Operationalizing Tumor Molecular Profile Reporting in Clinical Workflows and for Translational Discovery

By

Matthew John Rioth

Thesis
Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of

MASTER OF SCIENCE
in

Biomedical Informatics
May, 2016
Nashville, Tennessee

Approved:

Carlos Arteaga, MD

Paul Harris, PhD

Mia Levy, MD, PhD

Jeremy Warner, MD, MS

ACKNOWLEDGEMENTS

I would like to acknowledge everyone who has contributed to this work. Foremost are the members of my thesis committee Drs. Arteaga, Harris and Levy and Dr. Warner the chair of the committee and my mentor. David Staggs in Vanderbilt HealthIT was instrumental in implementing the clinical systems changes described here and troubleshooting the problems that inevitably arose. Lauren Hackett worked tirelessly to achieve the changes to the Laboratory Services Agreement that made the data transfer possible. Megan Cook helped define the ordering workflow. Mike Tod and Erich Habberman at Foundation Medicine set up the data transmission. Drs. Ramya Thota and Doug Johnson helped validate the parsing systems. This project began under the direction of Dr. William Pao. The support of the Division of Hematology and Oncology as well as the Department of Biomedical Informatics has been essential to the completion of this work and to my training in biomedical informatics and oncology. This work was supported by grants from the National Library of Medicine (T15 LM 7450-12) and the Vanderbilt Institute for Clinical and Translational Research (UL1 TR 000445). This work would not be possible without the support of my family: my wife Meghan, daughters Liana and Hattie, and son Niels who were patient with me and always supportive.

TABLE OF CONTENTS

Appendix

LIST OF ABBREVIATIONS

ASCO	American Society of Clinical Oncology
BAM	Binary Alignment Map
CDA	Clinical Document Architecture
CDS	Clinical Decision Support
CPOE	Computerized Physician Order Entry
CUID	Concept Unique Identifier
DNA	Deoxyribonucleic Acid
EHR	Electronic health record
eMERGE	electronic MEdical Records and GEnomics
FGF	Fibroblast Growth Factor
FHIR®	Fast Healthcare Interoperability Resources
FISH	Fluorescent *in situ* Hybridization
FMI	Foundation Medicine, Incorporated
Gb	Gigabit
HGNC	HUGO Gene Nomenclature Committee
HGVS	Human Genome Variant Society
HIPAA	Health Insurance Portability and Accountability Act
HL7®	Health Level Seven International
HTTPS	Hypertext Transfer Protocol Secure
HUGO	Human Genome Organization
ICD	International Classification of Diseases
IP Address	Internet Protocol Address
IRB	Institutional Review Board
Kb	Kilobit
MESH	Medical Subject Headings
MRN	Medical Record Number
NCBI	National Center for Biotechnology Information
NCI-MATCH	National Cancer Institute Molecular Analysis for Therapy Choice
NCIt	National Cancer Institute Thesaurus
NGS	Next Generation Sequencing
NLM	National Library of Medicine
PDF	Portable Document Format
PHI	Personal Health Information
SAM	Sequence Alignment Map
SFTP	Secure File Transfer Protocol
SMART®	Substitutable Medical Apps Reusable Technology
SNOMED-CT	Systemized Nomenclature of Medicine Clinical Terms
TCGA	The Cancer Genome Atlas
UMLS	Unified Medical Language System

VCF	Variant Call File
VICC	Vanderbilt-Ingram Cancer Center
VUMC	Vanderbilt University Medical Center
VUS	Variant of Unknown Significance
XML	Extensible Markup Language

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

Cancer is fundamentally a disease of disorder in the genetic code[1]. This has been appreciated since the 1970's; however, it was not until 2001, with the introduction of imatinib[2], that knowing the genetic characteristics of a cancer would inform its treatment. Since that time the knowledge of cancer genomics and the armamentarium of drugs that target specific mutations has grown significantly[3]. The use of molecular diagnostics such as DNA sequencing to inform the diagnosis, prognosis, and treatment of cancer is becoming ever more essential to the practice of oncology[4]. Mutations or other genetic derangements can indicate a prognosis, such as mutations in the promoter of the TERT gene convey different survival times in glioblastomas[5]. In other situations, the diagnosis can be aided by molecular profiling[6]. However, the most well-known uses of molecular profiling is in the identification of genetic variants that convey sensitivity to drugs specific to those mutations. Examples of these abound in the literature, and the use of genetic information in this way is commonly referred to as "precision oncology", and increasingly is applied in many clinical situations. Using molecular profiling to identify targetable mutations in many cancers is now the standard of care[7].

Concurrently with the rise of molecular profiling in oncology, the technology to obtain ever increasing quantities of genetic information has increased in scope[8]. Massively parallel next-generation sequencing (NGS) is an accurate and sensitive means to assay many genes for potential clinically relevant mutations in a high throughput manner[9–11]. Thus as more patients undergo molecular profiling of increasing amounts of genetic data, the challenge of integrating these molecular profile data into clinical workflows is considerable. Additionally, obtaining these substantial amounts of genetic data presents an opportunity to utilize them to further the knowledge of cancer biology. As such, the reporting of molecular profile information is important and is gaining increasing attention[12–15]. This thesis posits that **a framework for the data model and functional requirements for the electronic transmission of tumor molecular profile reports can inform the implementation in clinical systems and for secondary uses**.

Following this introduction, chapter two outlines a framework for the data elements, file types, transmission and information technology systems required for reporting. Data types and report formats are placed in the context of the Ackoff knowledge framework[16]. Additionally, a review of several molecular profiling laboratories in the context of this framework is reported.

Chapter three describes the implementation and improvement of an automated system to report the results of molecular profiling from a third party lab—a laboratory that is not integrated with the ordering institution--into an electronic health record (EHR). This chapter was published in the Journal of Oncology Practice as a "Quality in Action" article[17]. The text of this manuscript has been augmented with additional commentary on the security of transmissions and the later development of parsing variants of unknown significance (VUS) from portable document format (PDF) files for display within the enterprise EHR.

The fourth chapter describes the process to utilize the molecular profile data received and aggregate it into a clinical-genomic database. This database has been used to address multiple secondary use cases and has become a valuable tool for multiple aspects of cancer discovery, clinical trials and operations.

The systems described satisfy some aspects of the framework described in chapter one, but others are left unaddressed for future directions or different systems. The overall experience of this work is to demonstrate that molecular profile data is important and if structured and parsed carefully can inform not only the treatment of cancer patients, but also the science of oncology.

CHAPTER 2

Defining a framework for tumor molecular profile reporting

The clinical use of tumor genetic testing, also known as molecular profiling, is increasing in the routine practice of oncology. In a recent survey of practicing oncologists conducted by the American Society of Clinical Oncology (ASCO), only 6.8% of respondents reported rarely or never ordering cancer gene panels (private communication). The use of high throughput next generation sequencing (NGS) of tumors' DNA across panels of genes discovers many genomic variants that may be of clinical importance[2,18]. This sequencing is carried out in specialized laboratories with expertise at extracting DNA from cancer cells, sequencing the DNA, and reporting the variants back to the ordering provider[9]. The reporting of molecular profiling results is unique in several aspects: the use of high-throughput sequencing generates large volumes of data, much of this sequence data is not informative, germline variants with potential hereditary implications may be captured, and interpretation of what is and is not important in the context of a cancer patient is constantly and rapidly evolving[13,19]. As such, the accurate, secure, and concise reporting of molecular profiling testing is a non-trivial topic. Additionally, the genetic information contained within molecular profiles should be interoperable such that they can be utilized for inter-institutional collaboration and other secondary uses. In this section, current approaches and standards are examined and a framework for the reporting of tumor molecular profiles is proposed.

Realizing the full clinical and secondary use of tumor molecular profiling touches on several domains of biomedical informatics to create a comprehensive framework for molecular profiling. Foremost, workflow analysis of ordering and receiving patterns must to be conducted to inform clinical informatics changes. The molecular profile report must be integrated into the EHR, and best practices dictate that the ordering provider should be notified of its return via a notification mechanism[20]. Consistent data standards, not only for the ascertainment of genetic variants in the sequencing pipeline, but also for the reporting and transmission of the molecular profile results, are needed[21,22]. Full interoperability of molecular profile reporting will require definition of genetic reporting standards in commonly-

implemented standards such as Health Level Seven International (HL7)'s emerging Fast Healthcare Interoperability Resource (FHIR) standard. Additionally, using structured molecular data fields for clinical trials in clinical trial management systems would allow computerized rule-based screening for patients eligible for targeted therapy-based interventional clinical trials[23].

Utilizing structured genetic information as well as clinical metadata, clinico-genomic data warehouses can be built from aggregated molecular profile reports for secondary uses similar to other clinical research informatics projects that create linkages from germline DNA and clinical data, such as Vanderbilt University's BioVU[24–27]. However, the full utilization of the genetic data from molecular profile reports requires both the ability to deep-phenotype the patients to understand the clinical context, as well as use of sophisticated bioinformatics tools designed for research. The ability to phenotype patients relies on fields of natural language processing, clinical domain knowledge, and automated cohort discovery[28,29]. The ability to use bioinformatics tools relies on the ability to map variants to unambiguous genetic coordinates within a known reference genome and within a known scope of the test. Additionally, translational science using a clinico-genomic data warehouse is aided by sophisticated statistical and machine learning techniques to test hypotheses and analyze the data[30–32]. Without resources from multiple domains of biomedical informatics to create an integrated framework, the full utility of tumor molecular profiling cannot be realized.

As depicted in Figure 2.1, the clinical and secondary use of molecular profiling requires multiple steps. Molecular profiling is initiated according to the clinical situation, and ordered most often by the treating oncologist. There are several ordering options as depicted in Figure 2.1(B): computerized physician order entry (CPOE), via a web portal hosted by the sequencing laboratory, or on paper which is then transmitted to the sequencing lab. Importantly, clinical information such as diagnosis must accompany the order to influence annotation and billing codes often are included to allow for reimbursement. The order is often the source of much of the sample, patient, and provider metadata that then flows through the rest of the framework. Following the order the sample must be obtained, and there are two common workflows for

this: if the order goes directly to the sequencing laboratory, it must contain tumor sample information such that the laboratory can contact the entity that holds the tumor sample in order to obtain it for sequencing; alternatively, the order can go directly to the entity that holds the sample—frequently a pathology department or group—which then sends the tumor sample along with the order to the sequencing lab. Once the sample is in the lab, sequencing is performed and genetic variants are ascertained; this is a subroutine that has been described elsewhere[18] and is not the focus of this framework. However, it should be noted that this process is influenced by the clinical information that was provided in the original order.

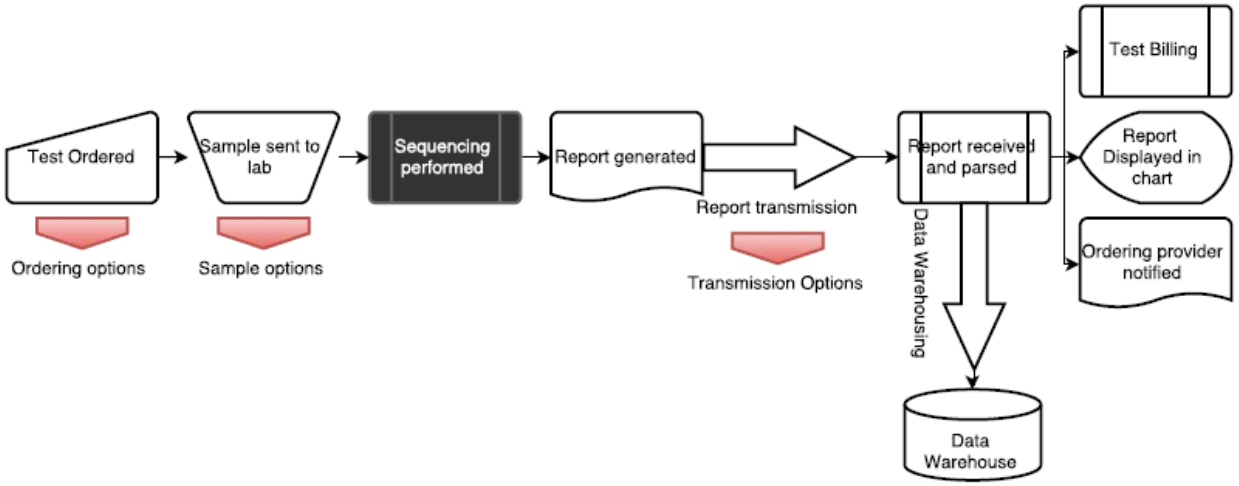
Following sequencing, a report must be generated and returned to the treating physician. Frequently this report contains annotation regarding therapeutic or other clinical information such as clinical trials; however, as is discussed later in this chapter, clinical annotation and variant information could exist as separate processes. As depicted in Figure 2.1(C), multiple options exist for returning the report to the provider. A paper report can be mailed or faxed to the ordering provider's institution, and then routed to the provider. Alternatively, a web portal provided by the sequencing laboratory can allow a provider to log in, view, and print results. An interface can also be created between the laboratory computer reporting infrastructures into the EHR used by the provider. Chapters 3 describes an implementation of such an interface. Regardless of the transmission means, the report must trigger a notification to be viewed by the provider and billing usually is initiated by the returned report.

