Sharing Clinical and Genomic Data in Cancer Research

INTRODUCTION AND OBJECTIVES

1 - Abstract

We aim to build a public platform for sharing data about cancer with the worldwide community of researchers, patients and citizen scientists. The vision is that a public data will help researchers around the world develop better models to study cancer, potential therapies, and to predict patient responses to therapies than if each participant maintained their data in isolation. The source of these datasets will be cancer patients themselves who are asked and choose to donate their genomic data and clinical records to this effort, with the assurance that the researchers will make every effort to protect the privacy of the patient.

The specific aims of this pilot project are 1) to determine the feasibility of consenting patients to donate their genetic (tumor) and clinical data; 2) to automate abstraction of clinical data, such as tumor pathology, disease staging, patient treatment history, treatment outcomes, and vital status from local medical records and clinical data bases; 3) to create algorithms to code and anonymize data for posting in a publicly accessible repository; and 4) to develop user-friendly web-based interface between the anonymized dataset and researchers.

2 - Background

DNA sequencing is transforming clinical oncology. Today, the treatment paradigms for lung cancer[1], melanoma[2], leukemia, breast cancer[3] and colorectal cancer[4] demand knowledge of somatic mutations prior to the initiation of first line treatment in the metastatic setting. We are presently at the inflection point between knowing nothing about the cancer genome in most of our patients and knowing all coding and relevant regulatory variants in the approximately 20,000 genes in the genome. While the utility, actionability and reimbursement of DNA sequencing in clinical oncology is debated from disease to disease, it is clear that the trajectory is towards more sequencing, done more deeply and more often in each and every patient.

In principle, it should be possible to dramatically accelerate medical progress by learning from the world's data on genome sequences and clinical phenotypes and illuminate the biological basis of cancer. By aggregating and analyzing large amounts of genomic and clinical data, it should be possible to discover patterns that would otherwise remain obscure. For example which mutations within a tumor predict response to treatment. Clinical interpretation of individual cancer DNA sequences will be powerfully enabled by comparison to extensive data on variation in cancer DNA sequence and phenotype. At present, it is generally not possible to predict which changes in DNA sequence lead to clinical consequences. When held against a large repository of other such data, however, robust patterns and relationships can be identified. Given the wide variety of disease endpoints, varied biogeography, and low frequencies of sequence variations, data from millions of samples will be needed.

Despite the clear benefits of data integration, the scientific and medical communities are not yet organized to seize this opportunity — nor are they on a path to do so, despite early attempts to start[5]. Currently, such data are typically analyzed in isolation, with sample sizes inadequate to make robust discoveries. Incompatible methods inhibit learning across datasets. Regulatory and ethical procedures could not anticipate, and thus were not designed to enable widespread comparison across studies and the sharing of information. Few clinical investigators have access to the analytical infrastructure needed to perform analyses for their patients; even the most sophisticated and well-resourced medical centers find it difficult to keep pace with rapidly evolving tools and pipelines.

There is a great unmet need in sharing knowledge about both the temporal sequence and the clinical settings in which somatic mutations occur in tumors. For example, unbiased sequencing of the tumors of lung cancer patients during the course of their treatment would reveal a reasonably high fraction of mutations at the T790M residue of the *EGFR* gene, a variant associated with resistance to treatment with tyrosine kinase inhibitors. We now know that this mutation is virtually never found prior to exposure to EGFR inhibitors, indicating that therapies targeting specific gene products can drive mutational resistance. Increasingly, many cancer patients are now receiving two, three or more lines of therapies, the outcomes of which may be due to the patients' mutation profile at baseline as well as to the selective pressure of the therapy. It will be imperative to collect genomic and clinical data longitudinally, in real time, to capture the complexities of the interaction between the cancer cells and their therapeutic exposures, and the effect on the health of the patient.

In order to accrue and share such complex cancer genomic and clinical data, it will be paramount to protect the rights and privacy of cancer patients. Recognition of the patients' ownership of their own histories and personal health information, and providing them with an opportunity to proactively and securely participate in this project is the first priority of this endeavor.

