Title

PRoBE the cloud toolkit: finding the best biomarkers of drug response within a breast cancer clinical trial.

Journal

JAMIA open, 4(2)

ISSN

2574-2531

Authors

O'Grady, Nicholas
Gibbs, David L
Abdilleh, Kawther
et al.

Publication Date

2021-04-01

DOI

10.1093/jamiaopen/ooab038

Peer reviewed

JAMIA Open, 4(2), 2021, 1–9
doi: 10.1093/jamiaopen/ooab038
Research and Applications

Research and Applications

PRoBE the cloud toolkit: finding the best biomarkers of drug response within a breast cancer clinical trial

Nicholas O'Grady [1], David L. Gibbs,[2,4] Kawther Abdilleh,[3,4] Adam Asare,[1] Smita Asare,[5] Sara Venters,[1] Lamorna Brown-Swigart,[1] Gillian L. Hirst [1], Denise Wolf,[1] Christina Yau,[1] Laura J. van 't Veer,[1] Laura Esserman,[1] and Amrita Basu[1]

[1]Department of Surgery, University of California San Francisco, San Francisco, California, USA, [2]Shmulevich Lab, Institute for Systems Biology, Seattle, Washington, USA, [3]General Dynamics, Department of Information Technology (GDIT), Rockville, Maryland, USA, [4]ISB-CGC, Seattle, Washington, USA and [5]Quantum Leap Healthcare Collaborative, San Francisco, California, USA

Corresponding Author: Amrita Basu, PhD, 550 16th St, 6th floor, San Francisco, CA 94158, USA; amrita.basu@ucsf.edu

Received 2 November 2020; Revised 5 January 2021; Editorial Decision 12 March 2021; Accepted 3 May 2021

ABSTRACT

Objectives: In this paper, we discuss leveraging cloud-based platforms to collect, visualize, analyze, and share data in the context of a clinical trial. Our cloud-based infrastructure, Patient Repository of Biomolecular Entities (PRoBE), has given us the opportunity for uniform data structure, more efficient analysis of valuable data, and increased collaboration between researchers.

Materials and Methods: We utilize a multi-cloud platform to manage and analyze data generated from the clinical Investigation of Serial Studies to Predict Your Therapeutic Response with Imaging And moLecular Analysis 2 (I-SPY 2 TRIAL). A collaboration with the Institute for Systems Biology Cancer Gateway in the Cloud has additionally given us access to public genomic databases. Applications to I-SPY 2 data have been built using R Shiny, while leveraging Google's BigQuery tables and SQL commands for data mining.

Results: We highlight the implementation of PRoBE in several unique case studies including prediction of biomarkers associated with clinical response, access to the Pan-Cancer Atlas, and integrating pathology images within the cloud. Our data integration pipelines, documentation, and all codebase will be placed in a Github repository.

Discussion and conclusion: We are hoping to develop risk stratification diagnostics by integrating additional molecular, magnetic resonance imaging, and pathology markers into PRoBE to better predict drug response. A robust cloud infrastructure and tool set can help integrate these large datasets to make valuable predictions of response to multiple agents. For that reason, we are continuously improving PRoBE to advance the way data is stored, accessed, and analyzed in the I-SPY 2 clinical trial.

Key words: clinical trials, cloud, breast cancer, data analysis, I-SPY 2

OBJECTIVES

Clinical trial centers need to address the untenability of data storage and analysis on local machines, as medical data grows in both size and complexity. To address these hurdles, a new paradigm is emerging that moves the conduct and management of clinical trials to cloud-based applications. With cloud-based technology, sponsors such as academic centers can build end-to-end data applications and transform their clinical development management in areas of data storage, aggregation, analysis, and sharing. Importantly, these cloud-based data platforms stowing genomic and clinical data must be interoperable so that information can be migrated and utilized

LAY SUMMARY

Multi-omics data in the world of medicine is ever increasing in size and complexity. Clinical trials are often challenged to develop methods for not only storing but also analyzing the molecular and clinical data generated. Due to the massive size requirements, local machines struggle, and data validation is often slow and inefficient. In this paper, we discuss leveraging cloud-based platforms to organize, visualize, analyze, and share data in the context of a clinical trial. Our cloud-based infrastructure, PRoBE, enables efficient analysis of data, such as predicting biomarkers of response, provides easy access to The Cancer Genome Atlas (TCGA), and allows for increased collaboration between researchers. Additionally, through a software integration our platform accommodates curation of pathology images so they can be directly accessed and viewed from the cloud. Through PRoBE, our stakeholders and trial managers now have a closer relationship to valuable clinical and biological data, under the umbrella of a consolidated and secure framework. We are heavily invested in further development of this platform as we continue to refine our ability to optimally identify biomarkers, discover therapeutic agents, and help every patient achieve an excellent long-term outcome.

between these complex systems. Leveraging cloud-based systems provides these benefits, all under a consolidated and secure framework.

Background and significance

Several recent initiatives have been focused on data sharing and collaboration between researchers through cloud platforms. A major part of this shift occurred in 2013 when the National Cancer Institute (NCI) initiated the Cloud Resources pilot program, which sought to democratize access to data generated through publicly funded research.[1] The initiative has made massive data sets publicly available as part of a public service known as the Cancer Research Data Commons.[2] These cloud resources aimed to make cancer genomics data sets, radiology and pathology images, together with tools and compute-power, available and accessible to a broad range of users using multiple access modes.

In the neoadjuvant breast cancer clinical trial, Investigation of Serial Studies to Predict Your Therapeutic Response with Imaging And moLecular Analysis 2 (I-SPY 2), we are leveraging the Institute for Systems Biology Cancer Gateway in the Cloud (ISB-CGC) platform to build a cloud infrastructure to be used within the context of a trial with multiple data collection timepoints and datatypes. ISB-CGC has made connecting to public databases, such as TCGA, straightforward through the use of Google tools and SQL commands. We aimed to use Google Cloud Platform (GCP) and the ISB-CGC infrastructure to store and analyze clinical, molecular, and pathology fingerprints from patient biopsies.