It should be noted that the use of molecular profiling is not standard of care for many cancers. The use of panels of genes for sequencing, rather than sequencing mutational "hotspots" of proven clinical significance exceeds the standard of care for most tumor types. As such performing NGS on panels of genes, is in itself a research endeavor to discover genetic variants that may be biomarkers for clinical trials, novel therapeutics or to inform translational discovery efforts. Molecular profile reports when aggregated can create a population-level view of the genetics of cancer. This can be facilitated by the electronic transmission of the reports in a structured format, as these can then be parsed and automatically

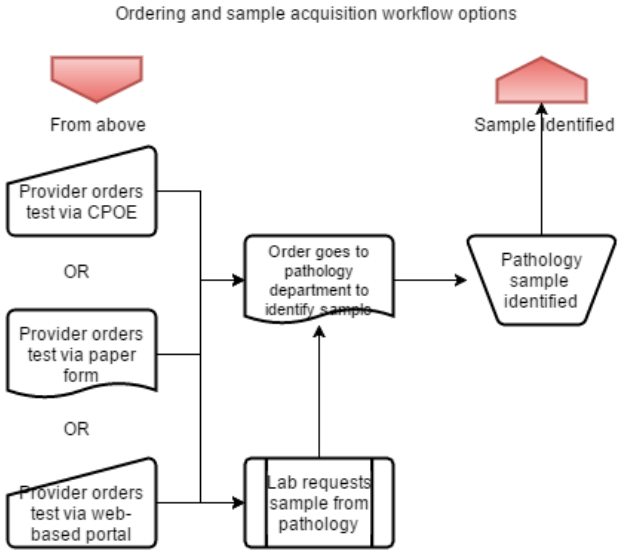
aggregated into a data warehouse for multiple secondary uses. As depicted in Chapter 4, one method of this report parsing and the creation of a population level database is described as well as the secondary uses it satisfies.

It should be noted that Figure 2.1 is a generalization of the framework for molecular profile reporting and ordering, and individual providers, practices, and institutions may have variations of these steps. However, as depicted in the rest of this chapter, there are many essential minimum required elements for ordering and reporting as well as many common considerations for the file types and transmission standards across clinical situations. These commonalities and differences are highlighted in a survey of the practices of molecular profiling laboratories in table 2.8. Standards organizations such as the Global Alliance for Genomics in Health[33] and HL7[34] are actively investigating the best practices for data formats and transmission types of molecular profile reporting, and as the clinical utility of molecular profiling continues to increase, there will likely be increased attention to this important aspect of cancer care.

A)


Test Ordered
Sample sent to lab
Sequencing performed
Report generated
Report transmission
Report received and parsed
Test Billing
Report Displayed in chart
Ordering provider notified
Ordering options
Sample options
Transmission Options
Data Warehousing
Data Warehouse

B)


Ordering and sample acquisition workflow options
From above
Sample identified
Provider orders test via CPOE
OR
Provider orders test via paper form
OR
Provider orders test via web-based portal
Order goes to pathology department to identify sample
Pathology sample identified
Lab requests sample from pathology

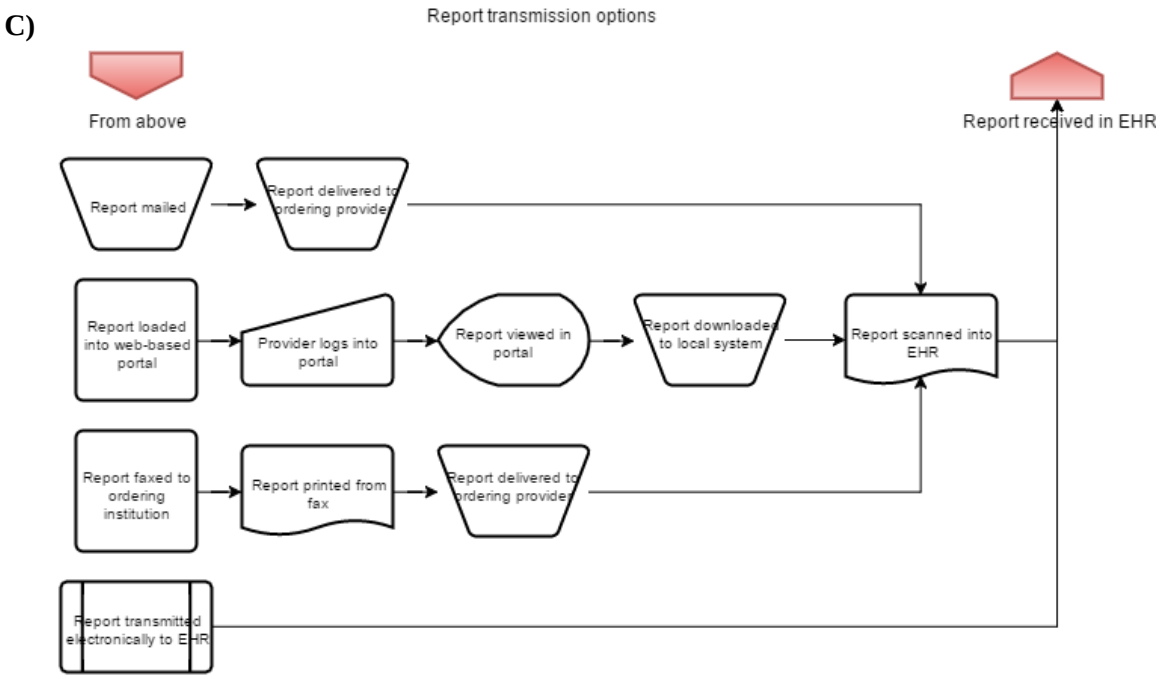


Figure 2.1 An information flow diagram for molecular profile ordering and reporting. In order to inform the overview of this work, a diagram of the flow of information from the ordering clinician through the sample acquisition, through sequencing and reporting and back to the ordering physician is reported. Importantly, the steps required to perform the sequencing (black rectangle) are out of the scope of this work, and have been well-summarized elsewhere *[18]*. Steps where multiple options for completing the task such as ordering, sample acquisition (B) and report transmission (C) are depicted by red arrows with separate subroutines. When molecular profile reports arrive back at the ordering institution they must be displayed in the chart, the provider must be notified, and billing must be initiated. Additionally, these reports can be directed to a data warehouse for secondary uses such as research.

Minimum data elements

Ordering of the molecular profile test

The initial step in molecular profile reporting is the order that initiates the testing. This order usually comes from the provider treating the patient in the course of clinical care, although some genetic testing is performed reflexively upon sample acquisition[35]. The order begins the clinical data capture and transfer that informs the annotation of the molecular profiling results as well as the secondary use utility subsequently. As depicted in figure 2.1(B), multiple options exist for ordering: a computerized physician

order (CPOE) system, a web-based portal hosted by the laboratory, or a paper order. The advantage of a CPOE system is that it keeps the order within the workflow of the physician and allows for clinical decision support (CDS) for test ordering. However, doing this requires the ordering to be built into the CPOE system, and if the order will be transmitted directly to the lab, then an electronic interface must be built which can be costly. Another more common option is the use of a web-based portal for a provider to log-in to, and complete an electronic ordering form which is sent directly to the lab. This has the advantage of not requiring an interface to be built between ordering systems and the laboratory. However, this is outside the typical workflow of the ordering provider, does not supply CDS, and requires manual typing of data from one system into another. Some ordering portals do little more than allow providers to print forms with the provider's ordering information already pre-populated. The most common means of ordering tests is filling out a paper form and faxing it. This has the advantage of not requiring a login and is within the workflow of most clinicians. However, it cannot supply CDS, requires the manual copying of information onto the ordering sheet, and requires information to be legible.

Following the placement of the order for molecular profiling, there are two potential options to initiate the execution of the order. If the order goes directly to the molecular profiling laboratory, it typically contacts the pathology department where the desired sample is located to obtain tissue for testing. However, in some workflows, the order goes first to the pathology department where the specimen is identified and sent along with the ordering requisition. If the laboratory, rather than the institution will be billing for the test, accompanying the order must be insurance information as well as billing codes for diagnosis. Traditionally these codes have been international classification of disease (ICD) codes; however, the clinical precision of ICD codes can be low requiring other data (such as stage or specific histology) that indicate to insurance companies the appropriateness of the test. For instance, within the ICD system there is no difference between adenocarcinoma and squamous cell carcinoma of the lung; however, these are separate clinical entities that have different prognosis, diagnostics, molecular landscapes, and

treatments[36,37]. As such ICD codes are minimally sufficient for determining clinical context for molecular profile reporting.

However, a domain-specific extension of the ICD codes, the International Classification of Diseases for Oncology (ICD-O), which is utilized by cancer registries, would be an alternative option[38]. The ICD-O version 3 is a postcoordinated terminology with multiple axes to designate site, morphology, behavior, and grading of neoplasms. The topography axis uses the ICD-10 classification of malignant neoplasms (except those categories which relate to secondary neoplasms and to specified morphological types of tumors) for all types of tumors, thereby providing greater site detail for non-malignant tumors than is provided in ICD-10[39]. The ICD-O classification system is widely used by tumor registries, and would be a reasonable choice for a codified diagnosis data element. A downside to this terminology is that it was last updated in 2000.

There exist several options for terminologies and a few ontologies for describing a clinical diagnosis, each with their own strengths and weaknesses[40]. The most widely used terminology is the Systemized Nomenclature of Medicine Clinical Terms (SNOMED-CT). The SNOMED-CT is an eleven axial heirarchical terminology that is maintained by the International Health Terminology Standards Development Organization and is indexed within the National Library of Medicine's (NLM) Unified Medical Language System (UMLS) metathesaurus[41]. It is a postcoordinated terminology which has the advantage of flexibility in terminology, but this conveys the weakness of myriad potential means of depicting the same clinical entity. As such many implementations impose a rule-based system on the terminologies available to use, thus implementing it in a precoordinated fashion[42,43]. This weakness is common to all postcoordinated terminologies. An example of a precoordinated terminology that attempts to cover all disease states is the medical subject heading (MeSH) terminology controlled by the NLM for the purpose of indexing medical articles for searching online[44]. Being precoordinated, MeSH has the advantage of constrained and standardized options for describing a disease; however, the weakness is that a user is limited to the options available within the terminology which may be insufficient in the realm of

oncology. Another potential precoordinated terminology option would be the terms included in the National Cancer Institute thesaurus (NCIt) which contains better precision than MeSH for clinically-relevant cancer terms[45,46]. Additionally, it is versioned, updated three times yearly, and indexed with UMLS metathesaurus. It is not, however, in widespread use outside of the Food & Drug Administration and a limited number of research-oriented standards development organizations (e.g., CDISC). In a practical implementation, one could map the terms from NCIt onto SNOMED-CT terms and use SNOMED-CT; however, in the context of this framework, NCIt would be a better choice.

In addition to the information regarding the patient and diagnosis, information about the specimen and provider are necessary. To ensure the correct specimen is tested, identifiers such as sample date, anatomic location, and pathology ascension number should be included to avoid ambiguity. The ordering provider and their institution is necessary to ensure the report is returned back to the correct provider. Institutions could be unambiguously identified by their institutional license number, while providers could be unambiguously identified by their national provider identification number.

Table 2.1: Minimum Data Elements for Tumor Molecular Profile Ordering

Element	Format	Rationale
Patient Identifiers	Name, medical record number, date of birth	Report must be able to be linked to patient
Payer information	Payer information, policy number	Required if laboratory will be billing for test
Provider Information	Name, National Provider Index ID (better)	Report must be routed back to ordering provider
Institution Information	Text, Institutional License Number (better)	Reports must be routed back to ordering provider at the correct institution
Diagnosis code	International Classification of Disease (ICD) 10, Clinical Modification billing code for test, National Cancer Institute Thesaurus (NCIt) concept unique identifier (CUID) (better)	ICD-10-CM required for billing and reimbursement. Although may be insufficient for testing in certain clinical settings.
Date of Order	Standardized date format	Required for billing and tracking
Sample identifiers	Date of acquisition, sample accession number	Ensures correct clinical sample is tested

Reporting of test results

With regards to reporting of molecular profile data, the first designation should be the minimum required data elements that constitute a report. Many aspects of these have been enumerated elsewhere[12,21,47], but in general they fall into three categories: patient metadata, sample metadata, and test result. Patient metadata incudes patient name, sex, date of birth, diagnosis or condition, and identifier number. Sample metadata includes type of sample, identifier number, and date of collection. Finally, the test results will vary based on the test performed. Multiple types of molecular data are relevant to cancer testing and each has a different means of reporting. Mutations, copy number variants, gene fusions, insertions, deletions,

splice site alterations, and other genetic variants can be detected using next-generation sequencing and included in a molecular profiling report.

Recently a joint effort by the College of Medical Genetics and Genomics and the Association for Molecular Pathology[21] have specified a joint consensus recommendation for the standards and guidelines for the interpretation of sequence variants. The European Society of Human Genetics has also published less-extensive guidelines for the same topic[12]. These publications outline many practical aspects of reporting molecular data with a focus on germline genetic testing. Additionally, they discuss the role of annotation in the report. However, both groups include more data elements (such as database resources, *in silico* prediction tools, and methodical considerations) that are beyond the scope of the actual report. Defined in Table 2.2 are the minimum essential data elements for molecular profile reports.

The description of variant annotation is described in a separate section. There is no current consensus on how to annotate variants nor a source of truth to ensure annotation accuracy. The practice of precision oncology is nascent enough that there exist a multitude of non-standardized approaches. However, regardless of the content of the annotation, the above data elements are sufficient to satisfy clinical uses and secondary uses.

