TITLE: Patient-driven efforts to liberate clinical cancer genomic data

AUTHORS: Stephanie A. Wankowicz^{1,2}, Brooke Maile³, Ken Deutsch³, Adam Hayden³, Erika Brown³, Tom Marsilje³, Beth Caldwell³, Tania Simoncelli⁴, Eliezer M. Van Allen^{1,2}

IN MEMORIAM: Tom Marsilje, Beth Caldwell

AFFILIATIONS:

1. Cancer Program, Broad Institute of MIT and Harvard; Cambridge, MA
2. Medical Oncology, Dana-Farber Cancer Institute; Boston, MA
3. Patient partner
4. Advocacy partner (current affiliation: Chan Zuckerberg Initiative; Palo Alto, CA)

CORRESPONDENCE:

Eliezer M. Van Allen, MD
Dana-Farber Cancer Institute
450 Brookline Avenue, LC9235
Boston, MA 02215
eliezerm_vanallen@dfci.harvard.edu

KEYWORDS: Cancer genomics, data liberation, patient engagement

WORD COUNT: 2,930

ACKNOWLEDGEMENTS: This project currently has no funding support and was completed with the generous donation of time and effort on behalf of the patient-partners represented in this article.

DISCLOSURES:

Eliezer M. Van Allen:

Consulting – Tango, Genome Medical, Invitae, Illumina, Foresite Capital, Dynamo
Unpaid consulting/travel – Roche/Genentech
Equity – Tango Therapeutics, Genome Medical, Syapse, Microsoft
Research Support – Novartis, BMS

ABSTRACT:

Cancer patients' data, including clinical and genomic sequencing, typically reside within incompatible electronic health records (EHR) and/or proprietary databases in clinical laboratories or companies. Multiple academic and commercial consortia have initiated "top down" approaches to encourage data sharingⁱ, though these consortia represent a small percentage of the cancer patient population and may not capture the genomic and phenotypic diversity of the larger cancer community. Critically, patients have the unique power to liberate and share their own data, since the Health Insurance Portability and Accountability Act (HIPAA) gives patients a right to access all data in a computable format. Here, we partnered with patients to examine the feasibility of leveraging the HIPAA right of access provision to access their cancer genomic data in a computable format from both academic and commercial laboratories (n = 5 patients; 3 academic labs, 5 commercial labs). We observed wide heterogeneity in responses to requests, deliverables to patients, and patient effort required to access data. Multiple labs reported that these requests were the first of their kind and subsequently instituted procedural changes to facilitate data sharing with patients. Furthermore, individual patients expressed significant meaning attributed to having access to and control over their cancer genomes. While this experiment exposed multiple difficulties for patient genomic data access, we argue that these difficulties can be overcome with a patient-centered platform to facilitate these efforts. Toward that end, we propose and describe a patient-driven cancer genome collective to directly address this need and empower data liberation and donation to advance cancer research and patient empowerment.

INTRODUCTION:

The Health Insurance Portability and Accountability Act (HIPAA) of 1996 gives patients the right to request, review, and obtain copies of their protected health information (PHI)ⁱⁱ. This right of access applies to information that exists within the "designated record set (or sets)" and includes any "information used to make decisions about individuals," with limited exceptionsⁱⁱⁱ. A 2014 amendment expanded this right of access to include information maintained by clinical laboratories^{iv}. Based on an additional clarification in 2016, this specifically includes a right to access genetic information: "...the full gene variant information generated by the test, as well as any other information in the designated record set concerning the test"^v.

HIPAA also states that covered entities, such as hospitals or clinical labs, must provide patients with their records in the form and format requested by the individual if records are "readily producible" in that format^{vi}. Specifically, if the records are kept in an electronic format, the data must be provided in an electronic format^{vii}. Additionally, to every extent possible, US Health and Human Services (HHS) expects covered entities to provide the individual with their genetic data in a machine-readable format^{viii}. HIPAA also states records must be given to the patient within 30 days of request^{ix} and gives patients the right to transmit a copy of their PHI, including clinical or genetic information, directly to a third party^x. Currently, the exact data format and data transmission procedures are left up to individual labs.