Investigators:

Eric Collisson, M.D. (UCSF) Principal Investigator - sample collection, patient consent

David Haussler (UCSC), PhD, Co-Investigator - web platform for data sharing

Ann Griffin, (UCSF) Co-Investigator - clinical data collection

Lawrence Fong, MD, (UCSF) Co-Investigator - sample collection, patient consent

Pamela Munster, MD, (UCSF) Co-Investigator - sample collection, patient consent

Amy M. Lin, MD, (UCSF) Co-Investigator - sample collection, patient consent

Hope Rugo, MD, (UCSF) Co-Investigator - sample collection, patient consent

Jessica Van Ziffle, PhD (UCSF) Co-Investigator - genomics data collection, curation, transfer

Nancy Joseph, MD, (UCSF) Co-Investigator - genomics data collection, curation, transfer

Maximilian Haeussler, PhD, (UCSC) Co-Investigator - web platform for data sharing, genomics data collection, curation, transfer

Ted Goldstein, PhD, (UCSF) Co-Investigator - web platform for data sharing, genomics data collection, curation, transfer

3 - Study Objectives

1. To demonstrate the feasibility of obtaining informed consent of patients to donate their genomic data and clinical records to the data sharing project.
2. To develop informatic tools to automate extraction of relevant clinical data, including but not limited to diagnosis, treatments, treatment outcomes and vital status, from local medical records and clinical databases.
3. To create algorithms to code and anonymize data for posting in a publically accessible repository.
4. Develop a user-friendly internet interface to share the de-identified/anonymized information with other researchers

Primary Outcomes

Confirmation of secure data transfer of anonymous, patient centric genomic and clinical data to Cancer Gene Trust

Secondary Outcome Variable

Rate of patient consent to data sharing

4 - Location

University of California San Francisco.

University of California, Santa Cruz.

STUDY DESIGN

5 - Research Design & Methodology

This is a prospective, longitudinal pilot study conducted in collaboration with patients treated for cancer at UCSF. Patients will be consented by Investigators or research coordinators with proper training. The sequencing results of somatic multi-gene panels and selected clinical data of patients who participate will be processed bioinformatically and initially made available with anonymized clinical annotation through the Cancer Gene Trust (CGT) on the internet. The CGT is a Global Alliance for Genomics and Health (GA4GH) project with an initial pilot network of servers including NKI (Netherlands), Melbourne, AUZ, and Singapore. Details on the project including a white paper, leadership, and FAQs can be found here: https://genomicsandhealth.org/work-products-demonstration-projects/cancer-gene-trust

The current pilot network can be viewed at: http://search.cancergenetrust.org/

We will record the number of potential patient data donors approached by investigators (screened) in the OnCore database. This will be the denominator to achieve the stated secondary outcome.

6 - Study Duration

The enrollment period will be 2 years (24 months). Cancer outcome data may be collected longitudinally by chart review, imaging review and clinical-adjudicated clinical outcome only until the end of the study. Total study duration will be 5 years.

7 - Inclusion and Exclusion Criteria

1. Inclusion Criteria
    a. Age 18 years of age or older.
    b. History of cancer or active malignancy.
    c. Ability to understand and provide informed consent.
    d. Next generation sequencing multi-gene panel assay results available or planned to inform cancer clinical care.
2. Exclusion Criteria
    a. Inability to provide consent.
    b. Objection to sharing data about themselves or their tumor with the larger research community.

8 - Vulnerable Populations

Although vulnerable populations are not specifically excluded from this study, there will no specific efforts to target these populations.

9 - Subject Accrual

Potential cases will be identified by screening patients in the principal investigator's and co-investigators' clinics by the study team. The team is comprised of clinical research coordinator, the study investigators and associated personnel. Ninety-nine patients will be included in this study.

10 - Subject Recruitment & Screening

Information to determine eligibility of patients to participate in the data sharing project will be collected by the study team prior to obtaining informed consent.

All adult patients meeting inclusion/exclusion criteria and scheduled to be seen in selected UCSF clinics will be screened for admission into the study by study investigators and team members. All prospective research subjects will meet with a study team member, the study will be explained, questions answered and the Informed Consent Process will be initiated, either by a study investigator or by a clinical research coordinator with documented consent training.