Here, we outline our work to improve the integration of research into clinical care within a multi-site breast cancer clinical trial. We have constructed a multi-cloud system, Patient Repository of Biomolecular Entities (PRoBE), which promotes the use of various cloud tools and data services to build custom applications for the I-SPY 2 trial. We highlight the implementation of PRoBE in several unique case studies including access to the Pan-Cancer Atlas, prediction of biomarkers associated with clinical response, and integrating pathology images within the cloud. Technical developments in our platform have adhered to the underlying guidelines of FAIR data access, and provide resources used to build an end-to-end cloud based clinical trial system. Our data integration pipelines, documentation, and all codebase will be placed in a Github repository. We address our stakeholders' requirements for a mature biorepository and outline current and future stages of this work. Our underlying goal through this technical integration effort is to identify biomarkers of clinical response to multiple drug regimens at differing timepoints.

The I-SPY 2 trial

The I-SPY 2 platform trial is an adaptively randomized, multicenter, multi-arm phase 2 study of investigational agents in combination with or in place of standard-of-care chemotherapy for breast cancer patients at high risk of recurrence.[3] One of the more innovative features of the trial is the ability to efficiently evaluate multiple experimental agents (or combinations of agents) simultaneously. I-SPY 2 employs a Bayesian adaptive randomization algorithm that preferentially assigns patients to agents that have accumulated evidence of efficacy in the same intrinsic subtype (Figure 1). Figure 1 displays the I-SPY 2 neoadjuvant trial schema that uses an adaptive randomization engine for targeted treatments specific to biomarker subtypes. Timepoints are denoted by "TX," in which various tests are performed, for example, magnetic resonance imaging (MRI), biopsies, and blood draws.

Serial imaging and biopsies performed over the treatment period inform adaptive randomization and are used as part of a wide-ranging biomarker discovery and validation program. The trial uses a neoadjuvant treatment model, which permits rapid assessment of tumor response to treatment (~6 months from beginning of treatment). As such, the primary endpoint of I-SPY 2 is pathological complete response (pCR), defined as the complete disappearance of invasive tumor, both in the breast and axilla.[4] A recently submitted manuscript demonstrates that pCR is a robust predictor of 3-year event-free and distant relapse-free survival in this population. We also use a residual cancer burden (RCB) index as a co-primary endpoint. The RCB index is a continuous variable (as opposed to pCR, which is binary) to quantify the extent of residual disease for patients who did not achieve pathologic complete response.[5] RCB index combines several factors, such as area of cancer, percent area that is invasive, number of positive nodes and largest node in area to name a few.[5] Both are useful prognostic factors for long-term survival in I-SPY 2 patients.

Continuously enrolling since 2010, I-SPY 2 is a mature trial ($n > 4000$), proven in its efficiency and meeting its primary objective of accelerating agent development and targeting.[6] To date, the trial has completed 16 experimental arms across 23 sites in the United States. Through the course of treatment, several diagnostic entities are collected such as tumor biopsies, MRIs, and a variety of molecular data at pre-treatment, inter-regimen, and post-treatment timepoints. The molecular data including gene expression, DNA and RNA sequencing, and protein modifications, are analyzed by a variety of biomedical vendors. Thus, we have an extensive database on I-SPY 2 patients and multiple requirements to be met across trial sites regarding data collection, sharing, privacy, and reporting.

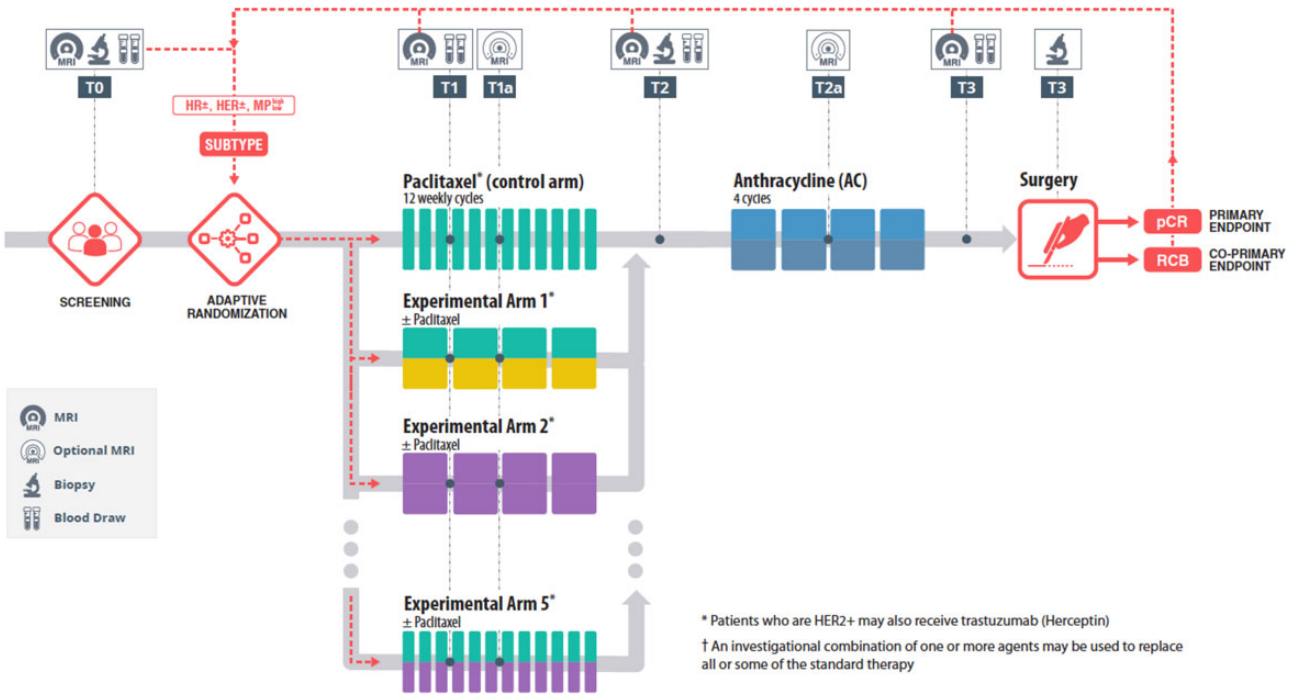

Figure 1. The I-SPY 2 neoadjuvant trial schema that uses an adaptive randomization engine for targeted treatments specific to biomarker subtypes. Timepoints are denoted by "TX," in which various tests are performed, for example, MRIs, biopsies, and blood draws.

