., Zinder, N., Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A.E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T.J., Higgins, M.E., Ji, R.R., Ke, Z., Ketchum, K.A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G.V., Milshina, N., Moore, H.M., Naik, A.K., Narayan, V.A., Neelam, B., Nusskern, D., Rusch, D.B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L.,

Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M.L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N.N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J.F., Guigo, R., Campbell, M.J., Sjolander, K.V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., Zhu, X., 2001. The sequence of the human genome. Science 291, 1304–1351.

Wagle, N., Berger, M.F., Davis, M.J., Blumenstiel, B., Defelice, M., Pochanard, P., Ducar, M., Van Hummelen, P., Macconaill, L.E., Hahn, W.C., Meyerson, M., Gabriel, S.B., Garraway, L.A., 2012. High-throughput detection of actionable genomic alterations in clinical tumor samples by targeted, massively parallel sequencing. Cancer Discov. 2, 82–93.

Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.J., Makhijani, V., Roth, G.T., Gomes, X., Tartaro, K., Niazi, F., Turcotte, C.L., Irzyk, G.P., Lupski, J.R., Chinault, C., Song, X.Z., Liu, Y., Yuan, Y., Nazareth, L., Qin, X., Muzny, D.M., Margulies, M., Weinstock, G.M., Gibbs, R.A., Rothberg, J.M., 2008. The complete genome of an individual by massively parallel DNA sequencing. Nature 452, 872–876.

Yap, T.A., Popat, S., 2014. Toward precision medicine with next-generation EGFR inhibitors in non-small-cell lung cancer. Pharmg. Pers. Med. 7, 285–295.