Table 2.2: Minimum Data Elements for Tumor Molecular Profiling Reports

Element	Format	Rationale
Patient Identifiers	Name, medical record number, date of birth	Report must be able to be linked to patient
Provider Information	Name	Report must be routed back to ordering provider
Ordering Information	Date of order, date of report, date of amendments (if any)	Useful for report tracking
Diagnosis	ICD10 code, Structured diagnosis text according to NCIt (better)	Diagnosis influences annotation
Sample Metadata	Structured text of anatomic location, date of acquisition, sample accession number	Information to identify the sample under study
Scope of Test	List of HGNC-compliant gene names and reference sequences assayed in test	Including scope of test allows discernment of negative results rather than results that could not be determined
Test Result		
Unambiguous DNA Change	HGVS syntax[48,49]	Allows interoperability to other bioinformatics tools
Reference sequence	RefSeq[50,51]	
Predicted protein change	HGVS syntax	Clinically-embraced syntax for reporting
Allele fraction	Numeric	Potentially relevant clinically
Technical Failure	Boolean	Report if a test failed and therefore should not be interpreted
Signing Authority	Name, credentials	Pathologist who signs out and approves the report
Laboratory Information	Name of laboratory director, CLIA number (if applicable), credentialing certification	Laboratory metadata according to regulatory requirements
Annotation	Multiple options	See section “Annotation” below

File Type

One of the initial specifications that must be made in the reporting of molecular profiling reports is the file format for the report. The options (Table 2.3) span a range of categories, data elements and sizes[22]. These options can be thought of as existing along the continuum of data, information, knowledge and wisdom as articulated by Ackoff[16] (Figure 2.1). File formats such as sequence alignment map (SAM) or binary alignment map (BAM) files that contain the raw sequence data or sequence data aligned to the reference genome could be considered data[52]. These formats are often the largest in size up to several gigabytes for whole genome sequences, contain the least amount of clinical information, and are not commonly used in clinical practice.

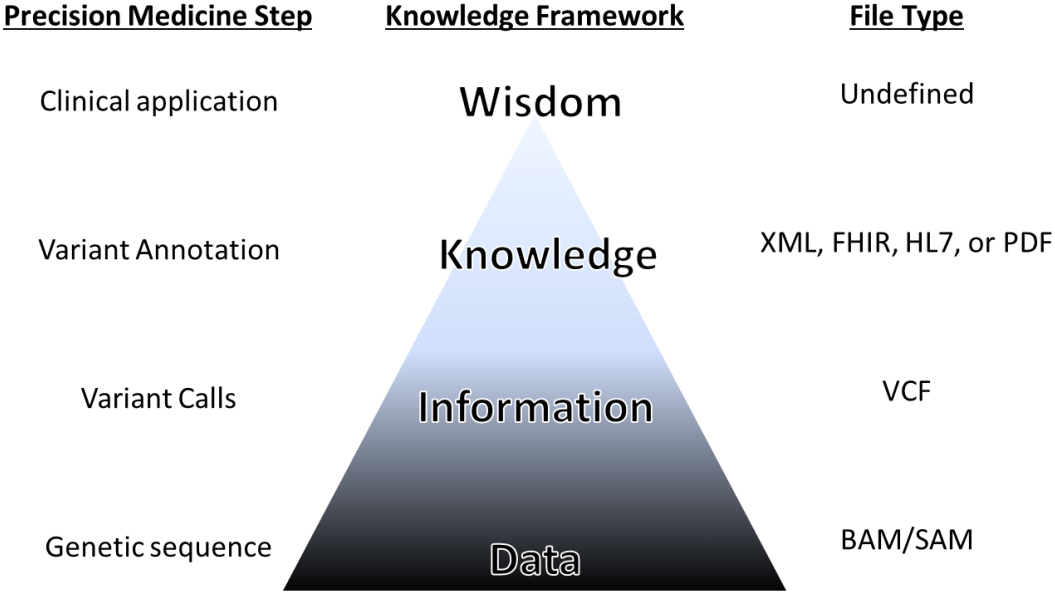


Figure 2.2 Knowledge framework for molecular profiling. The steps for precision medicine as well as the file formats can be viewed within the context of a knowledge framework postulated by Ackhoff. The Data layer is represented by the genomic sequence and contained within alignment mapping files. Variants are information and contained within a file such as a VCF. Annotation of the variants represents knowledge and is the most heterogeneous of potential file formats. Finally wisdom is the clinical application of the variants in the context of the individual patient and takes into consideration the entire clinical context including performance status, comorbidities, and patient preference, among others. This clinical context is not captured in the current framework of molecular profiling and as such the proposed file for this step remains undefined.

When sequence data is annotated with calls of genomic variants such as mutations, this constitutes information. Data files that correspond to this type of information would be variant call files (VCF) for

mutations or mean allele frequency files for copy number variants. These file formats are smaller than BAM or SAM files, often in the order of kilobytes to megabytes. They constitute the changes observed in a sample relative to a reference genome, and in combination with a reference genome the sample genome can be reverse engineered. These file formats are commonly used in bioinformatics and research endeavors, but they lack the ability to convey clinical annotations that would be necessary for clinical reporting. One attempt at a data standard, a "genomic" VCF which incorporates additional clinical annotation into the traditional VCF format, has been created but has failed to attain widespread use[53,54].

In this context, knowledge could be described as annotation of genetic variants with regards to their potential clinical significance. Variants could be annotated as prognostic, diagnostic, or therapeutic in the sense that they inform treatment. This is the level that most molecular profiling reports exist. Variants are annotated against a previously curated knowledgebase of clinical significance and this knowledge is reported to the ordering clinician[31]. Multiple possible file or messaging formats exist for this task: extensible document language (XML), HL7 V2, HL7 Clinical Document Architecture (CDA), and FHIR[34], or a paper-based format such as portable document format (PDF).

Wisdom is an emergent phenomenon that is dependent on clinical context and patient preferences, such that conveying it is beyond the scope of a molecular profiling report. However, in the context of CDS the molecular profile report could be included to help guide a wise decision. An ideal CDS system will be able to take the molecular profile report as well as additional clinical and contextual information and help guide a clinician to make the best treatment decisions for an individual patient.

Custom XML and the HL7 formats, especially FHIR, are best suited to molecular profile reporting because they can contain the data elements from Table 2.2 in addition to satisfying file transmission requirements.

Table 2.3: Options for molecular reporting file types

File Type	Advantages	Disadvantages
SAM, BAM[52]	Unambiguous, complete data	Largest file (~Gb), no annotation, not human readable
VCF[52]	Reasonable file size (~Kb), unambiguous variant representation, allows some annotation	Not human readable
Generic XML	Highly customizable, interoperable, structured	Not standardized, not human readable
HL7 v2.5.1[55]	Interoperable, structured, standardized, allows bidirectional interfaces	Less customizable, not human readable
HL7 CDA [34]	Structured, standardized, human readable	Not as interoperable as FHIR, large files with many data elements
FHIR[56]	Interoperable, structured, standardized, allows bidirectional interfaces, human readable	Less customizable, not yet widely implemented
Paper, PDF	Human readable, familiar format, interoperable	Not standardized, slow paper transmission, cannot be parsed for data elements

Abbreviations: BAM – Binary Alignment Map, SAM – Sequence Alignment Map, VCF – Variant Call File, XML – Extensible Markup Language, HL7 – Health Level 7, CDA – Clinical Document Architecture, FHIR – Fast Healthcare Interoperability Resources, PDF – Portable Document File

File transmission and security

Many clinical labs that perform molecular profiling are external to the information systems of the treating physician. Thus transmission of the report is necessary to inform treatment. As depicted in Figure 2.1(C), the transmission of reports can be faxing, as has been the standard for decades[57], the use of a web-based portal, electronic delivery through the internet, or physical paper report delivery though the mail. The transmission of molecular profiling reports falls under the jurisdiction of the Security Rule of the Health

Insurance Portability and Accountability Act (HIPAA)[58], and as such one of the primary concerns regarding the transfer of molecular profile reports is security.

Faxing of reports has some safeguards in terms of policies dictating which fax machines can be used, and who has access to them. However, in many instances this process is not auditable and therefore actually not in compliance with HIPAA (§ 164.312(b)). Additionally, faxing can produce documents of poor readability and data integrity. As such, faxing is not the ideal means of reporting.

One alternative is to create a web-based portal for ordering providers to access, view, and download reports. This has the advantage of preserving access controls, data integrity, and some degree of transmission security which are required under the Technical Safeguards of HIPAA[58]. Additionally, web portals are dynamic and can display additional information that a static paper report cannot. However, web portals are at a disadvantage in that they require log-ins, generally do not support single sign-on authentication if controlled by a third party laboratory, and contain user interfaces that can disrupt clinical workflow and lead to less utilization. While OAuth[59,60] would alleviate some of the login burden on the providers, such systems are limited in their implementation in the healthcare area. Additionally, web-based portals are not part of the patient's permanent chart and require some form of downloading or printing followed by scanning to place the report in the patient's chart.

Electronic transmission has many advantages: it can be performed securely, with structured data, which can be audited, with access controls, and integrated into clinical work flows. Using custom XML or one of the HL7 data types, servers with the clinical sequencing laboratory can communicate with clinical servers to transmit structured reports. For this process HTTPS rather than SFTP is advantageous for several reasons: HTTPS, and specifically a web-service that relies on POST rather than GET transmission protocol, has the benefit of accepting streaming data or transmissions at any time. Additionally, report files do not rest on an FTP server requiring encryption at rest. FTP servers require user accounts and keys which can be susceptible to social engineering attacks. HTTPS has centralized verification certificates

rather than relying trusting. Additionally, HTTPS connections can be created such that only communication from a trusted IP address (i.e., the laboratory server) is allowed. Although HTTPS has the disadvantage that the communication port must always be available, this can be masked with an external firewall that prevents probing the communication port. As such the preferred means of automated transmission, and one that satisfies the Technical Safeguard provisions of HIPPA, is an automated transmission of a structured document based on HTTPS security.

Another potentially utilized security protocol is the DIRECT project[61]. DIRECT specifies several standards-based security protocols for transmitting authenticated, encrypted health information directly to known, trusted recipients over the Internet. It relies on a verified health information service provider to serve as an honest broker of verification. This system has been championed by the Office of the National Coordinator of Health Information Technology; however, its implementation has not been wide spread. In the future this could serve as the best option for transmitting molecular profile reports.

Table 2.4: Options for molecular reporting transmissions from the lab to the ordering clinician

Transmission Modality	Advantages	Disadvantages
Paper/Fax	Leverages existing infrastructure	Unstructured, low information fidelity, not auditable, requires multiple manual processing steps including patient matching and scanning
Web portal	Access control, email notification, dynamic, comprehensive, potentially downloadable	Requires workflow deviation, not directly integrated into patient chart, not interoperable
Web-based transmission	Allows potential bi-directional interfaces, Access control, digital and potentially structured data, auditable, automation, can integrate into other clinical systems	Greater security risks than HTTPS, no data streaming, susceptible to social engineering attacks
Secure File Transfer Protocol		
Secure Hypertext Transfer Protocol	Access control, high security, digital and potentially structured data, auditable, automation, data streaming, can integrate into other clinical systems, no data at rest	Reveals receiving server IP address
DIRECT	Highest security, regulated by Office of National Coordinator, developed specifically for healthcare application	Not widely implemented, guides for implementation have not been updated in over a year.[61]

Healthcare information technology infrastructure requirements

For a molecular profiling report to be used clinically, it must make it to the treating physician[62]. However, reports of molecular profiling are returned asynchronously with the clinical workflow; often

results are returned several weeks following the initial request. Reports must make it back to the ordering provider at a later time. To do so, the file must make it from the point of receipt to the point in the clinical workflow where the physician can view it. This can be as simple as a nurse retrieving a faxed report, or as complicated as an automated incorporation into an EHR that triggers other events such as clinical decision support, provider notification, pathological re-review, and/or patient billing. Several reviews have outlined functional requirements for laboratory reporting[63–65]. These reviews have outlined several aspects necessary for reporting that are not specific to molecular profiling: patient matching, incorporation of the report into the chart, and provider notification. Additionally, the return of a molecular profiling report could potentially provoke a re-review of the case by the pathologist at the host institution to determine if new testing is required, if the original histology needs to be re-reviewed, or if a new clinicopathologic correlation should be made.

Some have argued that the return of molecular information could provide information that changes the diagnosis, or other information that has already been reported about the pathological specimen. For this reason, some have advocated routing all returning molecular pathology reports through the laboratory information system for initial review by an on-site pathologist before reporting it to the ordering clinician. However, this policy would produce several potential drawbacks: it could increase the time before the treating clinician receives the molecular information[66–68]; it introduces additional information interfaces that constitute additional points of system failure[64,69]; and likely would not be reimbursable as the original molecular profiling report has already been signed by a pathologist[21]. An increasingly common compromise is to the have the report return to a “molecular tumor board” which meets regularly, is composed of clinicians, molecular biologists, and pathologists, and provides expert interpretations for the clinical end user(s).

Table 2.5: Elements of information technology required for reporting

Information Technology Element	Rationale
Verify patient information	Returning reports must be matched to the correct patient
Accurately incorporate report into chart	Reports must be incorporated into the chart with descriptive titles and metadata to allow them to be found. Additionally, data elements within the report should be searchable from a patient's chart.
Notify ordering provider of receipt	Ordering providers should be notified when the results have been returned

Annotation

In order to translate information to knowledge, the variants that are observed in the molecular profile must be annotated regarding their clinical significance. While some variants have well-described clinical diagnostic, prognostic or therapeutic implications, many lack such evidence. The terminology surrounding variant classification into tiers of significance is beginning to develop[21]. However, the knowledge base of what constitutes clinical significance, including clinical trials and preclinical data, is constantly expanding. Thus, clinical laboratories that perform molecular profiling are also *de facto* knowledge management organizations of variant annotation information. Because of these dual demands, some have advocated the separation of primary molecular measurements from clinical interpretation[62]. In response to the demands of molecular data knowledge management, some companies have been founded to manage this aspect of molecular profiling, thus relieving the clinical laboratory of this task[70]. Broadly, two options (Table 2.6) exist for annotation of variants: annotation within the same report as the variant calls or annotation from a separate knowledge base external to the variant calls.

However, the division of variant calling from annotation creates an additional interface that could be the source of miscommunication or data loss. To prevent this, transmission of standardized, structured, and

interoperable variant information will be required to allow lossless creation of reports. This could be achieved by utilizing one of the data standards such as VCF or HL7 FHIR as outlined in table 2.3.