MATERIALS AND METHODS

Cloud-computing environment

Our electronic data capture software is hosted on Amazon Web Services (AWS), and so we use AWS to store clinical and raw molecular data (eg, whole-genome sequencing, exome arrays, etc.) (Figure 2). Figure 2 shows a map of data flow throughout a cloud-based platform. A cloud environment allows for fluidity of data through multiple analytic platforms. Data originates from the patient, passes through various cloud platforms for analysis and querying, and ends with the investigator. For analysis, we have the ability transfer clinical and biological data from AWS to GCP to explore and visualize data in real time.

Data preparation

GCP provides several options in various programming languages for exploring data in a specified project. The primary, and simplest, method is to use the built-in analytics data warehouse "BigQuery." SQL queries of BigQuery tables can be executed both interactively through the Google Cloud web interface and programmatically through Python or R, enabling researchers to connect with data analysis and data visualization algorithms. As the service is entirely managed by Google, users can take advantage of robust computing power without having to manage their own cluster or configure database software. The massively parallel backend query engine enables SQL queries to be processed at incredibly fast speeds allowing researchers to mine through terabytes and even petabytes of cancer data in a relatively short amount of time. Google BigQuery is self-scaling; it identifies resource requirements for each query to finish quickly and efficiently and provides those resources to meet the demand. Once the workload has completed, BigQuery reallocates those resources to other projects and other users. Multi-user analysis is not an issue, and all handled by Google's job prioritization. A BigQuery slot is a virtual CPU (vCPU) used by BigQuery to execute SQL queries. BigQuery automatically calculates how many slots are required by each query, depending on query size and complexity. At any point, 100 simultaneous queries can be performed in a project.[7]

To effectively use BigQuery, a number of limited pre-processing steps need to be performed before uploading the data. BigQuery utilizes tabular data sets and data tables, which can be organized in many ways. We transformed I-SPY 2 data into the "tidy" format, which additionally works well within R's "Tidyverse" and all its associated tools for data science. The tidy data format requires: (1) each variable have its own column, (2) each observation must have its own row, and (3) each value must have its own cell.[8,9] This allows specific and unique results to be data mined through BigQuery in a consistent format. Once transformed, the data was uploaded and catalogued into various datasets, for example, clinical, biological, and pathology imaging data. In addition to I-SPY 2 data, ISB-CGC makes a large number of datasets publicly available in BigQuery, which allows for easy access and comparisons to private and public data. For example, ISB-CGC has made Pan-Cancer TCGA data publicly available using BigQuery. A web-based application that allows users to explore data available in ISB-CGC BigQuery tables is provided in the "Resource Availability" section at the end of the paper. Other databases, such as dbSNP[10] and ClinVar,[11] are also accessible through BigQuery, enabling I-SPY 2 researchers to query annotations for a particular SNP and compare gene or protein expressions. With I-SPY 2 data uploaded into BigQuery tables, we were able to discover various methods of data exploration and visualization, including an interactive and custom data query system using RShiny in conjunction with BigQuery, which leverages cloud computing power in the background. These applications are discussed in detail in the "Results" section.

Cloud script conversion

Pre-cloud implementation, I-SPY 2 pre-processing data functions were performed with local scripts using the R language.[12] Processing was slow, taxing on computer storage, and made it difficult for mul-

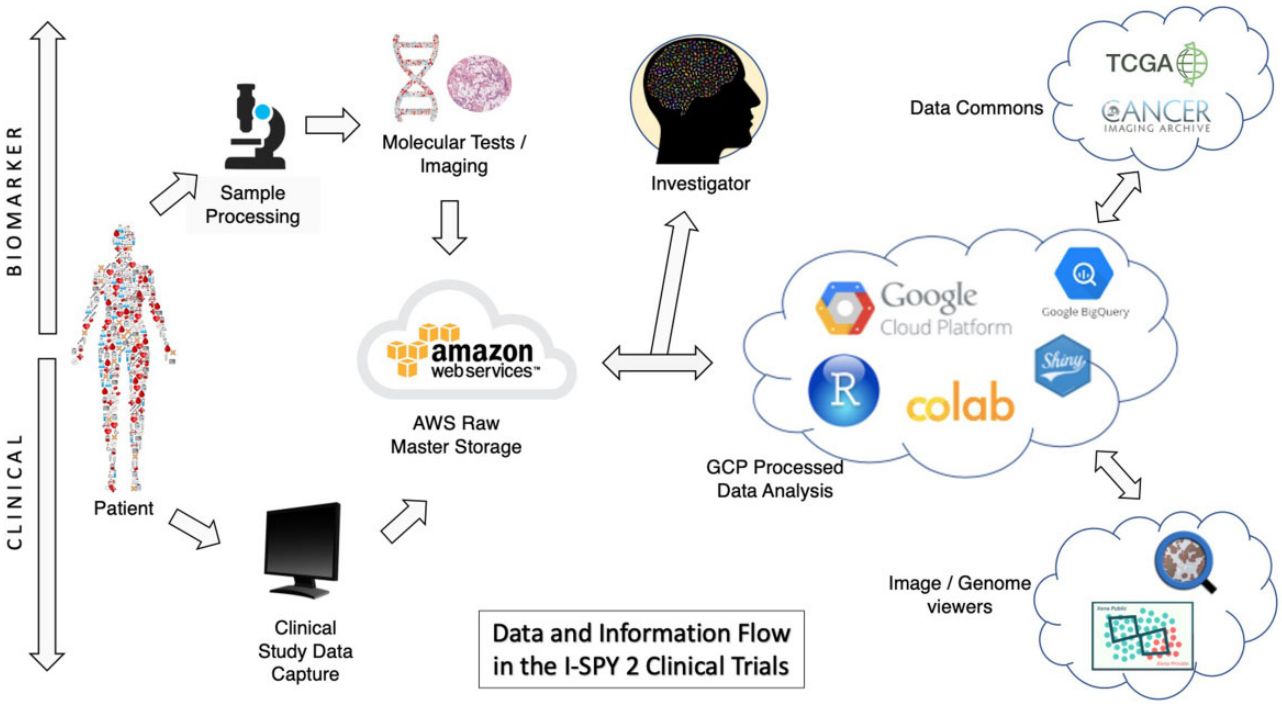

Figure 2. A map of data flow throughout a cloud-based platform. A cloud environment allows for fluidity of data through multiple analytic platforms. Data originates from the patient, passes through various cloud platforms for analysis and querying, and ends with the investigator. Cloud analytic platforms have connection capabilities to public genomic databases like the TCGA.