11 - Data

This project will collect two main types of data and optionally a third type:

1. Genomic sequencing data from patient tumors will be provided by genetic testing laboratories. For data provided by the UCSF lab, the data will not include germline variants, as both germline and tumor are sequenced and any variant seen in the germline is subtracted from the result. For data provided by certain outside tumor sequencing companies, only the tumor is sequenced so the data may include variants that may have originated in the germline.
2. Clinical data about the patient will be provided by the cancer registrar at UCSF from data gathered by the cancer registry as a part of routine reporting to the State as per California Law. This data will be de-identified as detailed below before submission. Additionally, we will obtain de-identified clinical information from the UCSF OMOP database as further detailed below.
3. In some cases, patients may have had outside third parties such as friends, family, or for-profit companies or not for-profit organizations collect manage and/or abstract their medical records. In such cases and if the patient is willing, we may in the future provide portals through which the patient or his voluntary or paid designate may upload this accessory data to a database we in part manage and use for research. UCSF investigators will not prevent the patient from sharing data with the research community if he chooses.

STUDY PROCEDURES

12- Procedures

Consent

This will be performed in person by investigators or their trained coordinators.

Data Collection and De-Identification:

The cancer registrar will receive from genetic testing laboratories genomic data files and provide these to an application the research team is developing that will automatically extract the tumor variant data for submission. In the case of Foundation Medicine this is the ResultsReport/ResultsPayload/variant-report from their XML file. We will obtain clinical information from the cancer registrar who will export from their registry software (CNEXT) a data file to the submission application which will extract the fields listed in the appendix that have been deemed 'de-identified' by the IRB, and only those will be included in the submission. Additionally, we will obtain clinical information from the de-identified EMR at the University of California, San Francisco which are stored in the Observational Health Data Sciences and Informatics (OHDSI; https://www.ohdsi.org/) Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM; https://www.ohdsi.org/data-standardization/the-common-data-model/). We have developed a pipeline to automatically extract relevant de-identified clinical information from these records for integration into CGT that will function with any de-identified clinical EHR data in the CDM format. Clinical data from both of these sources will be de-identified in accordance to our protocol and the UCSF IRB standards. This data are also listed in the informed consent document. For example 'age at diagnosis' or 'disease staging' will be included, 'name', or 'birthdate' will not.

Data Upload

Sequence and clinical summaries will be uploaded from UCSF to the CGT under the auspices of the GA4GH (http://ga4gh.org/#/[6]). UCSC will develop a SQL query matching the approved clinical fields allowing export from the cancer registrar's system (CNeXT) and/or the UCSF OMOP database. The exported meta-data along with the genomic file from Foundation Medicine will then be submitted to the UCSF CGT steward via a simple drag and drop web application.

We will verify that the cryptographic hash of the original files submitted by UCSF investigators matches the hash of the data in the UCSF CGT steward server. Additionally we will verify that those data can be accessed, mirrored, and downloaded by another CGT steward, and that the downloaded files match the strong cryptographic hash of the original files. The latter will prove sharing beyond the UCSF's steward server.

The GA4GH along with UCSC will provide the necessary technical and operational support for submission and hosting for UCSF's CGT server.

STATISTICAL PLAN

13 - Sample Size Determination

We expect to enroll up to 99 patients in this pilot program. We will enroll for two years. We expect each investigator to enroll 5-10 patients a year.

14 - Statistical Methods

We are unable to conduct formal statistical analysis given the pilot nature of this study.

SAFETY & ADVERSE EVENTS

15 - Potential Risks

The main risk involves loss of confidentiality. Subject confidentiality is protected by selecting only data fields that cannot be used to identify a patient. Only cancer, not germline, mutations are shared, which makes patient identification from mutational data alone very unlikely. It is illegal to discriminate on the basis of genomic or genetic data, but loss of privacy may cause distress to the participant or family members.

16 - Potential Benefits

Discovery of novel and innovative methods for the sharing genomic and clinical data has the potential to significantly alter the current management of patients with cancer. This study may lead to the development of new methods, standards and future directions in cancer data sharing.

17 - Confidentiality & Data Storage

Precautions will be taken to ensure that strict confidentiality is maintained within the research team. All basic research materials will be inaccessible to anyone other than the investigators at UCSF and UCSC. Unique study ID numbers (unique for this project) will be prepared for all source documents.