In general, this rule is underutilized for clinical care. The lack of action is partially due to individuals having an incomplete understanding of patients' rights, along with bureaucratic difficulties of obtaining the records^{xi}. More recently, a nascent patient movement has grown out of frustrations about the difficulties in accessing and sharing patient medical data (with a focus on electronic health record data)^{xii}, and likely others will emerge.

In contrast to clinical EHR data requests, little movement has been made towards using this right of access to obtain genetic data. One of the first attempts to obtain PHI genetic data was by a group of individuals that had undergone germline genetic testing by Myriad Genetics for cancer risk genes, including *BRCA1* and *BRCA2*. For this attempt, requests were made for the "entire

Designated Record Set” and included specific language requesting BAM and VCF files. Myriad Genetics initially denied these requests, stating “...those amendments [2016 amendment to include sequencing information as part of PHI] did not expand the definition of the Designated Record Set, e.g. to include any of the additional items listed in your letter. We have therefore not provided information or files outside of the above detailed Designated Record Set.”^{xiii} Only after learning an administrative complaint would be filed against them with the HHS Office of Civil Rights did Myriad Genetics comply with the request. Nevertheless, they only provided a PDF version of the report, which was not machine readable, and they continued to refuse to acknowledge that it is a patients’ right to have this data. After multiple additional exchanges, patients were able to obtain a machine readable variant call list, however it still was missing key information (i.e. chromosome position) that are necessary to properly analyze these data^{xiv}.

The 2016 Myriad Genetics effort informed patient-driven genetic information liberation of germline genetic data. In parallel, somatic genomic data from tumors has become increasingly utilized in clinical cancer care. We sought to test the ability of individual patients to obtain their own cancer genomic data from academic and/or commercial labs that was used for clinical purposes. Toward that end, we partnered with individuals who approached genetic sequencing labs to attempt to liberate their data and share their experiences.

METHODS:

We partnered with five cancer patients on a self-directed information gathering exercise to document their attempts to obtain clinically generated genetic/genomic sequencing data. Between 1/2016 and 8/2017, patients themselves requested genetic/genomic sequencing data from eight different clinical labs (three academic clinical labs and five commercial labs). They first sent out an email (**Appendix A**) to the general client representative or personal contact at each clinical genetics lab requesting the following items:

- 1) Full list of all variants identified (benign, unknown significant, and/or pathogenic)
- 2) Results of any rearrangements, insertions, and/or deletions.
- 3) Underlying data formats (genome build) and reference files (bed files), where applicable.

The email also contained language from HIPAA explaining the patient’s right of access to individual health information, including their genomic data. The file format of the data was not specified as it was unclear what kind of data (from a BAM file that contains all raw data, to a VCF or TSV file that comprised of a processed list of genetic variants) the clinical genetic/genomic labs would be able to provide. As per regulations, data was requested to be sent within 30 days in a machine-readable format to the patient’s email. Patients then followed up as necessary and as they were able, either electronically or through phone communications.

RESULTS:

In total, five patients made requests to eight clinical sequencing vendors (three academic labs and five commercial labs) (**Fig. 1**). Two academic clinical labs provided a full variant call format (VCF) file containing all somatic genomic variants, although the timeline and effort required to obtain the data varied greatly (range: brief email or an in-person meeting to sign a legal release form). A third academic lab would only provide a scanned PDF of the test results despite multiple inquiries.

Regarding the five commercial clinical labs, requests were made to three somatic genomic labs and two germline genetic labs. Among somatic genomic testing labs, responses ranged from immediate distribution of raw genomic data (in the form of a BAM file), distribution of a BAM file without necessary reference files to enable interpretation after

Figure 1. Summary of request attempts for academic and commercial labs. File types include Variant Call Format (VCF), Portable Document Format (PDF), Tab Separated File (TSV), and Binary version of Sequence Alignment Map file (BAM). Num. = number, Mo. = months.

repeated inquiries, and distribution of a TSV file (provided only after the request was reassigned to senior members of the vendor). Responses from the two germline genetic labs resulted in one outright refusal, while the other provided a file that did not include reference files for interpretation. Again, the effort required to obtain this data varied greatly, with some companies sending data within a day of the request and others only sending a PDF after 16 different email and phone communications. Additionally, patients received several data formats, including BAM files, VCFs, and PDFs, via physical thumb drives through the mail and emails with links to data in cloud storage. With one exception, raw data was sent without concomitant reference files or information to indicate how the data was processed. Narrative details of specific requests are provided in **Appendix B**.