tiple analysts to work on the same datasets. We have leveraged the use of virtual machines in GCP to run RStudio directly in the cloud. This conversion gives us a high-speed connection to data stored in the cloud, as opposed to needing it on local machines. Working "within region," meaning the data and VMs are co-located, makes data transfers free and extremely fast. Users can scale between several virtual machine options, including memory size, vCPU count, and persistent disk limits. These VMs can be prioritized between general purpose, memory optimized, and compute optimized machines, well beyond our requirements (up to 224 vCPUs and 224 GB memory as of this writing).[13] Several R packages are available, predominantly those built by developer Mark Edmondson, such as googleAuthR, googleComputeEngineR, and googleCloudStorageR, that aid in transferring local scripts to a google RStudio server[14] (Github posted in "Resource Availability" section). Running R scripts in the cloud not only removed the need for local storage space, but drastically sped up our pre-processing functions.

For collaborative programming and analytical work, GCP also has a free python-based Jupyter style notebook called Colaboratory notebook. ISB-CGC has created a community notebook repository that contains a rich compendium of codes for specific bioinformatics tasks (GitHub posted in "Resource Availability" section).

High-throughput image data analysis

A major advancement made possible through cloud integration is our ability to share and view pathology images. Google's Health Care API,[15] which supports standard based data formats and protocols of existing healthcare technologies, allows for real time integration with various software programs. We utilize the Google Cloud Healthcare Datasets and Datastores, which accepts Digital Imaging and Communications in Medicine (DICOM) images, in sync with QuPath[16] (Github posted in "Resource Availability" section). QuPath is a software platform for whole slide image analysis, and a tool we have often used to view I-SPY 2 pathology images. Viewing pathology images from the cloud is discussed in detail in the results section.

I-SPY 2 controlled data access and sharing

An I-SPY 2 Data Access and Publications Committee (DAPC) accepts proposals from internal and external investigators interested in accessing clinical trials data, including radiology and pathology imaging for original research. If an investigators concept is approved, access to I-SPY 2 curated data sets can be granted, including any PRoBE applications that have been created, after a data use agreement has been signed. This allows researchers fast access to multi omics data acquired through I-SPY 2, as well as any data sets ISB-CGC has transformed and made available in BigQuery format, as explained in the results section. A recommended best practice is to additionally execute a Google Cloud Business Association Agreement, that covers the scope of the project. GCP is HIPAA compliant[17] and uses role-based privileges to grant access to individual buckets. Additionally, the DAPC ensures no I-SPY 2 data released to investigators can be traced back to an individual patient, in keeping with Section 164.514(a) of the HIPAA Privacy Rule.[18] As an extra security measure, we have instituted a double-blind patient identifier. This double-blind ID is applied and known to a select group at the I-SPY 2 trial's central location, unknown to hospital sites who have applied the original patient identifier. All datasets are double blinded before being released to interested investigators or for publications.

Before a dataset gets uploaded to PRoBE, it must pass through internal quality controls from a panel of biostatisticians and scientists, field dependent, as well as members of the lab that originally generated the data. Both clinical and biological data must be approved and locked, before being made accessible to interested researchers. In this way, we are better able to control versioning of data, as approved