Study data with Private Health Information PHI or potentially identifiable information will be maintained on a study-dedicated computer whose files are password protected, on a portable device that is encrypted, on a secure network, and on a secure/encrypting website that is available only to UCSF researchers. Data will be shared in a de-identified manner with outside investigators.

To ensure to as great an extent as possible that no identifiable data are uploaded publically, we have developed an automated procedure for de-identification. The details of the process are outlined in the 'Data De-Identification' section above.

For the purposes of designing optimal interfaces for both collecting and disseminating patient data, third party, commercial entities may be invited to collaborate. Institutional review will be sought prior to sharing any patient data with commercial entities, but patients will not be re-consented to share with commercial entities.

18 - Data Safety & Monitoring

The study coordinator will review accrued data manually (data curator), and will bring any discrepancies to the attention of the study team for investigation and resolution.

19 - Risk/Benefit Ratio

Although there is no immediate, direct benefit to the individual participants of this study, there is considerable potential benefit to future subjects, including participants in this study, and to the community as a whole. Physicians will be able to better share data about cancer.

Our ideal scenario is that outside investigators from other institutions find a data record with a mutation that was not actionable at the time of diagnosis at UCSF but that has become actionable later. These outside investigators can then contact UCSF and suggest enrollment of the patient into a clinical trial or suggest treatment with a recently approved drug.

REGULATORY & ETHICAL CONSIDERATIONS

20 - Informed Consent

Informed consent will be obtained during the patient's inpatient or outpatient visit at UCSF. The goals of the study will be described by the participating investigator, a study nurse or qualified clinical research coordinator at UCSF. The consent document will be reviewed and provided to prospective subjects. Comprehension of the consent document will be assessed by clinical investigators. Participants must understand the consent process in order to participate.

STUDY FINANCES

21 - Compensation to Subjects for Participation

No compensation will be provided for the participants.
Study participants will not share in intellectual property emanating from this project.

22 - Conflict of Interest

There is no conflict of interest for the principal investigator or any member of the study staff.

PUBLICATION PLAN

23 - Publication Plan

Results of the proposed research may be presented at conference sessions within the Cancer Center, and in various oral, print and electronic publications. The principal investigator holds primary responsibility for presentation of the results of the study.

References

1. Lynch, T.J.*, et al.* Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *N Engl J Med* **350**, 2129-2139 (2004).
2. Chapman, P.B.*, et al.* Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *N Engl J Med* **364**, 2507-2516 (2011).
3. Slamon, D.J.*, et al.* Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpressed HER2. *N Engl J Med* **344**, 783-792 (2001).
4. Van Cutsem, E.*, et al.* Cetuximab and chemotherapy as initial treatment for metastatic colorectal cancer. *N Engl J Med* **360**, 1408-1417 (2009).
5. Grossman, R.L.*, et al.* Toward a Shared Vision for Cancer Genomic Data. *N Engl J Med* **375**, 1109-1112 (2016).
6. Global Alliance for, G. & Health. GENOMICS. A federated ecosystem for sharing genomic, clinical data. *Science (New York, N.Y* **352**, 1278-1280 (2016).

Appendix

Information that will and will not be shared as part of this study

	Information that **will** be shared	Information that **will** <u>**not**</u> be shared
Locations:	• Birth • Diagnosis • Residence • Death	• City or towns • Addresses
Dates	• Age at diagnosis • Year in which treatment began • Last patient contact • Age at time of Death	• Date of Birth • Date of Death
Durations	• Hospital admissions/discharges • Diagnostic procedures • Disease staging • Surgeries • Treatments • Recurrence	
Health	• Diagnosis • Diagnostic procedures • Disease staging • Presence of other diseases or conditions • Summary of previous treatments (including surgery, chemotherapy, radiation therapy, hormone therapy, and immunotherapy) • Summary of treatments at UCSF • Summary of reconstructive surgeries • Current health status • Cause of death • Tumor genetic data • Tumor imaging data • Photomicrographs of your tumor cells	• Medical record number(s) • Health plan numbers • Photographs of your face • Physician name(s)
Demographic and Other	• Gender • Race/ethnicity • Religion	• Name • Phone numbers • Email addresses • Social Security number