After completing data requests, we reached out to representative labs to evaluate the impact of individual patient requests on internal lab policy. Multiple labs reported that the patients making requests were the first such patients to do so, and as a result there were no pre-existing formal processes in place for handling such requests. For example, to address this challenge prospectively, the Dana-Farber/Brigham and Women's PROFILE initiative created a process to deliver VCF files to patients (after a discussion with a lab representative and agreeing to waiver), typically on a flash drive. Among commercial labs, Personal Genome Diagnostics provided the patient with raw data two days after the request and had processes in place to support this form of data release. Similarly, Foundation Medicine reported that patients can complete a data release form through a client services group to request and authorize the release of their data, which may include structured variant level data on a case by case basis. From Guardant Health, the laboratory director became involved in one patient's request from 2017; after supplying the raw data to the patient and noting this was the first such request, the director described instituting an internal standard operating procedure to deal with future patient requests for their raw data and ensure efficiency. Notably, the majority of labs expressed a willingness to create systems for data sharing, though no uniform process for data sharing was described between labs.

PERSONAL IMPACT OF REQUESTING TUMOR GENETIC DATA:

During the process of the data requests, the participants were asked to share their thoughts about having access to this data. Transcripts of their remarks are provided below:

Ken Deutsch: *I was diagnosed with bladder cancer when I was 52 (four years ago) and within months learned that it had metastasized to my bones. Having lost multiple family members to cancer, including my father when he was 52, I was tested and found that I was positive for a BRCA mutation. Just by coincidence, the standard of care treatment for metastatic bladder cancer*

is a type of chemotherapy (cisplatin) that has a higher chance of response for those with a BRCA mutation. My tumor showed both the BRCA mutation that is in my family and a second mutation somatically.

As a cancer survivor who happened to respond well to treatment, I view it as my obligation to help unlock the mysteries of which mutations might be helped by which treatments. In my own experience, I have found that some labs clearly want to share data, but do not have the systems in place to make sure data sharing happens easily and just want to be sure they have the right safeguards in place. However, I have also found that other labs have put whatever obstacles they can find in the way of sharing with patients our own data. While the first group of institutions can benefit by putting better systems in place, the second group has developed business models around hoarding data that is both ethically and legally suspect.

Brooke Maile: *It is difficult to think straight when your doctor has just told you that you have cancer. I learned that angiosarcoma was extremely rare, aggressive and often fatal. A forward thinking oncologist suggested that I send my tumor sample to a specialized genetic testing lab to determine the exact mutation. That made perfect sense to me and gave me hope that science had evolved such that my treatment could be personalized to my particular type of cancer and specifically to my tumor.*

I felt excited and hopeful when I contacted the genetic testing lab for my test results. However, I was stunned when they told me they would not give me my data. I remember wanting to scream because surely, they were kidding. My thinking was: 'My body, my tissue, my data!'

Fortunately, my doctor got the information quickly and put it immediately into my hands. The good news: I know the exact genetic mutation – STK11. The bad news: four years later, this valuable data is not being utilized because labs refuse to share information. I know because I spent months trying to get my data. I appreciate that there are challenges but system platform issues can be overcome if egos are set aside and everyone has a willingness to share potentially lifesaving data and information.

We say we want to cure cancer to keep our loved ones alive and healthy. The only way to do this is for patients, testing labs, clinicians and scientists to share complete and usable information. We have the power and the data – let's make this happen!

Adam Hayden: *I was diagnosed with glioblastoma (GBM) in June 2016. This aggressive form of brain cancer is well-known to the medical oncology community for its heterogeneity, therapeutic resistance, grim prognosis, and neurological decline. The Cancer Genome Atlas (TCGA) selected GBM for its first full sequencing, given the complexity just mentioned. This is to say, the genetic profile presents an obstacle to the development of efficacious therapies. Connecting genomic data to clinical outcomes presents a significant opportunity to accelerate research and improve patient care. My neuro oncologist informed me of the mutations and methylation status of my particular tumor and the bearing on prognosis. Though news of an IDH1+ mutation status was "good," for example—as good as can be for such a lethal cancer, this genetic information represents only a "snapshot" of my tumor, restricted to a static point in time.*

Innovations on the horizon such as liquid biopsies to monitor disease progression and changes in mutation status promises to benefit patients, but it is not yet clear to me how I may access this data in the future, nor how data may be aggregated. Educating and informing patients about their right to access and the benefit to their care and care for others with similar diagnoses is essential

to develop a framework as genomics continue to improve in scope, efficiency, and cost. Our call to establish standards benefits patients today and builds structure for the future.