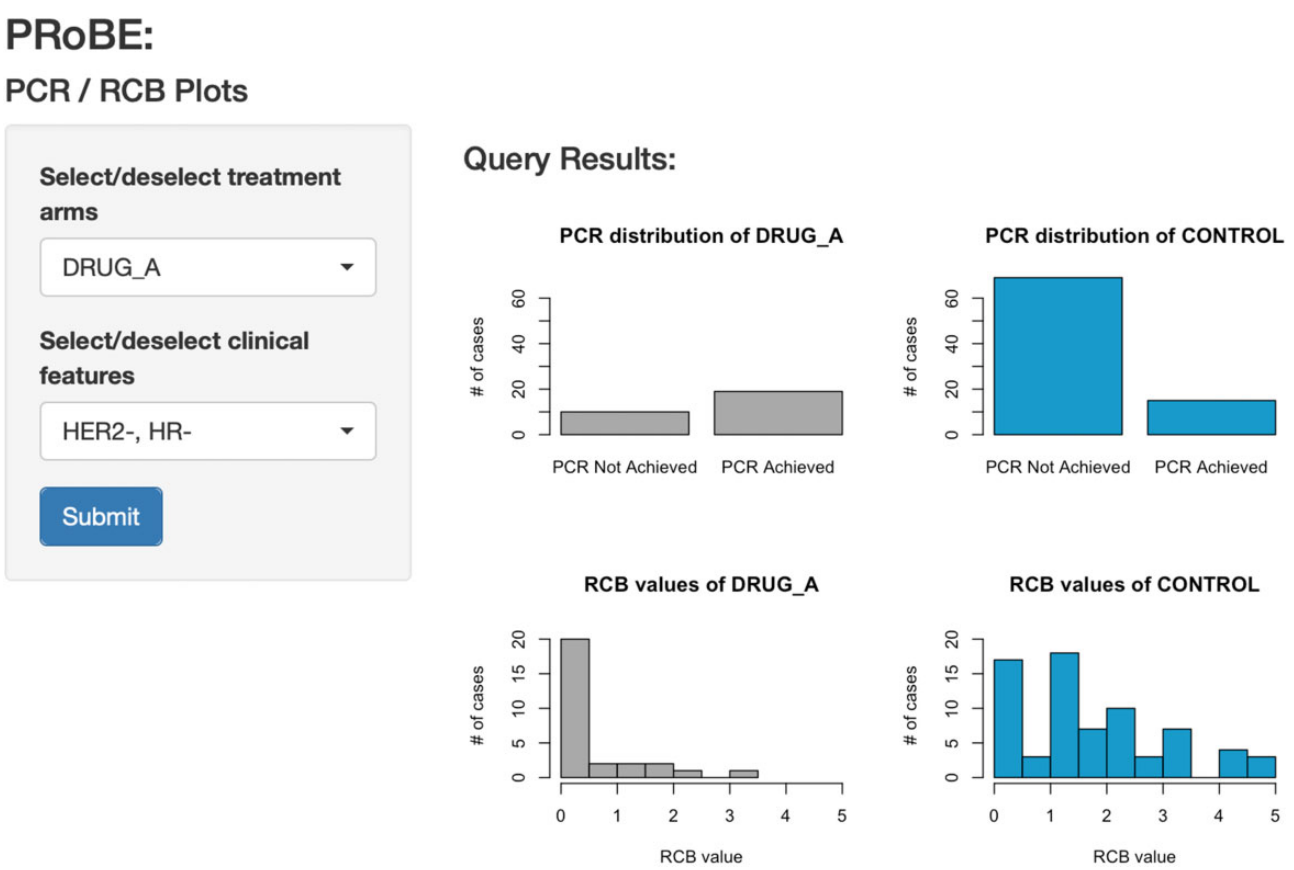

Figure 3. R Shiny application that accesses BigQuery tables in the background. Application displays pCR and RCB plots for treatment and control arms, with filters on drug and clinical subtype.

users are pulling from a controllable, single source of truth. Although it is possible for other sites to contribute data to this platform, PRoBE currently only uploads data directly related to the I-SPY 2 trials.

While leveraging the computing power and massive storage capabilities of cloud environments provides an enormous opportunity to bioinformaticians, it is important to remember these tools may not be as easily accessible to those who are less technically adept. User-friendly tools must be developed so data can be accessed and analyzed by additional members of any research team. Due to the number of clinical variables that may need to be queried in parallel, we have developed a few web-based applications for routine use by I-SPY 2 clinical trials team. Additionally, these applications will enable investigators both within and outside of I-SPY 2 to corroborate and compare efficacy and molecular signatures under both public and controlled access. We have leveraged Google's access control and account permissions to assign view, edit, and grant privileges based on a person's role in the trial. Should a researcher be approved for full access to data, this would not prevent a user's capability to download data and work on local machines.

Below, we discuss these applications to help I-SPY 2 case coordinators, investigators, and the research community at large to access all public datasets that we are hosting.

RESULTS

Case study: evaluating clinical efficacy data with PRoBE

To examine the efficacy of a given drug versus the control, PRoBE offers an application that displays the distribution of select clinical variables collected in the I-SPY 2 trial. In this example, we focus on 2 co-primary endpoints, pCR and RCB index. We leverage PRoBE's visualization pipeline and R Shiny to make interpretable plots for both clinicians and researchers (Figure 3) (Github posted in "Resource Availability" section). Figure 3 shows an R Shiny application that accesses BigQuery tables in the background and displays pCR and RCB plots for treatment and control arms. Data is mined through use of SQL commands, for any tables uploaded in BigQuery. We show a distribution of those who achieved and did not achieve pCR in the experimental and control arm. We also plot the distribution of RCB indices, which tangentially relates to pCR (ie, pCR = 1 equivalent to RCB = 0). In this example, patients in the "Drug A" arm have a higher pCR rate, and lower RCB index (RCB = 0 = no cancer; RCB = 5 = extensive residual cancer) than patients in the control arm. Other clinical variables such as subtype (eg, HER2, receptor status) can also be used to filter through the application.

Case study: accessing public genomic datasets

The Pan-Cancer Atlas was the penultimate project following The Cancer Genome Atlas, where 33 types of cancer were investigated, producing over a petabyte of publicly accessible data.[19] The Pan-Cancer Atlas project contains batch-corrected RNA-seq data for 20 502 genes across 9921 patients. The ISB-CGC BigQuery resource was created to mirror the data, but in a format that facilitated processing, analyzing, and integrating "big data," thus shifting the large-scale compute power needed away from local systems and into Google cloud.[20]

Figure 4. Distribution of Gene Expression Values from data collected in the I-SPY 2 Trial and TCGA Cohorts.

We have leveraged the instantiation of the Pan-Cancer Atlas resource, available now in BigQuery, so that our investigators can compare and validate findings within a large cohort of multiple cancers. I-SPY 2 is particularly interested in whether single genes are over or under expressed, especially in complicated and difficult diseases such as basal-like breast cancer that are hard to treat. Given immune signatures are emerging as useful predictors of therapeutic response in breast cancer,[21] we display expression of PD-L1 (CD274), a well-known immune modulator across multiple subtypes including basal-like, HER2, and luminal A (Figure 4) (Github posted in "Resource Availability" section). Figure 4 shows distribution of Gene Expression Values from data collected in the I-SPY 2 Trial and TCGA Cohorts.

Data of both public and private tables are easily mined through the use of SQL commands. The boxplot on the top displays I-SPY 2 data and bottom displays the Pan-Cancer Atlas data (BRCA cohort). We also included the number of observations seen in each subtype (n), as well as a t-statistic between subtypes (in the same dataset). This interface quickly allows for a rapid visualization of expression levels in any gene of interest in I-SPY 2 patients, and larger public databases like Pan-Cancer Atlas.

Case study: discovering biomarkers of resistance across different arms using machine learning

For every consented I-SPY 2 patient, tumor biopsies are collected, analyzed via microarray, and sequenced (among many other tests, clinical arm dependent). PRoBE offers a pipeline, using RWEN, a weighted elastic net method[22] to predict response of treatment across any experimental arm of choice. The resulting output is a list of candidate gene expression markers, predictive of sensitivity or resistance. Through systematic benchmarking and literature surveys, we show that our method has an overall lower median root-mean-squared error (RMSE) of response compared to traditional statistical methods that do not predict well on sample outliers (median RMSE=0.18).[22] PRoBE allows this analysis to be performed for a single subtype through one simple operation, and for each arm separately.