Table 2.6: Approaches to variant annotation

Modality	Advantages	Disadvantages
In the same report as the variants	Does not require additional steps, direct linkage of variant to annotation	Requires clinical laboratory to maintain real-time knowledge of annotations
Queried from a knowledgebase external to the report	Dedicated annotation service	Requires linking variants to annotation, creates additional interface

Secondary use of molecular profiling data

As described by Masys et al, a goal of including molecular profiling data into the EHR is to satisfy potential secondary use cases[62]. At Vanderbilt University, utilizing EHR data for secondary uses has been successfully implemented for many years[24,25]. There are several specific use cases that should be supported by molecular profile reports. Molecular data can be used as inclusion criteria for clinical trials; however, these type of biomarker driven clinical trials are different from the traditional histology based clinical trials with which oncologists are most familiar[23]. As such, identification of patients by molecular inclusion criteria across cancer populations will aide in the conduct of these novel trials. This will require that variant information in molecular profiling reports can be searched across populations.

Additionally, collaborations across institutions such as the electronic Medical Records and Genomics (eMERGE) network, will require molecular data that can be normalized so that cohorts can be combined[71]. This will require standardized terminologies, data representations, and metadata. The data elements such as unambiguous description of DNA changes with standardized terminologies coupled with publically available reference sequences (Table 2.2) will allow data normalization and accurate merging of cohort populations.

Scientific discovery and cancer biology research should also be facilitated by molecular profile reporting. The genetic data contained within each molecular profile report constitutes potentially novel evidence of cancer biology. Many tools exist for interrogating the genetics of cancer, but they are frequently built for data types other than clinical reports. However, clinical reports should support the ability to accurately reverse-engineer the data types that are used as substrates for bioinformatic investigation. This requires unambiguous and standardized terminology for DNA changes, as well as indication of reference sequences. Additionally, including data elements such as allele frequency (Table 2.2) could lead to new evidence regarding the clinical significance of tumor heterogeneity.

Table 2.7: Secondary use cases for molecular profiling data

Use case	Description
Clinical trial matching	An increasing number of clinical trials are requiring molecular biomarkers as inclusion criteria. Molecular profiling reports should be able to be utilized for identifying patients eligible for clinical trials. The application of this is described in Chapter 4.
Clinical trial planning	Biomarker data can be used to determine the feasibility and projection of accrual for clinical trials with biomarker-based inclusion criteria.
Quality improvement	Aggregating data for assessment of utility and cost effectiveness is important for payers and administrators.
Operational needs	Monitoring the utilization and results of molecular profiling for an organization can be necessary for contractual reasons, for grant applications, and for financial analysis.
Inter-institutional collaboration	Collaboration across institutions will require the normalization of molecular data. The variant information contained in molecular profiling reports should be interoperable across institutions.
Data transactions	Molecular data from clinical specimens can be valuable to pharmaceutical, diagnostic, and/or biotechnology companies. These data may have monetary value or non-monetary value, such as early access to promising investigational agents.
Scientific discovery	The variant information contained in molecular profiling reports should be interoperable with tools used for bioinformatic discovery efforts and map to reference data sets.

Current Overview of Molecular Profile Providers

Characteristics of clinical tumor molecular profile ordering at multiple laboratories were ascertained by review of cancer or somatic tests listed in the National Centers for Biotechnology Information's (NCBI) Genetic Testing Registry[72]. Table 2.8 is not a complete list but represents a sample of laboratories and testing modalities. The potential for secondary use is categorized by the presence or absence of structured electronic data transmission as advertised on the NCBI website of individual laboratory websites as referenced in the table.

Table 2.8: Framework elements for providers of clinical tumor molecular profiling

Provider	Test Description	Ordering	Reporting	Potential for secondary use
Foundation Medicine, Inc.[73]	Targeted exome sequencing	Paper	Paper, portal, or XML interface	With structured XML reporting
Caris Life Sciences[74,75]	Targeted exome sequencing, immunohistochemistry, fluorescesent *in-situ* hybridization (FISH), among others	Paper, portal or interface	Paper, portal, or EHR interface via Quest platform	Possible with EHR interface
PathGroup[76]	Targeted exome, cytogenomic array	Portal, paper	Paper, portal or EHR interface with HL7 or PDF support	Possible with EHR interface
Guardant Health, Inc.[77]	Targeted exome from circulating cell-free DNA	Paper	Paper, portal	Requires manual curation
Genomic Health, Inc.[78]	Targeted RNA panel sequencing	Paper, portal	Paper, portal	Requires manual curation
University of Washington[79]	Targeted exome panel	Paper	Paper	Requires manual curation
BioReference Laboratories[80]	Targeted exome sequencing	Paper, EHR interface	Paper, EHR interface	Possible with EHR interface

Washington University[81]	Targeted exome sequencing	Paper	Paper	Requires manual curation

Many laboratories, both independent and based at academic centers, offer tumor molecular testing[72]. Most of these tests consist of targeted panels of exomes sequenced by NGS; however, notable exceptions are Caris biosciences which performs multiple other types of tests and integrates these results into their reports, and Genomic Health which utilizes RNA-based expression analysis for its predictive test. By reviewing sample reports, all tests appear to utilize HGNC-compliant gene names and HGVS-compliant protein syntax for reporting variants. Only Guardant360, the only cell-free DNA test commercially available, reports the allelic fraction it detects. All tests accept paper ordering and reporting, but many utilize a physician portal for reporting. The larger commercial labs have the advertised ability for EHR interfaces, but not the university laboratories. Caris Life Sciences advertises its EHR interface for ordering as well.

Conclusions

As depicted in Figure 2.1, the reporting of molecular profile data is complex due to the volume of potential data, the multiple informatics systems involved, the numerous participants in the system, and the increasing amount of knowledge required for annotation. However, a comprehensive framework for molecular profile reporting can support the care of cancer patients and the science of oncology. This requires defining data elements contained within the report, utilizing a standard information exchange format such as custom XML or one of the HL7 standards that supports loss-less data transmission, a secure and efficient transmission method that complies with federal regulations, and systems within the medical information technology infrastructure that support clinical care. The question of variant annotation is evolving, but separating the laboratory observation from the annotation knowledgebase will allow each task to be addressed with focused expertise. With the evolving state of molecular biomarker

annotation, the conversion of genetic data into clinical wisdom will require multiple layers of expert informatics systems.

CHAPTER 3

Implementing and improving automated electronic tumor molecular profile reporting

Introduction

Molecular profiling of tumors is becoming the standard of care in an increasing number of cancer types to inform the practice of precision medicine[18,82]. However, the implementation of somatic gene testing into clinical practice can be difficult as such testing can be reported asynchronously and separately from other pathology information[83–85]. Additionally, tumors are often tested at laboratory facilities that are not integrated with oncologists' offices and their electronic health records (EHRs). This disconnect can create challenges in reporting important genetic information from the laboratory to the treating oncologist[15].

Oncologists at Vanderbilt University Medical Center (VUMC) have partnered with a provider of tumor exome sequencing, Foundation Medicine Inc. (FMI, Cambridge, MA), to analyze patients' tumors for genetic variants that may inform clinical decision making[9]. This collaboration, like most with 3rd-party labs, has relied on the use of faxed documents to report test results. While faxing has for decades been the standard of 3rd-party lab reporting, the use of black and white faxes with subsequent paper copies that are scanned into an EHR leads to a process that requires many manual steps and results in a loss of color information and poor readability in the EHR[57] (Figure 3.1). This lack of information interoperability has been cited as a target to improve the practice of precision medicine[86]. At VUMC, an assessment of faxed molecular profile reporting demonstrated poor information fidelity and lack of provider notification; these issues prevented optimal utility of molecular profiling. We hypothesized that automated electronic reporting from FMI directly into the VUMC EHR could be feasibly implemented and enable provider notification. We describe the design, evaluation, and implementation over one year of this system.

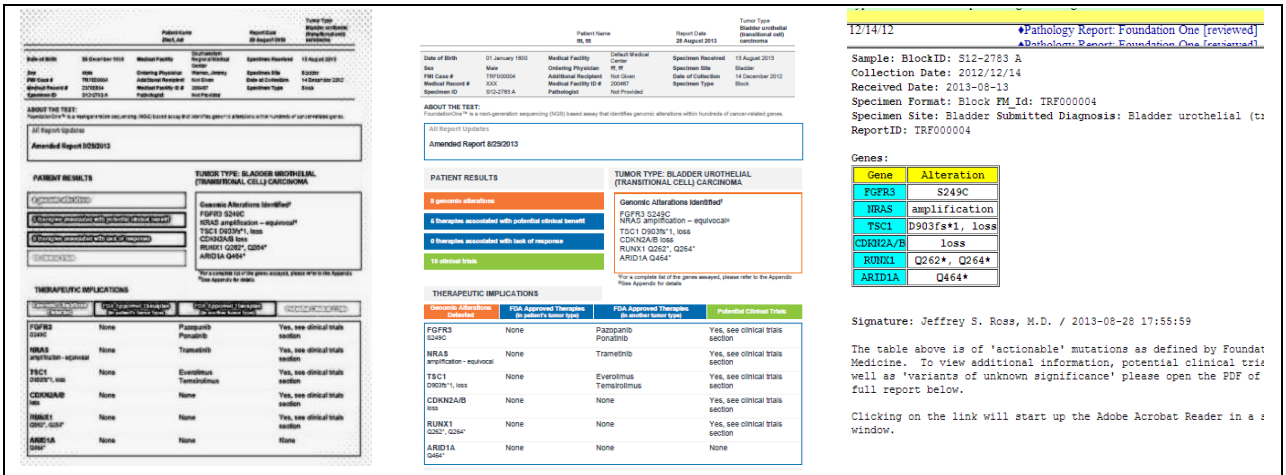


Figure 3.1 Report information fidelity is improved with automated transmission. Faxing and scanning of tumor molecular profiling reports (left) results loss of color information and reduced readability compared to the original color PDF report (center). Additionally, parsed XML information can be displayed directly in the EHR (right) thus not impeding the practitioner's workflow when finding results of genetic testing. Data shown is dummy data created for demonstration purposes; dates and accession numbers do not correspond to any patient personal health information.

Methods

The initial evaluation of somatic gene testing at VUMC was to perform a lean-based assessment with a modified value-stream mapping of test ordering and reporting workflow[87]. To address the shortcomings identified in this assessment, we defined a document structure and a data transfer protocol for electronic lab reporting. Through an iterative design and testing process utilizing lean methodology, a post-processed laboratory report was packaged into a standardized extensible markup language (XML) document that included demographics, ordering information, and the FMI-designated actionable variants that recapitulate the "first page" of the paper report. A full-color portable document format (PDF) report was attached to the XML file. Using multiple layers of security, these results were transmitted nightly from FMI to VUMC servers. The files were parsed, matched and incorporated into VUMC patient charts (APPENDIX 1). Provider information was matched to trigger existing notification methods[88]. A log of all transmissions and any detected errors was maintained to evaluate the process (APPENDIX 3). Differences in error rates were compared using the chi-squared test. This work was determined to be non-human subjects due to its quality improvement nature (VUMC IRB #140813).

The multidisciplinary team consisted of: a nurse who serves as the "tissue librarian" helped conduct the workflow analysis. Physician members of the Vanderbilt Ingram Cancer Center (VICC) Research Informatics Core helped define the data and transmission structure as well as perform the evaluation and data analysis. Administrators from VICC helped to define functional and security requirements. Members of the VUMC HealthIT department created the XML receiver and parser to incorporate the report into the EHR and trigger automated notification. Staff from FMI performed the molecular profiling, helped define the data and transmission structure, and worked with VUMC HealthIT to implement the transmissions.

Results

The ordering and reporting workflow analysis revealed that the faxed and scanned lab reports were difficult to read and providers were not consistently being notified when the reports returned. The data structure and transmission protocol addressed these issues.

Transmission feasibility was demonstrated with an initial test transmission of 524 reports that had been created prior to September of 2014. We alerted providers that they would receive multiple notifications via the EHR as these old reports were delivered. Logs of files transmitted by FMI and received by VUMC were identical with 100% of files being received, without evidence of files being lost in transmission. However, 33 files (6.29%) failed to be incorporated into the EHR were investigated (figure 3.2). All reports that integrated into the EHR demonstrated text quality identical to other EHR elements, and the PDF contained color information at resolution comparable to internet publications (figure 3.1). The system went live with nightly transmission of new reports in October 2014.

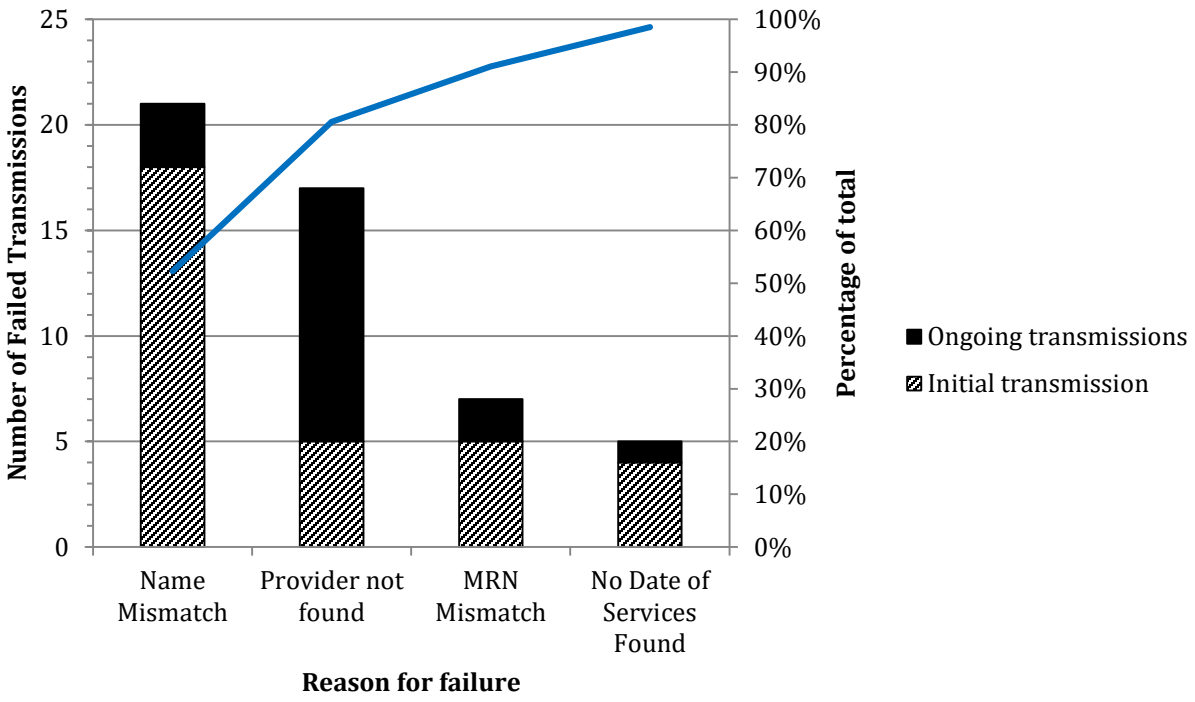


Figure 3.2 Pareto diagram for 50 report transmission failures. The bars are subdivided by type of transmission. The initial transmission was evaluated to determine causes of failures to improve the ongoing transmissions. The transmissions that could not match a provider were still accepted into the patient's chart, but were also routed to a team to manually assess the proper physician to report.