Universally, the participants strongly believed that they should have full access to this data and it should be easily accessible.

TOWARD ACCESSIBLE CANCER GENOMIC DATA LIBERATION:

The results of this pilot demonstrate that while patients have the right to obtain all forms of their cancer genetic/genomic data, numerous issues prevent effective data liberation, including: 1) inconsistent interpretation of the HIPAA right of access rules amongst labs generating these data, 2) lack of guidelines to inform the proper mode of data sharing, and, most importantly, 3) lack of scalable mechanisms for patients to initiate, receive, and act upon the data they liberate.

We believe these challenges can be overcome through a cancer genome *collective* of patients empowered to exercise their HIPAA rights and liberate these data. Based on the success of online communities accessing and sharing genetic data from germline genetics products (e.g. <https://dna.land/>), we propose creating a sustainable patient-partnered project whereby cancer patients can learn how to request and access their genomic data, as well as use that data for purposes they deem appropriate (**Fig. 2**). We envision this platform performing three major functions: 1) Educate patients about their HIPAA right of access and help patients identify the genetic/genomic tests they have received for data requests; 2) Give patients the option to donate

Figure 2. A cancer genome collective that empowers patients to liberate and share their genomic and clinical data: (1) Educate patients about right of access; (2) Consent to research project to donate data; (3) De-identify and aggregate data for further research.

their data (genetic, genomic, and/or clinical) for research; and 3) De-identify and aggregate that data with corresponding clinical data to be used by the research community to advance cancer research. Patients can also consent to be re-contacted in the case of eligibility for a clinical trial or a specific cancer therapy based on the genetics/genomics they have liberated and donated, since the data was generated in a clinical setting.

To make data access feasible, this platform would be created in close collaboration with participating genetic/genomic labs to establish guidelines for fulfilling patient requests for genetic data. This effort would include enforcing standard operating procedures for handling and acting upon patient data requests. As genetic/genomic data is often stored in structured, machine readable formats (e.g. VCFs), these labs could create infrastructure for encrypted transmission of interoperable data to various entities at the direction of the patient, such as clinical providers and the patients themselves.

Such an entity would also be suited to having educational content that can overcome the general confusion about HIPAA right of access identified in this project. Specifically, this web presence

would have consolidated multimedia educational content to advise all impacted parties – patients, providers, hospitals, laboratories – about patient data rights, modes of data sharing, and meaningful ways these data can make an impact.

Finally, the recent decision by Centers for Medicare & Medicaid Services (CMS) to reimburse for somatic genetics testing^{xv} from verified labs for patients with advanced cancers will greatly expand the number of cancer patients in the community (non-academic medical center) setting who will have access to these tests. The lack of concomitant requirement by CMS that genetics labs make the data freely available to the research community may be rectified by this patient-driven data-sharing movement to complement and reinforce data sharing activities happening from academic medical centers (i.e. AACR Project GENIE). As more patients to exercise their right of access to obtain their data, the creation of a patient-driven web presence that addresses these challenges in data liberation – education, guidelines, access – will further an ecosystem of data sharing to accelerate new discoveries that may directly impact the care of cancer patients.

APPENDIX A: Sample data request letter

To Whom It May Concern,

I am writing to request my genetic test records, including list of all variants identified (benign, unknown significant, and/or pathogenic) and results of any large-scale rearrangement/insertion/deletion analysis conducted with your company, [GENETIC SEQUENCING COMPANY/CLINICAL LABORATORY].

Recent regulations and guidance from the U.S. Department of Health & Human Services make clear that under HIPAA, patients have a right to genetic information that laboratories obtain during the testing process as well as any information used to make decisions about genetic test results. [79 Fed. Reg. 7,290 (Feb. 6, 2014); 65 Fed. Reg. 82,606 (Dec. 28, 2000); U.S. Department of Health & Human Services, Individuals' Rights Under HIPAA to Access Their Health Information; <http://www.hhs.gov/hipaa/for-professionals/privacy/guidance/access/index.html>].