PRoBE also offers analysis of various external datasets such as the Cancer Therapeutics Response Portal[23,24] and Sanger Genomics of Drug sensitivity[25] datasets. In each of these datasets, cell lines were treated with a single agent and cellular response was measured. The cell lines are characterized for gene expression as well as other genetic variants. For each arm, we compared the genes in our list to resistance genes that are present in CTD2 and Sanger datasets (Figure 5) (Github posted in "Resource Availability" section). Figure 5 shows a heatmap of significant resistance genes and clinical status for selected patients in a single arm of the trial (left), and distribution of RCB values (right). We observed overlaps between the 2 lists; often, common driver genes were identified by both versions of the pipeline.

Since the pre-processing pipelines are instantiated in the Cloud, we run the pipeline on gene expression data, accessing gene signature scores for comparison, and generating plots through an R Shiny web interface. We use the MSigDB[26] database in BigQuery to extract annotated gene sets and are currently implementing this automation within our framework.

Case study: viewing pathology images from the cloud

The I-SPY 2 clinical trial collects pathology H&E images from pretreatment breast tumors on all consenting patients. Since the inception of the trial in 2010, we have gathered over 500 GB of images, a storage size that would be near impossible to work with for most local machines. Through the use of Google's command line tools, we are able to search and stratify these images into separate cloud stor-

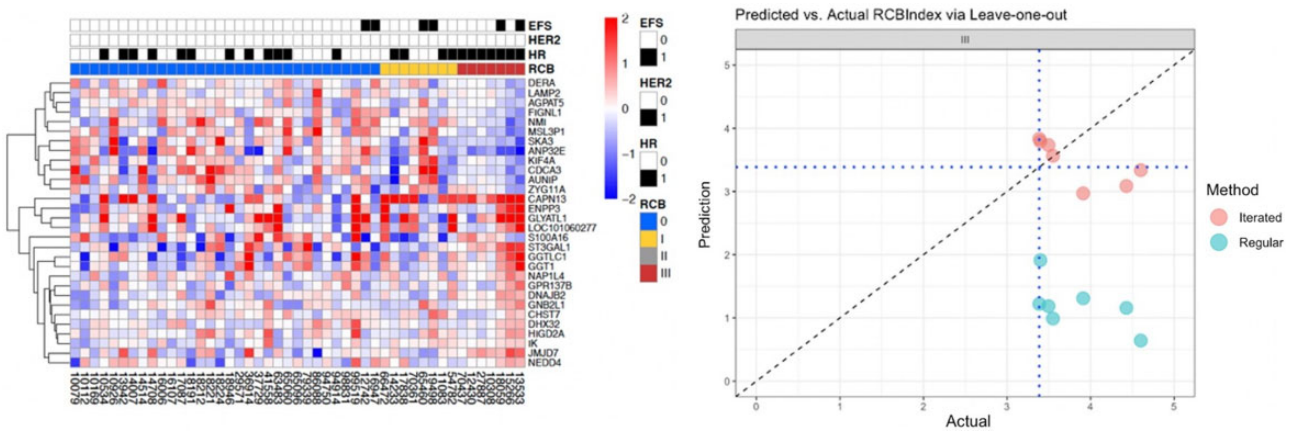

Figure 5. Heatmap of significant resistance genes and clinical status for selected patients in a single arm of the trial (left) and predicted versus observed RCB values via two statistical methods (right).

age buckets. Each bucket has independent permissions settings, which enables us to easily and securely share data for a multitude of investigators. Through Google Cloud Healthcare and QuPath, we can additionally provide these images without the need for local downloading. Many of I-SPY 2 investigators have proposals to view images from multiple drug arms, which without a cloud streaming solution, would result in downloading hundreds of images.

Linking QuPath to Google Cloud Healthcare's API requires 2 extensions. The first allows a connection from QuPath to your datasets (images) in GCP (Github posted in "Resource Availability" section).[27] This enables an authorization process, which is a way to ensure users also have appropriate access to a set of images. The second extension[28] converts images into a DICOM format (Github posted in "Resource Availability" section). The pipeline is capable of converting any file supported by openslide.org (eg, .svs, .tif, .scn, and .mrxs) to a DICOM whole slide image. The fields of radiology, cardiology, and ophthalmology have already adopted DICOM as the standard, and DICOM format seems to be an emerging standard for pathology images as well, despite most microscopes and scanners not producing DICOM images by default. The I-SPY 2 in-house Aperio scanner produces svs images, so this converter was necessary for us to upload DICOM images.

With the appropriate file formats and a cloud connection, investigators and pathologists have access to view their requested images, directly from the cloud, through QuPath (Figure 6) (Github posted in "Resource Availability" section). Figure 6 displays a QuPath connection to GCP, which allows users to view images directly from the cloud.

Although this process is in its nascent stage of the I-SPY 2 trial, several benefits seem clear. Pathologists and researchers alike will not have to download large sets of images for their research, saving both time and physical memory space. Sharing sets of data can be as easy as permission regulation, which will increase collaboration across varying professions. Additionally, annotations can be saved and synched to the cloud for all collaborators to view. Clinical personnel can benefit from this process by having pathologists review biopsy sections and stratify into stroma versus tumor. While QuPath is a great option for the I-SPY 2 trial, there are other methods available such as the DICOMWeb WSI Viewer,[29] which is a proof-of-concept image viewer that is also offered through the Google Health Care API. The Open Health Imagine Foundation (OHIF), which is generally more popular with radiologists, is also supported through GCP.

Additional pipelines available: VCF to BigQuery

A select number of I-SPY 2 tumor biopsy samples were sent for full exome sequencing and subsequent variant calling. Between the raw FASTQ, BAM/BAI files, processed VCF, and other flat mutational files, over 8 TB of data had been generated. GCP has made large-scale variant analysis relatively seamless with the use of Google native tools. Namely, loading and storing thousands of VCF files into BigQuery is simplified using Google Genomics' Variant Transforms tool (Github posted in "Resource Availability" section).[30] This tool can transform and load VCF files into Google's BigQuery platform giving researchers the flexibility to search through VCF files from one central table. Using custom SQL queries in BigQuery, researchers can search through variants from thousands of VCF files with relative ease. This can be useful as an initial quality assessment, such as checking known mutation rates in breast cancer for genes like TP53.