Errors during the initial transmission included reports failing to integrate into the EHR due patient name misspellings, incorrect medical record numbers (MRN), lack of a date of service, and missing provider information (figure 3.3). During the initial transmission there were also two periods of time in which the receiving VUMC server was unable to communicate with a separate system that verifies patient information, resulting in those reports being rejected. Examples of MRN errors that resulted in reports being rejected included using the pathology accession number for MRN, including the "#" symbol, and typographical errors. Providers who were new to the division accounted for the majority of provider errors; however, these were not rejected outright, but rather incorporated into the EHR and routed to a results reporting team for manual provider notification. Taken as a whole, many errors were attributable to mistakes in the manual ordering process when the test was first ordered.

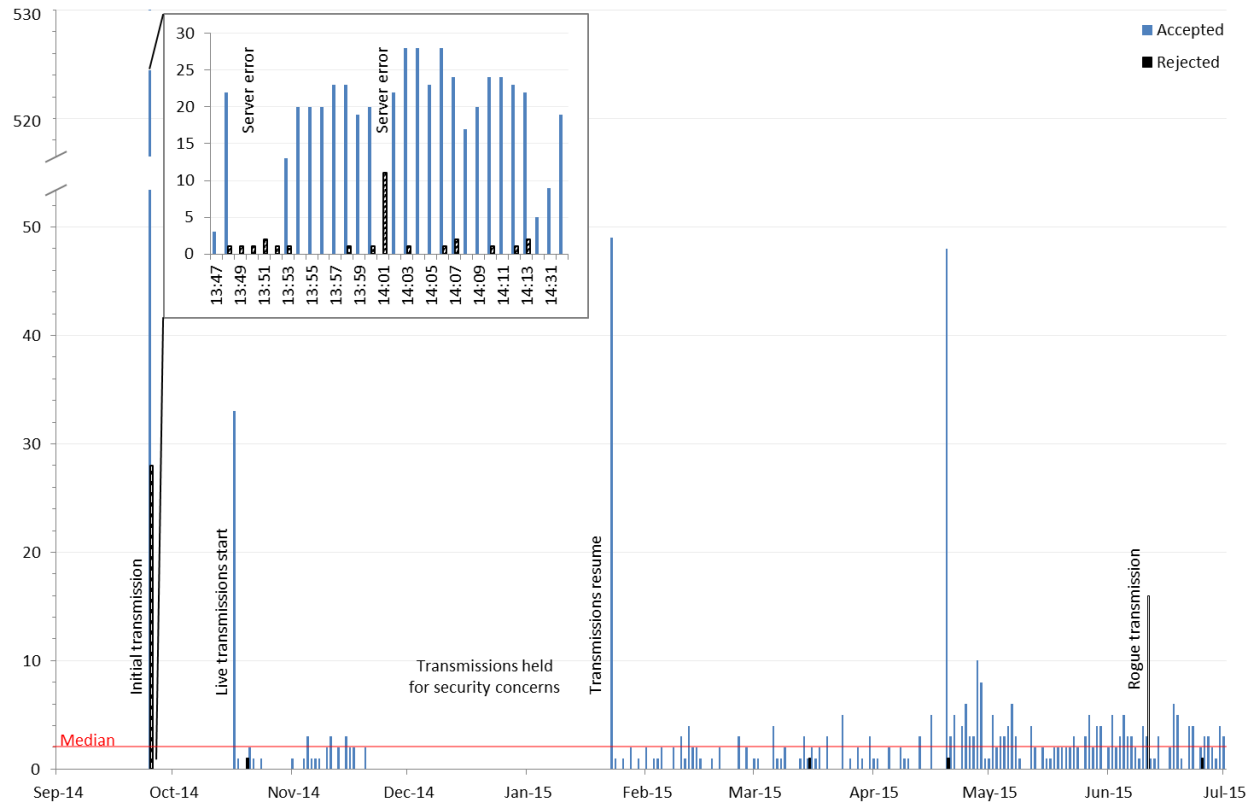


Figure 3.3 Run chart demonstrating the failed and successful genomic report transmissions over one year. Median transmissions were two per day. An initial transmission over two hours of 524 reports is further depicted (inset). At two points during the transmission, the receiving system at VUMC could not communicate with the server that validates patient information, resulting in those reports being rejected. From late November 2014 to mid-January 2015, transmission was halted over security concerns. An erroneous transmission of 16 charts in June 2015, represents the only deviation from only sporadic rejected charts over seven months of continuous operations.

Following reporting this conclusion to the ordering providers, the error log demonstrates fewer files were subsequently rejected as a result of increased diligence in ordering. Mean rejection rate decreased from 6.29% [95% CI: 4.21- 8.38%] to 3.84% [95% CI: 1.92-5.74%, $p<0.001$]. In the fall of 2014, the transmissions were halted due to security concerns until January of 2015. When transmissions resumed in January of 2015, continued low rates of rejected files were observed until in June a transmission of 16 files of non-VUMC patients were transmitted and rejected by VUMC servers (figure 3.3). This "rogue" transmission also contained the first file that failed for technical reasons with corrupted contents and is under investigation.

Security continued to be a concern over the continued operation of this system. As the authentication was based on IP address, file type confirmation and trust, when the FMI servers began behaving unusually there was high concern within VUMC. The FMI servers would undertake a "brute force" probing of available ports on the VUMC receiver server as well as testing for system files. Other healthcare business associates of VUMC do not evidence such behavior. The response from FMI to the concerns of VUMC was to assert that this behavior was part of their security systems to scan all other machines that their systems were connected to. However, this behavior did not immediately stop when asked. Clearly one of the most concerning aspects of this situation is the possibility of someone other than FMI, e.g., a "bot" or other malware, was controlling their systems, which had the ability to write files to our systems. Utilizing a third-party lab does open the medical institution to greater risk as data must transit outside of the institution and this creates the potential for data loss, as has been seen at other labs[89]. For this reason, application programming interfaces (APIs) that create a secure interface layer and only "pull" exposed data on demand, such as SMART on FHIR, may be an attractive alternative to the "push" architecture employed in this pilot project.[90]

Currently, the system operates to receive a median of two reports daily, and has successfully integrated over 900 test reports into patient charts. Providers have expressed appreciation of the automatic EHR-based notification as well as better readability of the color PDF. Additionally, providers were able use the EHR's text search functionality to quickly find a patient's genomic results within their chart, a feature previously unavailable with scanned documents.

With the success of parsing variants of unknown significance from PDFs in the research enterprise, we undertook to add this functionality to the enterprise EHR. Using the code created to parse the files (APPENDIX 2), this was implemented to parse out the VUS and display them directly in the chart with the following text accompanying them: "One or more variants of unknown significance (VUS) were detected in this patient's tumor. These variants have not yet been adequately characterized in the scientific literature at the time this report was issued and/or the genomic context of these alterations make their

significance unclear." This was successfully implemented and appears to be functioning as would be expected.

Discussion

This system demonstrates the feasibility and evaluation of an automated, secure electronic reporting solution of molecular profiling reports from a 3rd-party lab into an EHR that improves information fidelity by preserving text quality and report color and allows utilization of an existing physician notification system. The accuracy of incorporating results into the correct chart was significantly improved by informing ordering providers regarding the importance of accurate information during the ordering process.

While there were concerns that prompted multiple reviews of the system, this system demonstrates that files can surmount these concerns. There is the possibility of files lost in transmission that would not be logged; however, the initial test was accurate, and monitoring charts for undiscovered errors in the system has not revealed missing files. Many failures of the system were attributable to propagated ordering errors, and can be improved with awareness by of the importance of accurate information during ordering.

This system uses common document types (XML and PDF) that could be extended to other practices. The importance and challenges of reporting cancer genetic testing was highlighted recently by the College of American Pathology with proposed XML-based "electronic cancer checklists" which could be implemented with a similar system as we have described here[15]. However, the XML document that was developed would not necessarily be extensible to other 3rd-party labs or clinical sites. A potential future solution would use the interoperable Health Level 7 International's Fast Healthcare Interoperability Resources (FHIR) standard, using the SMART platform and genomics extensions (SMART on FHIR Genomics)[56]. With increasing attention to the importance of standards in the interoperability of cancer-specific data, it is likely that the SMART on FHIR Genomics framework will be a solution for future exchanges of genomic information[91]. While using such a standard would not necessarily resolve the

process issues encountered during this implementation, it would facilitate implementation with vendor EHRs which support such data standards[92–94].

Future efforts to extend this system at VUMC could include using computerized physician order entry (CPOE) to order the molecular profiling test. Closed-loop systems such as that offered by Syapse (Palo Alto, CA) incorporate CPOE into the physician's workflow on certain EHR platforms. However, CPOE can require prohibitively costly interfaces between EHR, laboratory information systems, and 3rd-party labs. In an ideal setting, CPOE can automatically populate names, MRNs, and other identifying information; thus eliminating typographical errors in ordering. Additionally, CPOE can provide an opportunity for clinical decision support in the increasingly complex practice of precision oncology. As the use of molecular profiling of tumors increases, solutions to integrate results reporting into clinical workflows will allow greater utility from these tests.

CHAPTER 4

Pragmatic precision oncology: the secondary uses of clinical tumor molecular profiling

Introduction

Cancer is fundamentally a disease that arises from disorder in our DNA and represents the complex convergent end state of a cacophony of genetic variants[1]. While large scale sequencing projects have recruited thousands of patient samples[95], the heterogeneity of cancer ensures that ongoing study of tumor genomics will be required to understand not only the genetic origins of the disease, but also its multiple means of escape from treatment[96–98]. Tumor genetic testing, also known as molecular profiling, is becoming the standard of care for an increasing number of cancer types[8,82,97,99]. Molecular profiling of patient tumors is expanding beyond individual base pair sequencing into next generation sequencing (NGS) panels of the exons of numerous potential genetic drivers of the cancer[9]. The increasingly routine use of NGS molecular profiling in oncology clinics generates data to inform the care of the individual patient, but can also in secondary use inform the science of precision oncology. Using informatics tools to automatically parse and aggregate molecular profiling reports into a real-time database represents a pragmatic use of clinical data to create resources that can be used for robust secondary uses.

In the course of clinical care, the results of molecular profiling are reported back to the ordering provider. The mechanism of results reporting varies from faxes, online portals internal or external to an electronic health record (EHR), or direct integration into the EHR. While there exists an HL7 V2.5.1 standard for the transmission of some clinical genomic results, the use cases for molecular profile reporting are not implemented widely[55]. Previously, we have described a clinical transmission system of tumor molecular profiling reports from a 3rd-party sequencing laboratory into an EHR[17,100] for real-time clinical practice. In addition, this structured data feed allows molecular results to be parsed, restructured and stored into a database that can be used to further inform the science and practice of precision

oncology. This is similar to other pragmatic databases that are automatically populated from clinical data from EHRs[25]. Herein, we describe the creation and use cases for an automated, real-time database of annotated tumor variants.

Methods

Oncology providers at Vanderbilt University Medical Center (VUMC) order tumor molecular profiling of metastatic, advanced, or recurrent cancers according to the individual clinical situation (Figure 4.1). The tumors undergo molecular profiling at a 3rd party laboratory of a panel of exons according to the laboratory's protocol. The results of the molecular profiling are packaged into a custom extensible markup language (XML) document which is transmitted securely from the laboratory sever to VUMC servers on a nightly basis. Variants are represented with Human Genome Variant Society standardized syntax for naming and genomic alterations[48,49]. Variants with known significance for the diagnosis, prognosis, or therapy of the specific cancer type are structured in the XML tags, while variants lacking that established clinical significance, "variants of unknown significance" (VUS), are only found within the attached PDF. The reports are parsed and displayed in the EHR for clinical use, but also reflected to a secure server for database storage. This database creation mechanism was approved by the VUMC institutional review board as an information repository preparatory to research.