Please send this information to me as soon as possible, and in any case within 30 days, in an electronic, machine-readable format via email at [patient email].

Sincerely,

[Patient Name]

APPENDIX B: Detailed narratives of data request attempts

Requests for genomic data from academic labs

Two patients requested data from three academic genomic testing labs. One patient received their full variant call format (VCF) file via email within 48 hours of requesting the file, while another patient never received their machine-readable sequencing data. Another academic site was only able to provide a scanned PDF after filling out multi-page handwritten form. From the third academic medical center, the patient had to fill out additional handwritten forms and was initially told that this institution had no information on the genetic sequencing data. After a follow-up email to the sequencing lab, the patient received a phone call from a representative stating that 1) this individual was the first patient who had ever requested their data, 2) the hospital acknowledged that they were unsure what to do with the request and had checked with lawyers about what they could give the patient, and 3) the hospital would be happy to give the patient their data in whatever form the patient requested. However, the patient was then required to come to the hospital in person to sign two legal release forms prior to receiving a variant call file (VCF) on a thumb drive.

Requests for genomic data from commercial labs

In total, data was requested from five commercial clinical labs, including three somatic sequencing commercial labs and two germline sequencing commercial labs. From the germline sequencing commercial labs, one vendor stated that while they would like to provide the data, they did not have the infrastructure possible to do so and denied the request. A second company responded after three email attempts stating that the patient had not filled out the proper paperwork. The patient promptly filled out the forms, but did not hear back from the company for another 38 days. The company then sent the patient the results in an encrypted file that the patient could not open. Finally, the company sent the data in a format the patient could use. However, the data received did not include pertinent genetic reference information, such as start and stop site or cDNA change, making it difficult to interpret with data from other labs.

Regarding somatic genetic testing labs, one lab stated that the file size was too large to transmit electronically, and mailed a thumb drive to the patient that contained a BAM file. The patient requested information, including the alignment algorithm and BED file to enable effective interpretation of the raw data, but never heard back from the company. Another lab, contact through the client services office, initially stated that they could not share the data with the patient as the data was proprietary. After replying with the HHS guidelines, the lab provided a BAM file. When data was requested from a third commercial lab, the lab responded immediately with options to send a BAM file electronically or through the mail.

ⁱ AACR Project GENIE Consortium. "AACR Project GENIE: powering precision medicine through an international consortium." *Cancer discovery* (2017).

ⁱⁱ The Health Insurance Portability and Accountability Act (HIPAA), enacted by Congress in 1996, sought to address challenges for protecting the confidentiality of health information posed by advances in health information systems technology. The Act called on Congress to enact a medical privacy statute and provided the HHS Secretary with the authority to promulgate regulations on health care privacy in the event that Congress did not act within three years. Congress did not meet this imposed deadline, so HHS issued its proposed "HIPAA Privacy Rule" in November 1999 and the final rule in December 2000. FR 65 (December 28, 2000), 82469.

ⁱⁱⁱ 45 CFR 164.524(a)(1) – (a)(3)

^{iv} Department of Health and Human Services. *CLIA Program and HIPAA Privacy Rule; Patients' Access to Test Reports; Final Rule*. FR 79 (February 6, 2014), 7290-7316.

^v Individuals' Right under HIPAA to Access their Health Information 45 CFR § 164.524

^{vi} 45 CFR § 164.524(c)(2)

^{vii} 45 CFR §164.524(c)(2)(ii)

^{viii} HHS considers machine readable data to mean digital information stored in a standard format enabling the information to be processed and analyzed by computer. This would include providing the information in the format of MS Word or Excel, text, HTML, or text-based PDF, among other formats. FR 78 (January 25, 2013), 5631.

^{ix} 45 CFR§ 164.524(b)(2)

^x 45 CFR § 164.524(c)(3)(ii)

^{xi} Lye, Carolyn T., et al. "Assessment of US Hospital Compliance With Regulations for Patients' Requests for Medical Records." *JAMA Network Open* 1.6 (2018): e183014-e183014.

^{xii} [Getmyhealthdata.org](https://getmyhealthdata.org)

^{xiii} Personal Correspondence, Tania Simoncelli

^{xiv} Personal Correspondence, Tania Simoncelli

^{xv} Centers for Medicare and Medicaid, Administrative File: CAG #00450N