DISCUSSION AND CONCLUSION

I-SPY 2 demonstrates personalized medicine by leveraging the power of early endpoints and connecting the early endpoints to recurrence and survival, using genetic determinants to better predict response to emerging therapeutics. Importantly, over the past decade the trial has accumulated an incredibly rich and deep compendium of biomarkers that can inform how we learn and treat patients in the future. Currently, patients are adaptively randomized based on FDA cleared or standardly used biomarkers, and response to treatment based on both MRI volume change (an automated measure of functional tumor volume) as well as the rate of complete pathologic response. The objective of this biomarker-rich trial is to use all the available clinical and genetic information to predict the likelihood of an individual patient's response to a set of treatment options. Thus, it is mandatory for the trial to make available data to investigators using a common framework and applications.

While data sharing is often challenging given strict regulations, we are committed to making data from the trial public when possible, with and following primary publications, and accessible at any time to interested researchers. From the start, we developed a structured governance and oversight process through our DAPC. Though all data currently shared is double blinded and in compliance with Section 164.514(a) of the HIPAA Privacy Rule, there will be hurdles to address as the I-SPY 2 trials being to collect sequencing data. While it is common practice to remove patient identifiers from data before releasing it to the public, there are instances where identifica-

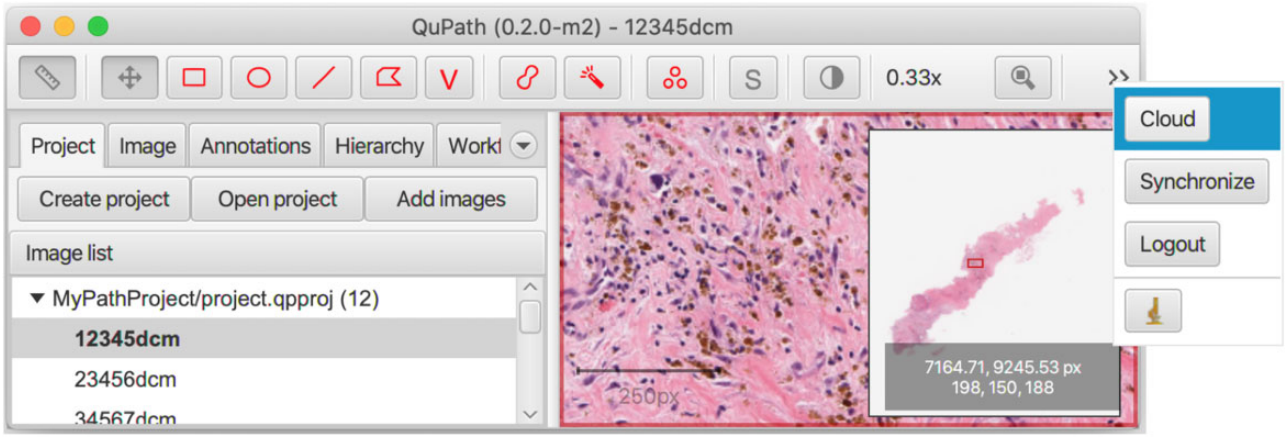

Figure 6. Connecting QuPath to Google Cloud Platform allows users to view images directly from the cloud.

tion of individuals happened anyway through combining anonymous gene sequences with genealogical databases and other public information such as age, state, or surnames.[31] These problems can be solved with further protections, but they require constant vigilance. Our DAPC must carefully review what patient data may be safely shared when in conjunction with sequencing data. Despite stricter regulations and considerations for sequencing data, we need more data sharing rather than less, because the benefits of publicly available data often outweigh the costs.[32] We will constantly work towards minimizing privacy-related risk, balanced against the benefits of the innovations that may arise from increased data availability through tools like PRoBE. Governance of I-SPY 2 data covers data to be released to the public, data shared directly with investigators, and any tools developed by PRoBE that access data on the cloud.

In the future, we hope to develop better risk stratification by integrating additional molecular, MRI, and pathology viewing tools into PRoBE to better elucidate biomarkers that predict drug response and outcome in these early-stage, high risk patients. PRoBE has proven useful in scalability, as we have begun directly ingesting pathology images from I-SPY 2 sites across the country. By sharing permission-based links to I-SPY 2 sites, we can directly receive pathology images into our GCP platform, send them through our pipeline to QuPath, and ultimately transmit images to our pathologists more rapidly. We highlighted the use of the TCGA database in one case study application, however ISB-CGC has uploaded data from many more public data sources. Future applications could incorporate data from these public databases, for example, Ensembl,[33] Gene Ontology,[34] or AACR GENIE[35] databases. For example, cross validation of any significant biomarker signatures, in a cohort outside of I-SPY 2, could strengthen our discoveries. In addition to cloud tools, future analytic processes include advanced sequencing workflows, automated partitioning and clustering pipelines, and incorporating additional machine learning algorithms for response prediction, tumor subtype analyses, and prediction of clinical outcomes. A robust cloud infrastructure can help integrate these tools and workflows to make valuable predictions across multiple agents. For that reason, we are continuously improving PRoBE to advance the way data is accessed and analyzed in the I-SPY 2 clinical trial. We are heavily invested in developing PRoBE so that we can continue to refine our ability to optimally target agents and help every patient achieve a complete response an excellent long-term outcome.

ACKNOWLEDGMENTS

We thank Dusan Zelembaba and team from Google Cloud for their instrumental help on pathology image integration. We thank Sheila Reynolds for fruitful discussions on the ISB-CGC platform.

AUTHOR CONTRIBUTIONS

NO and AB prepared the manuscript and figures. DLG and KA contributed knowledge, expertise, and language of the ISB platform. SV was integral in setting up the QuPath integration effort. AA, SA, LB-S, GLH, and LVV are members of the I-SPY Biomarker Working Group and provided many useful ideas. CY and DW provided use case ideas for implementation. LE is the Principal Investigator of the I-SPY 2 trial.