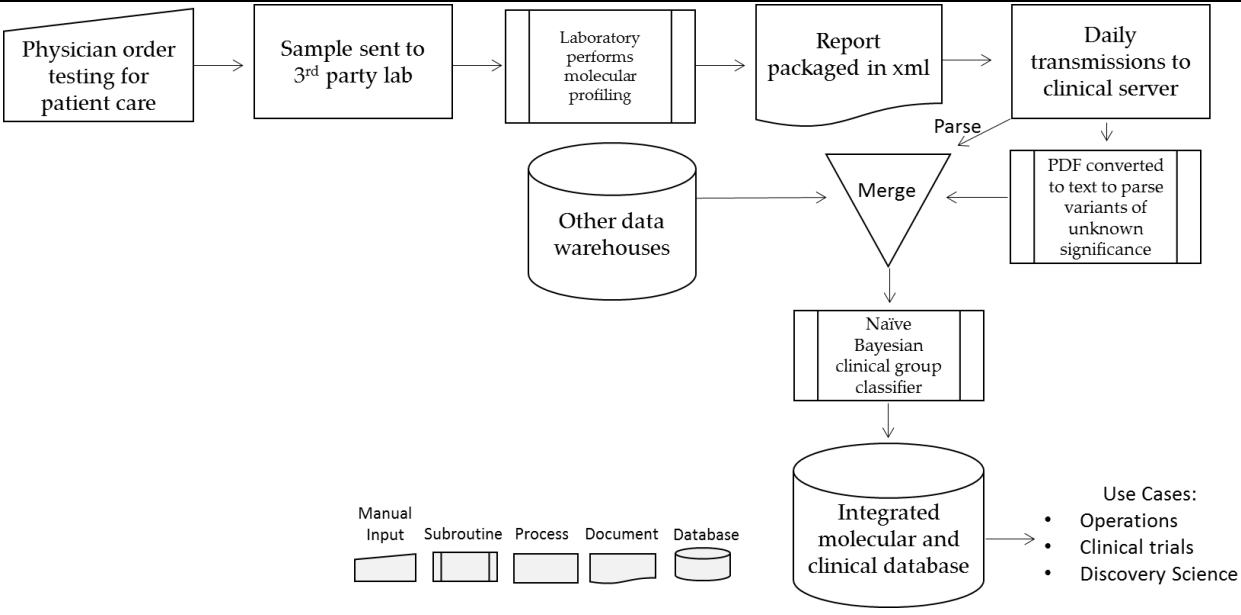


Figure 4.1: Flow diagram of data processing. Tumors undergo molecular profiling according to the clinical situation. Molecular profiling reports are packaged into custom XML documents and transmitted daily to clinical servers to incorporate into the chart. The XML is parsed for "actionable" variants and sample metadata. A PDF of additional information is converted to text and parsed for "variants of unknown significance". These data sources are merged with extant clinical data warehouses. A classifier takes heterogeneous metadata strings and classifies reports into one of 18 clinical categories. The resultant data is stored in an integrated molecular and clinical database for multiple use cases.

The daily incoming XML documents are parsed by a PERL (v 5.14.2) script using the XMLSimple package (v2.20) for variants of known clinical significance as well as metadata regarding histology, anatomic site, and logistical information (APPENDIX 2). The PDF undergoes command line conversion to a text document. This text document is then processed using regular expressions in PERL to extract the VUS. The data is then stored in multiple tables in a relational database in a secure server maintained by the Vanderbilt-Ingram Cancer Center's Research Informatics Core.

For operational and other needs, it was necessary to classify cases into "clinical groups" which roughly correspond to the subject matter expertise of the treating physicians (e.g. lung, breast, central nervous system). To classify incoming samples to the clinical groups, multiple potential classifiers were tested. A naïve Bayesian classifier was trained with 10-fold cross validation using R (v 3.2.2), the e1071 package (v. 1.6), and the ranger (0.3.0) package. Ordering provider, patient gender, and submitted histology text

variables for 800 expertly-annotated samples were parsed and trained for classification into 18 clinical categories.

To evaluate the accuracy of the parsing and storage protocols, results were compared to a published dataset manually abstracted from the reports of some of the same patients[101]. Samples of 80 reports were manually re-evaluated by two investigators and compared to those generated by the parsing algorithm. The R statistical package was used to calculate descriptive statistics as well as Cohen's kappa, accuracy, precision and recall. All appropriate statistical tests were two-sided.

The existence of the database was communicated to basic science, translational, operations, and clinical researchers within VUMC. Data use cases information was collected and managed using REDCap electronic data capture tools hosted at VUMC [102]. Institutional review board approval was required for data requests from investigators for subsequent research involving personal health information.

Data for comparison to The Cancer Genome Atlas (TCGA) was obtained using the cgsdr (v1.2.5) package in R (v3.2.2) which was also used for Mann-Whitney U and Chi-square statistical analysis. All appropriate statistical tests were two-sided.

Results

The database created by parsing molecular profile reports is accurate and useful across a broad range of precision medicine secondary uses. Over one year of operation it contained 819 unique molecular profiling reports. The reports were parsed to 3435 variants of known significance and 7185 VUS, with a median of 13 variants per report (4 known and 9 VUS per report). Duplicate reports were purged by requiring unique pathology sample identifiers. Reports received were dependent on the test ordered by providers.

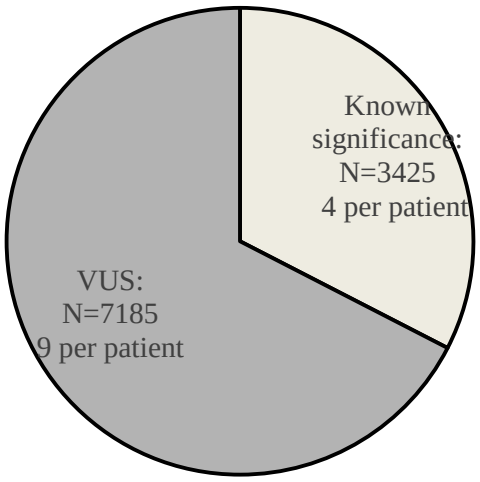


Figure 4.2: Variants of known and unknown significance observed over one year of operation, 3435 variants of known significance and 7185 VUS were parsed and stored. A median of 13 variants were found per sample: median 4 of known significance (range 0-26) and median 9 of unknown significance (range 1-95).

In comparison to the manually extracted cohort of 103 patients[101], the automated XML parsing method was concordant to those data (kappa=0.95) with 13 discordant variants (median 0, range 0-7) out of 365 variants of known significance (figure 3). All discordant variants were false negatives by the manual method and verified to be correctly parsed by the algorithm by repeat manual review of the original molecular profiling report. Of the VUS parsed from text using regular expressions, 100% (308 of 308) from 20 charts were verified from the original report to be accurate. No false positive variants were observed.

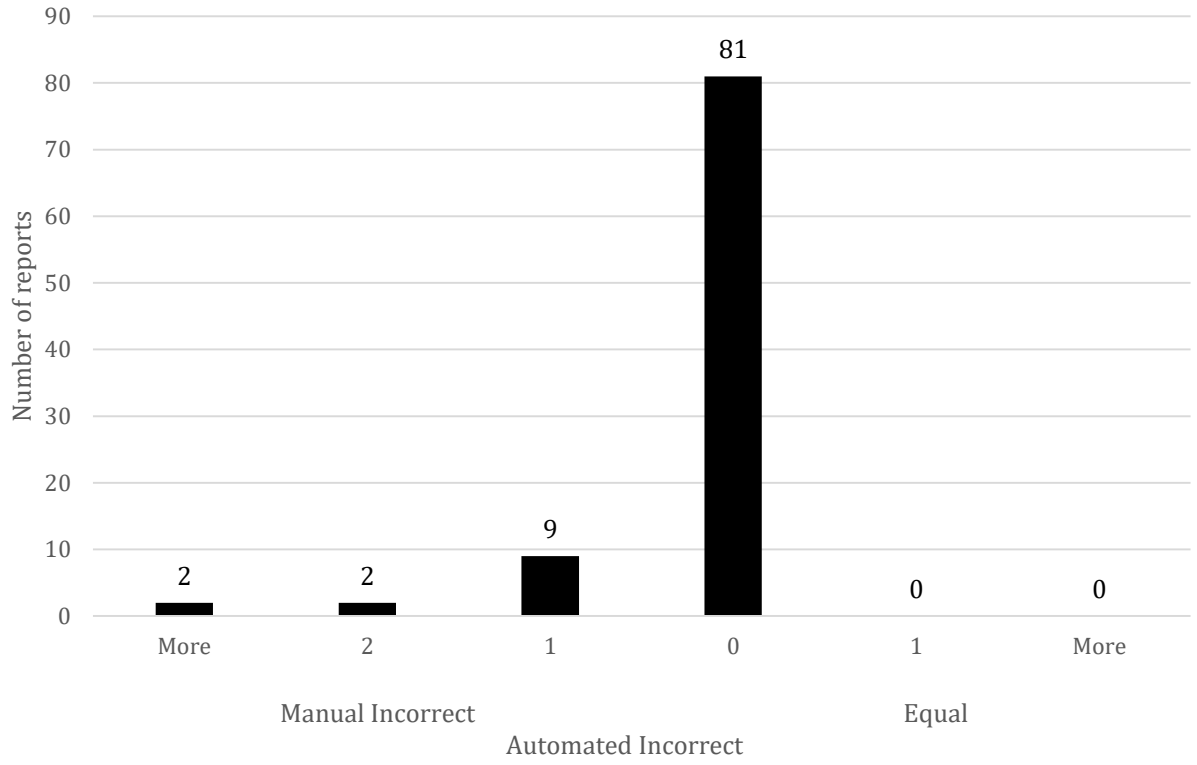


Figure 4.3: Automated parsing of variant information is more accurate than manual abstraction. A histogram of discordant reports between manual and automated data extraction. Both methods were concordant in 81 of 94 charts reviewed (kappa = 0.95). All observed discordant reports (n=13) were false negative variants of the manual method verified by review of original report. The automated method did not produce any detected false negative nor false positives.

A Bayesian classifier was trained to classify incoming samples into one of 18 clinically relevant groups, performing with accuracy of 0.994 (795 of 800 samples correctly classified) (Table 4.1). The classifier utilized three sample metadata elements to classify samples. The factors were provider (48 levels), submitted diagnosis (149 levels), gender (2 levels). Of the 5 misclassified examples, 2 were sarcoma likely reflecting the disparate anatomic sites and nomenclatures for sarcomas. Three other misclassifications reflected uncommon tumor types seen by a provider who predominantly sees a common tumor type. In prospective evaluation of 106 cases, the classifier performed with 0.953 accuracy.

Table 4.1 Naïve Bayesian and ensemble tree classifiers confusion matrices

(A)

Predicted↓ \ Reference↓	Breast	Cholangio	Central nervous	Gastrointestinal	Genitourinary	Gynecologic	Hepatocellular	Hematologic	Head and Neck	Lung	Melanoma	Neuroendocrine	Pancreatic	Renal	Sarcoma	Skin	Thyroid	Unknown primary
Breast	186	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cholangiocarc.	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Central nervous	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Gastrointestinal	0	0	0	145	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Genitourinary	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0
Gynecologic	0	0	0	0	0	76	0	0	0	0	0	0	0	0	0	0	0	1
Hepatocellular	0	0	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0
Hematologic	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0
Head and Neck	1	0	0	0	0	0	0	0	58	1	0	0	0	0	0	0	0	0
Lung	0	0	0	0	0	0	0	0	0	65	0	0	0	0	0	0	0	0
Melanoma	0	0	0	0	0	0	0	0	0	0	72	0	0	0	0	0	0	0
Neuroendocrine	0	0	0	0	0	0	0	0	0	0	0	16	0	0	0	0	0	0
Pancreatic	0	0	0	0	0	0	0	0	0	0	0	0	14	0	0	0	0	0
Renal	0	0	0	0	0	0	0	0	0	0	0	0	0	13	0	0	0	0
Sarcoma	1	0	0	0	0	0	0	0	0	1	0	0	0	0	41	0	0	0
Skin	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0
Thyroid	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	0
Unknown prim.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	30

Confusion matrix for naïve Bayesian classifier of samples into clinical groups (A). Overall accuracy 99.4% (795 of 800 correctly classified). Variables included for classification were patient gender, ordering provider name, and submitted diagnosis. Other variables that were considered but reduced performance were biopsy site, tokenized diagnosis string with stop words removed, and including genetic variant data. This classifier was evaluated prospectively on 106 cases since this manuscript was originally submitted and performed with 95.3% accuracy on those cases.

Other classification systems attempted include decision tree and random forest (B). Most tree-based classification packages fail at this task as the "ordering provider" feature contain 48 levels and the "submitted diagnosis" feature contains 149 as they can only process upwards of 32 levels. However, the ranger (v0.3.0) package in R is able to provide a classification estimate using an ensemble tree method. However, this classifier proved to only be 74% accurate.

(B)

	Breast	Cholangio	CNS	GI	GU	GYN	HCC	Heme	HN	Lung	Melanoma	Neuroendocrine	Pancreatic	RCC	Sarcoma	Skin	Thyroid	Unk
Breast	187	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cholangio	0	0	0	9	0	9	0	0	1	0	0	0	0	0	0	0	0	0
CNS	3	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GI	0	0	0	132	0	6	0	0	1	0	2	0	0	0	0	0	0	0
GU	1	0	0	3	0	1	0	0	0	0	2	0	0	0	0	0	0	0
GYN	0	0	0	0	0	77	0	0	0	0	0	0	0	0	0	0	0	0
HCC	0	0	0	5	0	5	0	0	1	0	0	0	0	0	0	0	0	0
Heme	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HN	0	0	0	0	0	26	0	0	32	0	0	0	0	0	0	0	0	0
Lung	1	0	0	0	0	2	0	0	2	60	1	0	0	0	0	0	0	0
Melanoma	1	0	0	2	0	4	0	0	0	0	65	0	0	0	0	0	0	1
Neuroendocrine	3	0	0	8	0	5	0	0	0	1	1	0	0	0	1	0	0	0
Pancreatic	0	0	0	6	0	4	0	0	0	0	0	0	4	0	0	0	0	0
RCC	0	0	0	1	0	5	0	0	0	0	7	0	0	0	0	0	0	0
Sarcoma	2	0	0	3	0	17	0	0	0	0	1	0	0	0	18	0	0	0
Skin	0	0	0	7	0	2	0	0	0	0	0	0	0	0	0	0	0	0
Thyroid	0	0	0	4	0	4	0	0	0	0	3	0	0	0	0	0	0	0
Unk	0	0	0	11	0	7	0	0	0	0	4	0	0	0	0	0	0	8

Multiple users presented requests for data to answer varied precision oncology use cases. Fifteen use cases were classified into operational, clinical trial, and basic science research projects [table 4.2]. Data extraction cohorts ranged from 17 reports to all reports (median = 83 reports). Variables of interest included ordering statistics by month and histology, variants information for evaluation of a novel assay, molecular biomarker estimates for clinical trial feasibility estimates, visualization of VUMC population variant data relative to other published data sets, and variant information from a previously-defined cohort of interest. Many informal comments were positive regarding the existence, utility and accuracy of this database.

Table 4.2 Precision oncology use cases from aggregated tumor molecular profiling data.

TYPE	DATA REQUEST USE CASE	OUTCOME
OPERATIONS	1. Figures for NCI cancer center support grant	Successful application
	2. Molecular profiling utilization statistics	Institutional directive to increase utilization
	3. Validation of novel NGS assay	Ongoing development of assay
TRIAL	1. Prioritizing future clinical phase 1 trials of targeted therapeutics	Better collaborations with pharmaceutical companies to open trials
	2. Solid tumors with MET and NTRK1,2,3 variants	Successful identification of potential trial patients
	3. Breast cancer patients with FGFR mutations or 11q amplifications	Successful identification of potential trial patients
	4. Solid tumors with MAP2K, MAP2K2 variants	Successful identification of potential trial patients
	5. Tumors with ERBB family mutations	Successful identification of potential trial patients
	6. Solid tumors with novel BRAF mutations	Successful identification of potential trial patients
RESEARCH	1. Network analysis of VUS in breast cancer	Work presented as abstract at breast cancer conference
	2. Triple negative breast cancer patients with PIK3CA variants	Ongoing mechanistic discovery work
	3. Novel ERBB2 mutations for validation	Ongoing mechanistic discovery work
	4. Potential germline implications of molecular profiling	Preliminary data for R01 application
	5. Genetics of GI malignancies with signet ring histology or peritoneal metastases	Preliminary data for career development grant
	6. SRC variants in lung cancer	Ongoing mechanistic discovery work

Data for stage from TCGA specimens was obtained from data regarding clinical stage across multiple tumor types with the number of metastatic tumors was taken as a percentage and an absolute number. Number of metastatic tumors from VUMC are given in Table 4.3.

Table 4.3. Metastatic (Stage IV) samples in TCGA vs VUMC

Tumor Type	TCGA Stage IV (N)	VUMC N
Breast	2% (15)	230
Colorectal	16% (44)	133
Hepatocellular carcinoma	1.5% (6)	12
Lung	4% (9)	77
Cholangiocarcinoma	19.5% (7)	17
Ovarian	14.2% (79)	43
Pancreatic	0.5% (1)	19
Melanoma	76.8% (367)	102

Discussion

In addition to informing clinical care, tumor genetic testing can be used to advance the practice and science of precision oncology. A system to parse and aggregate structured tumor genetic variant information is accurate in comparison to a manually curated reference standard and useful in addressing multiple investigations across an academic medical center. The discrepancies between the reference and the automated parser likely indicates filtering during the creation of reference standard of variants that were not clinically relevant to that analysis. This discrepancy illustrates that in practice, the definitions of "known significance" or "clinically relevant" are constantly evolving entities.

The evolution of variant classification will continue to be a challenge in the annotation of molecular profiling results. Of particular note is the role molecular biomarkers will play in future precision oncology trials. One illustrative example is that of an experimental drug, lucitanib, that targets fibroblast growth factor (FGF) pathway alterations (Table 4.2, Trial example #3). This first-in-class drug was studied through a clinical trial (NCT02202746) with molecular inclusion criteria of FGF receptor mutations or chromosome 11q amplifications[103]. Any patients meeting the molecular inclusion criteria prior to the trial opening (and thus prior to incorporation into the knowledge base used to annotate the reports) would not have been identified as eligible. Though the methods we have described, we were able to identify all patients who met these inclusion criteria so they could be screened for this trial.

Since the methodology we describe has been implemented, additional investigators have requested data to determine feasibility and to assist in screening for patients meeting molecular inclusion criteria. Precision medicine will increasingly rely on "basket trials" such as NCI-MATCH[104,105], where patients are included not by histology but rather by molecular inclusion criteria[23,106]. In such trials, being able to screen across populations for patients meeting molecular inclusion criteria will be critical for efficient study conduct. Investigators can use this database to query across populations of cancer patients for molecular biomarkers to determine if a clinical trial is feasible, and once open, to identify eligible patients. Investigators conducting precision oncology clinical trials, particularly early phase clinical trials, have already used this database effectively.

Generally, the practice of precision oncology will require a greater awareness of molecular biomarkers that may fall outside the daily practice of an oncologist. Maintaining current knowledge of molecular inclusion criteria of basket trials is challenging as novel or low frequency biomarkers may be unfamiliar to the physician. This becomes an increasingly difficult task in clinical practice as both molecular profiling and precision oncology trials increase. Additionally, future annotation of variants as clinically relevant relies on the molecular diagnostic lab to maintain a continually updating knowledgebase of all molecular biomarkers for clinical trials. The failure at any of these steps prevents patients from being

enrolled on precision oncology trials or offered potentially therapeutic treatments. A system such as this can provide another layer of knowledge to ensure all patients who may benefit from a molecularly defined clinical trial are identified.

In addition to clinical trial use cases, basic science and translational researchers have requested data to develop hypotheses for foundational research. Several requests for data came from cancer biologists studying genetic mechanisms of cancer genetics and biomarkers. As demonstrated in Table 4.3, the TCGA has a large number of samples, but in most histologic types the predominance of samples are early stage patients rather than metastatic. In contrast, the VUMC population is almost entirely metastatic patients. Therefore the absolute numbers of metastatic samples are larger in the pragmatic VUMC cohort than in the TCGA. This is important because it would be hypothesized that patients in the pragmatic cohort who have late stage, aggressive disease, and are likely to have been previously treated would be an enriched population to discover genetic means of aggressive phenotype, treatment resistance, and rapidly fatal disease.

When a population of 925 breast cancer samples in TCGA is compared to 230 VUMC samples the number of mutations is significantly increased (Figure 4.4). In addition, the VUMC cohort demonstrated substantially more mutations in genes associated with treatment resistance and aggressive phenotype: ESR1 (7.0% vs 0.11% $p=4.9x10^{-8}$), PIK3CA (42.1% vs 34.2% p=0.036), ERBB2 (7.4% vs 2.5% $p=5.4x10^{-4}$), TP53 (59.1% vs 34.8% $p=1.0x10^{-12}$). When these data are combined with a greater ability to phenotype the patients using clinical data within the EHR, the power of a pragmatic approach to cancer clinicogenomic database creation is evident.

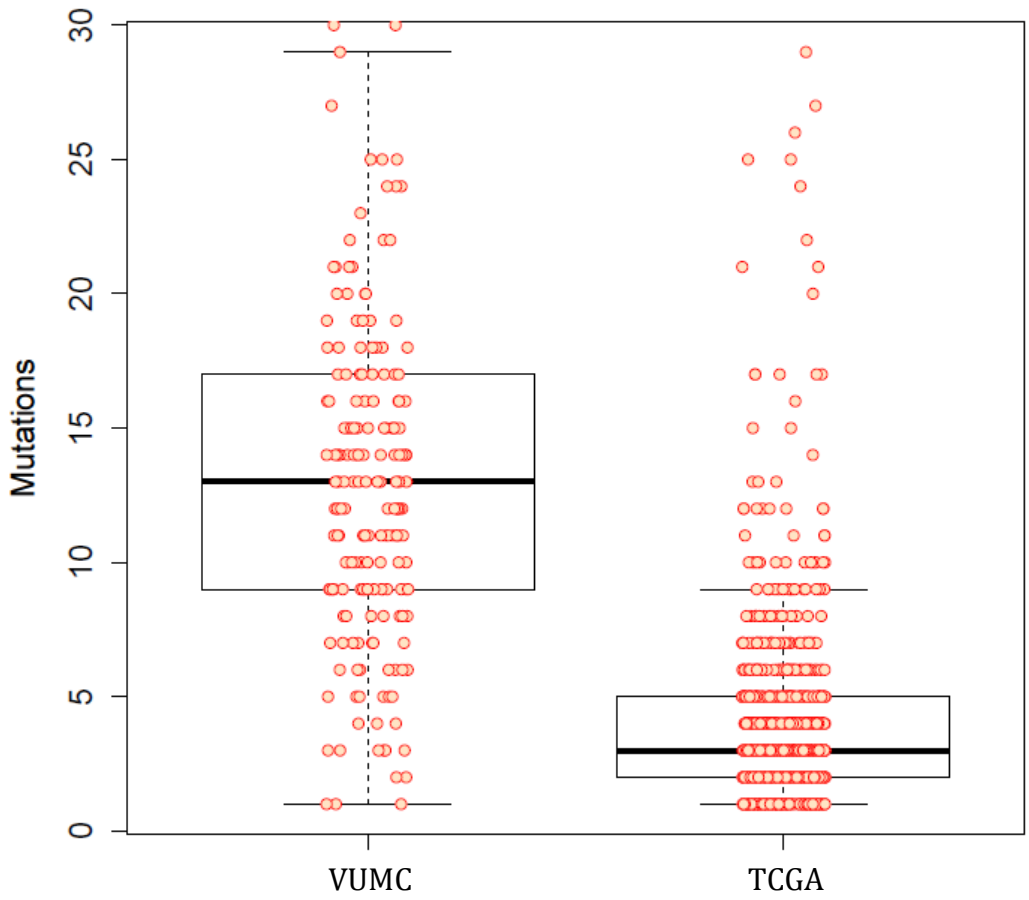


Figure 4.4. Mutations per patient are higher in the pragmatic VUMC cohort than TCGA. The VUMC cohort has a median 13 vs 3 mutations in TCGA ($p=1x10^{-13}$). Mutations were assessed across the same panel of genes as are assayed in the clinical VUMC panel.

Aggregating and structuring clinical data for secondary use represents a pragmatic and cost-effective means to facilitate translational science.[25,27] As more patients accrue in the database, the power to discover novel cancer biomarkers and to drive discovery science increases. The strengths of this approach are modeled on the successful strategy of the Electronic Medical Records and Genomics (eMERGE) network by linking molecular information to existing clinical data warehouses[71]. Furthermore, the patients who undergo molecular profiling in a pragmatic manner represent unique clinical substrates of exceptional responders or resistant tumors in many cases. Patients within this database constitute more advanced cancers than those represented in databases such as The Cancer Genome Atlas. These aggressive, refractory or unusual clinical situations may reveal new insights into cancer biology.

Our methodology is limited by the custom nature of the parser that was created for the custom XML document. As such, it is not extensible to other molecular profiling reports and constitutes a limitation of these methods. This limitation illustrates the need for a consensus means of structured molecular profiling reporting, such as the efforts underway with HL7 FHIR[56] so that the information is interoperable and standardized tools can be built.

This system represents a pragmatic approach to the practice and science of precision medicine in oncology. Structuring and aggregating accurate molecular profiling data facilitates operational, clinical trial, and discovery science use cases that further the field and improve patient care. As more molecular profiling is utilized in cancer care, the data it produces can feed into a learning health system that continues to drive precision oncology practice and discovery.

CHAPTER 5

Conclusions

This work demonstrates the implementation of an electronic transmission of tumor molecular profile reports to clinical systems and for secondary uses can be informed by a data and transmission framework. With regards to specific aspects of the framework, this system illustrates several positive aspects as well as some shortcomings. Ordering was largely beyond the scope of the intervention, but much of the clinical data that accompanied the reports was obtained at the ordering stage. Furthermore, improvement in the accuracy of ordering information improved the accuracy of the PHI matching for the receiver.

The data types that were contained within the file transmission were structured within the XML for easy parsing. Genetic information was reported according to HGNC nomenclature standards. The largest shortcomings for the data type that was implemented were the lack of an unambiguous depiction of the DNA level change, nor information regarding allelic fraction. This was of little consequence to the clinical use cases; however, it prevented the unambiguous mapping of genetic information that is required for many other bioinformatics tools and for interinstitutional collaborations.

The annotation for genetic variants was contained within the report. The annotations for the variants of known significance were structured within the XML document, but those for VUS could only be parsed from the PDF. However, the genetic information was able to be parsed and structured accurately from the PDF, allowing discrete capture of the VUS as well as the variants of known significance.

Transmission was achieved via a secure web service employing HTTPS security. This system functioned well; however, there were persistent concerns of the security and behavior of the FMI systems. In the future, a system such as DIRECT would be more secure and standardized. The parser took data from the receiver and displayed it directly into the EHR as is the custom of HealthIT. As such, the pathology laboratory was bypassed. To remedy the lack of reporting to pathology, a secure portal within VUMC was set up to monitor transmissions and initiate billing.

Perhaps the greatest success of this system is the diversity of use cases the aggregated database was able to satisfy. This is not necessarily due to the design nor structure of the database, but rather attributable to the value of the data. The pragmatic population of patients represents a unique but clinically important subset of aggressive, refractory and pretreated patients that benefit most from clinical trials, but also represent a genetic substrate that is likely enriched in mechanisms of treatment resistance and aggressive phenotype. The population is over 1000 patients as of this writing, and it continues to grow by 30-40 each month. With the increase in biomarker-driven precision oncology, the data from these patients will hopefully continue to improve the treatment of cancer.