FUNDING

NO and AB were supported by NCI grant number 5P01CA210961. ISB-CGC has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, Task Order No. 17X053 under Contract No. HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY

Please contact the corresponding author amrita.basu@ucsf.edu for details. Custom PRoBE Applications: https://github.com/NickOGrady. ISB-CGC: http://www.isb-cgc.org. Datasets in BigQuery (ISB-CGC Resource): https://isb-cgc.appspot.com/bq_meta_search/. QuPath cloud connection: https://github.com/GoogleCloudPlatform/qupath-chcapi-extension. DICOM converter: https://github.com/GoogleCloudPlatform/wsi-to-dicom-converter. RStudio in cloud: https://github.com/MarkEdmondson1234. VCF to BigQuery Converter: https://github.com/googlegenomics/gcp-variant-transforms.

REFERENCES

1. Hinkson IV, Davidsen TM, Klemm JD, Kerlavage AR, Kibbe WA, Chandramouliswaran I. A comprehensive infrastructure for big data in cancer

research: accelerating cancer research and precision medicine. *Front Cell Dev Biol* 2017; 5:83.
2. Kerlavage T. Advancing a national cancer knowledge system. *Medium*, 2016.
3. Barker AD, Sigman CC, Kelloff GJ, Hylton NM, Berry DA, Esserman LJ. I-SPY 2: An Adaptive Breast Cancer Trial Design in the Setting of Neoadjuvant Chemotherapy. *Clin Pharmacol Ther* 2009; 86 (1): 97–100.
4. Cortazar P, Zhang L, Untch M, Mehta K, *et al*. Pathological complete response and long-term clinical benefit in breast cancer: the CTNeoBC pooled analysis. *Lancet* 2014; 384 (9938): 164–72.
5. Symmans WF, Wei C, Gould R, *et al*. Long-term prognostic risk after neoadjuvant chemotherapy associated with residual cancer burden and breast cancer subtype. *J Clin Oncol* 2017; 35 (10): 1049–60.,
6. Harrington D, Parmigiani G. I-SPY 2 — a glimpse of the future of phase 2 drug development. *N Engl J Med* 2016; 375 (1): 7–9.
7. Google. *BigQuery* Quotas and Limits. *Google*, 2020.
8. Soltoff B. *Tidy Data*. Chicago, IL: University of Chicago, 2021.
9. Wickham H, Data T. Tidy data. *J Stat Soft* 2014; 59 (10): 59.
10. Sherry ST, Ward M, Sirotkin K. dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res* 1999; 9 (8): 677–9.
11. Landrum MJ, Lee JM, Benson M, *et al*. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* 2018; 46 (D1): D1062–d1067.
12. RCoreTeam. R*:* A *Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2019.
13. Google. Compute engine machine types. *Google*, 2020.
14. Edmondson M. *Launch RStudio Server in the Google Cloud with two lines of R*, 2016. https://code.markedmondson.me/launch-rstudio-server-google-cloud-in-two-lines-r/
15. Google, *Cloud Healthcare API | Google Cloud*, 2019. https://cloud.google.com/healthcare
16. Bankhead P, Loughrey MB, Fernández JA, *et al*. QuPath: open source software for digital pathology image analysis. *Sci Rep* 2017; 7 (1): 16878.
17. Google. *HIPAA Compliance on* Google *Cloud Platform*, 2019. https://cloud.google.com/security/compliance/hipaa
18. Malin B. *Guidance on De-identification of Protected Health Information*, 2012. https://www.hipaajournal.com/de-identification-protected-health-information/
19. Hoadley KA, Yau C, Hinoue T, Cancer Genome Atlas Network, *et al*. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* 2018; 173 (2): 291–304.e6.
20. Reynolds SM, Miller M, Lee P, *et al*. The ISB cancer genomics cloud: a flexible cloud-based platform for cancer genomics research. *Cancer Res* 2017; 77 (21): e7–e10.
21. Kim J-Y, Lee E, Park K, *et al*. Immune signature of metastatic breast cancer: Identifying predictive markers of immunotherapy response. *Oncotarget* 2017; 8 (29): 47400–11.
22. Basu A, Mitra R, Liu H, Schreiber SL, Clemons PA. RWEN: response-weighted elastic net for prediction of chemosensitivity of cancer cell lines. *Bioinformatics* 2018; 34 (19): 3332–9.
23. Seashore-Ludlow B, Rees MG, Cheah JH, *et al*. Harnessing connectivity in a large-scale small-molecule sensitivity dataset. *Cancer Discov* 2015; 5 (11): 1210–23.
24. Basu A, Bodycombe NE, Cheah JH, *et al*. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell* 2013; 154 (5): 1151–61.
25. Yang W, Soares J, Greninger P, *et al*. Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res* 2013; 41 (Database issue): D955–61.
26. Subramanian A, Tamayo P, Mootha VK, *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *PNAS* 2005; 102 (43): 15545–50.
27. Zelembaba D. qupath-chcapi-extension. *GitHub*, 2019.
28. Zelembaba D. wsi-to-dicom-converter. *GitHub*, 2019.
29. *DICOMweb WSI Viewer*, 2019. https://github.com/GoogleCloudPlatform/dicomweb-wsi-viewer
30. Google. *Variant Transforms Tool*, 2019. https://cloud.google.com/life-sciences/docs/how-tos/variant-transforms.
31. Gymrek M, McGuire AL, Golan D, *et al*. Identifying personal genomes by surname inference. *Science* 2013; 339 (6117): 321–4.
32. Deming D. Balancing privacy with data sharing for the public good. *The New York Times*, 2021.
33. Hubbard T, Barker D, Birney E, *et al*. The Ensembl genome database project. *Nucleic Acids Res* 2002; 30 (1): 38–41.
34. Harris MA, *et al*. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* 2004; 32 (Database issue): D258–61.
35. AACR Project GENIE Consortium. AACR Project GENIE: powering precision medicine through an international consortium. *Cancer Discov* 2017; 7 (8): 818–31.