REFERENCES

1 Hanahan D, Weinberg RA. Hallmarks of Cancer: The Next Generation. *Cell* 2011;**144**:646–74. doi:10.1016/j.cell.2011.02.013

2 Garraway LA, Verweij J, Ballman KV. Precision Oncology: An Overview. *JCO* 2013;**31**:1803–5. doi:10.1200/JCO.2013.49.4799

3 Overview of Targeted Therapies for Cancer - My Cancer Genome. https://www.mycancergenome.org/content/molecular-medicine/overview-of-targeted-therapies-for-cancer/ (accessed 29 Feb2016).

4 Salto-Tellez M, Gonzalez de Castro D. Next Generation Sequencing: A Change of Paradigm in Molecular Diagnostic Validation. *J Pathol* 2014;:n/a – n/a. doi:10.1002/path.4365

5 Mosrati MA, Malmström A, Lysiak M, *et al.* TERT promoter mutations and polymorphisms as prognostic factors in primary glioblastoma. *Oncotarget* 2015;**6**:16663–73. doi:10.18632/oncotarget.4389

6 Kerr SE, Schnabel CA, Sullivan PS, *et al.* Multisite validation study to determine performance characteristics of a 92-gene molecular cancer classifier. *Clin Cancer Res* 2012;**18**:3952–60. doi:10.1158/1078-0432.CCR-12-0920

7 NCCN Clinical Practice Guidelines in Oncology. http://www.nccn.org/professionals/physician_gls/f_guidelines.asp#site (accessed 22 Aug2014).

8 Van Allen EM, Wagle N, Levy MA. Clinical analysis and interpretation of cancer genome data. *J Clin Oncol* 2013;**31**:1825–33. doi:10.1200/JCO.2013.48.7215

9 Frampton GM, Fichtenholtz A, Otto GA, *et al.* Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nat Biotech* 2013;**31**:1023–31. doi:10.1038/nbt.2696

10 Cheng DT, Mitchell TN, Zehir A, *et al.* Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT). *The Journal of Molecular Diagnostics* 2015;**17**:251–64. doi:10.1016/j.jmoldx.2014.12.006

11 Pritchard CC, Salipante SJ, Koehler K, *et al.* Validation and Implementation of Targeted Capture and Sequencing for the Detection of Actionable Mutation, Copy Number Variation, and Gene Rearrangement in Clinical Cancer Specimens. *The Journal of Molecular Diagnostics* 2014;**16**:56–67. doi:10.1016/j.jmoldx.2013.08.004

12 Claustres M, Kožich V, Dequeker E, *et al.* Recommendations for reporting results of diagnostic genetic testing (biochemical, cytogenetic and molecular genetic). *Eur J Hum Genet* 2014;**22**:160–70. doi:10.1038/ejhg.2013.125

13 Green RC, Berg JS, Grody WW, *et al.* ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet Med* 2013;**15**:565–74. doi:10.1038/gim.2013.73

14 Hehir-Kwa JY, Claustres M, Hastings RJ, *et al.* Towards a European consensus for reporting incidental findings during clinical NGS testing. *Eur J Hum Genet* 2015;**23**:1601–6. doi:10.1038/ejhg.2015.111

15 Simpson RW, Berman MA, Foulis PR, *et al.* Cancer biomarkers: the role of structured data reporting. *Arch Pathol Lab Med* 2015;**139**:587–93. doi:10.5858/arpa.2014-0082-RA

16 Ackoff RL. From data to wisdom. *Journal of Applied Systems* 1989;**15**:3–9.

17 Rioth MJ, Staggs DB, Hackett L, *et al.* Implementing and Improving Automated Electronic Tumor Molecular Profiling. *J Oncol Pract* Published Online First: 26 January 2016. doi:10.1200/JOP.2015.008276

18 Allen EMV, Wagle N, Levy MA. Clinical Analysis and Interpretation of Cancer Genome Data. *JCO* 2013;**31**:1825–33. doi:10.1200/JCO.2013.48.7215

19 Bombard Y, Robson M, Offit K. Revealing the Incidentalome When Targeting the Tumor Genome. *JAMA* 2013;**310**:795–6. doi:10.1001/jama.2013.276573

20 Tarczy-Hornoch P, Amendola L, Aronson SJ, *et al.* A survey of informatics approaches to whole-exome and whole-genome clinical reporting in the electronic health record. *Genet Med* 2013;**15**:824–32. doi:10.1038/gim.2013.120

21 Richards S, Aziz N, Bale S, *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* 2015;**17**:405–24. doi:10.1038/gim.2015.30

22 Tenenbaum JD, Sansone S-A, Haendel M. A sea of standards for omics data: sink or swim? *Journal of the American Medical Informatics Association* 2014;**21**:200–3. doi:10.1136/amiajnl-2013-002066

23 Sleijfer S, Bogaerts J, Siu LL. Designing Transformative Clinical Trials in the Cancer Genome Era. *JCO* 2013;**31**:1834–41. doi:10.1200/JCO.2012.45.3639

24 Denny JC, Giuse DA, Jirjis JN. The Vanderbilt Experience with Electronic Health Records. *Seminars in Colon and Rectal Surgery* 6;**16**:59–68. doi:10.1053/j.scrs.2005.08.003

25 Danciu I, Cowan JD, Basford M, *et al.* Secondary use of clinical data: The Vanderbilt approach. *Journal of biomedical informatics* 2014;**52**:28–35. doi:10.1016/j.jbi.2014.02.003

26 Roden D, Pulley J, Basford M, *et al.* Development of a Large-Scale De-Identified DNA Biobank to Enable Personalized Medicine. *Clin Pharmacol Ther* 2008;**84**:362–9. doi:10.1038/clpt.2008.89

27 Bowton E, Field JR, Wang S, *et al.* Biobanks and Electronic Medical Records: Enabling Cost-Effective Research. *Sci Transl Med* 2014;**6**:234cm3–234cm3. doi:10.1126/scitranslmed.3008604

28 Wei W-Q, Teixeira PL, Mo H, *et al.* Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *J Am Med Inform Assoc* Published Online First: 2 September 2015. doi:10.1093/jamia/ocv130

29 Warner JL, Levy MA, Neuss MN, *et al.* ReCAP: Feasibility and Accuracy of Extracting Cancer Stage Information From Narrative Electronic Health Record Data. *J Oncol Pract* 2016;**12**:157–8. doi:10.1200/JOP.2015.004622

30 Jones S, Anagnostou V, Lytle K, *et al.* Personalized genomic analyses for cancer mutation discovery and interpretation. *Sci Transl Med* 2015;**7**:283ra53–283ra53. doi:10.1126/scitranslmed.aaa7161

31 Tenenbaum JD, Shah NH, Altman RB. Translational Bioinformatics. In: Shortliffe EH, Cimino JJ, eds. *Biomedical Informatics*. Springer London 2014. 721–54.http://link.springer.com.proxy.library.vanderbilt.edu/chapter/10.1007/978-1-4471-4474-8_25 (accessed 23 Nov2015).

32 Gao J, Aksoy BA, Dogrusoz U, *et al.* Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. *Sci Signal* 2013;**6**:pl1. doi:10.1126/scisignal.2004088

33 Our Work | Global Alliance for Genomics and Health. https://genomicsandhealth.org/working-groups/our-work (accessed 14 Mar2016).

34 Dolin RH, Alschuler L, Beebe C, *et al.* The HL7 Clinical Document Architecture. *J Am Med Inform Assoc* 2001;**8**:552–69.

35 Innovative Translational Research Shared Resource SNaPshot Genotyping - Vanderbilt-Ingram Cancer Center. http://www.vicc.org/research/shared/translational/services/snapshot.php (accessed 12 Jan2014).

36 Meador CB, Micheel CM, Levy MA, *et al.* Beyond Histology: Translating Tumor Genotypes into Clinically Effective Targeted Therapies. *Clin Cancer Res* 2014;**20**:2264–75. doi:10.1158/1078-0432.CCR-13-1591

37 Pao W, Hutchinson KE. Chipping away at the lung cancer genome. *Nat Med* 2012;**18**:349–51. doi:10.1038/nm.2697

38 SEER ICD-O-3 Coding Materials. http://seer.cancer.gov/icd-o-3/ (accessed 15 Mar2016).

39 WHO | International Classification of Diseases for Oncology, 3rd Edition (ICD-O-3). http://www.who.int/classifications/icd/adaptations/oncology/en/ (accessed 23 Mar2016).

40 Wu T-J, Schriml LM, Chen Q-R, *et al.* Generating a focused view of disease ontology cancer terms for pan-cancer data integration and analysis. *Database (Oxford)* 2015;**2015**. doi:10.1093/database/bav032

41 SNOMED CT. https://www.nlm.nih.gov/healthit/snomedct/index.html (accessed 13 Mar2016).

42 Lee D, de Keizer N, Lau F, *et al.* Literature review of SNOMED CT use. *J Am Med Inform Assoc* 2014;**21**:e11–9. doi:10.1136/amiajnl-2013-001636

43 Cornet R, Nyström M, Karlsson D. User-directed coordination in SNOMED CT. *Stud Health Technol Inform* 2013;**192**:72–6.

44 Home - MeSH - NCBI. http://www.ncbi.nlm.nih.gov/mesh (accessed 13 Mar2016).

45 De Coronado S, Haber MW, Sioutos N, *et al.* NCI Thesaurus: using science-based terminology to integrate cancer research results. *Stud Health Technol Inform* 2004;**107**:33–7.

46 De Coronado S, Wright LW, Fragoso G, *et al.* The NCI Thesaurus quality assurance life cycle. *J Biomed Inform* 2009;**42**:530–9. doi:10.1016/j.jbi.2009.01.003

47 Herr TM, Bielinski SJ, Bottinger E, *et al.* A conceptual model for translating omic data into clinical action. *J Pathol Inform* 2015;**6**. doi:10.4103/2153-3539.163985

48 Human Genome Variation Society. http://www.hgvs.org/ (accessed 15 Nov2015).

49 Dunnen JT den, Antonarakis SE. Mutation nomenclature extensions and suggestions to describe complex mutations: A discussion. *Hum Mutat* 2000;**15**:7–12. doi:10.1002/(SICI)1098-1004(200001)15:1<7::AID-HUMU4>3.0.CO;2-N

50 RefSeq: NCBI Reference Sequence Database. http://www.ncbi.nlm.nih.gov/refseq/ (accessed 15 Nov2015).

51 Pruitt KD, Brown GR, Hiatt SM, *et al.* RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res* 2014;**42**:D756–63. doi:10.1093/nar/gkt1114

52 Clarke L, Zheng-Bradley X, Smith R, *et al.* The 1000 Genomes Project: data management and community access. *Nat Meth* 2012;**9**:459–62. doi:10.1038/nmeth.1974

53 Chute CG, Ullman-Cullere M, Wood GM, *et al.* Some experiences and opportunities for big data in translational research. *Genet Med* 2013;**15**:802–9. doi:10.1038/gim.2013.121

54 GATK | GATK | Guide Article #4017. https://www.broadinstitute.org/gatk/guide/article?id=4017 (accessed 26 Feb2016).

55 Clinical Genomics. http://www.hl7.org/Special/committees/clingenomics/ (accessed 15 Nov2015).

56 Alterovitz G, Warner J, Zhang P, *et al.* SMART on FHIR Genomics: Facilitating standardized clinico-genomic apps. *J Am Med Inform Assoc* Published Online First: 21 July 2015. doi:10.1093/jamia/ocv045

57 McGowan GK. The Signing of laboratory reports. *J Clin Path* 1974;**27**:427–9.

58 The Health Insurance Portability and Accountability Act of 1996.

59 <eran@hueniverse.com> EH-L. The OAuth 1.0 Protocol. https://tools.ietf.org/html/rfc5849 (accessed 29 Feb2016).

60 OAuth Web Authorization Protocol - ProQuest. http://search.proquest.com/openview/3cea88d176de039188098a5f13bed1ff/1?pq-origsite=gscholar (accessed 29 Feb2016).

61 The Direct Project | Policy Researchers & Implementers | HealthIT.gov. https://www.healthit.gov/policy-researchers-implementers/direct-project (accessed 26 Feb2016).

62 Masys DR, Jarvik GP, Abernethy NF, *et al.* Technical Desiderata for the Integration of Genomic Data into Electronic Health Records. *J Biomed Inform* 2012;**45**:419–22. doi:10.1016/j.jbi.2011.12.005

63 Sepulveda JL, Young DS. The Ideal Laboratory Information System. *Archives of Pathology & Laboratory Medicine* 2012;**137**:1129–40. doi:10.5858/arpa.2012-0362-RA

64 Blank GE, Virji MA. Development and implementation of an electronic interface for complex clinical laboratory instruments without a vendor-provided data transfer interface. *J Pathol Inform* 2011;**2**. doi:10.4103/2153-3539.77176

65 Guidi GC, Lippi G. Will 'personalized medicine' need personalized laboratory approach? *Clinica Chimica Acta* 2009;**400**:25–9. doi:10.1016/j.cca.2008.09.029

66 Overhage JM, Grannis S, McDonald CJ. A Comparison of the Completeness and Timeliness of Automated Electronic Laboratory Reporting and Spontaneous Reporting of Notifiable Conditions. *Am J Public Health* 2008;**98**:344–50. doi:10.2105/AJPH.2006.092700

67 Singh H, Thomas EJ, Mani S, *et al.* Timely Follow-Up of Abnormal Diagnostic Imaging Test Results in an Outpatient Setting: Are Electronic Medical Records Achieving Their Potential? *Arch Intern Med* 2009;**169**:1578–86. doi:10.1001/archinternmed.2009.263

68 Singh H, Vij MS. Eight Recommendations for Policies for Communicating Abnormal Test Results. *Joint Commission Journal on Quality and Patient Safety* 2010;**36**:226–32.

69 Guidi GC, Lippi G, Solero GP, *et al.* Managing transferability of laboratory data. *Clinica Chimica Acta* 2006;**374**:57–62. doi:10.1016/j.cca.2006.06.009

70 Digital Architects of Genomic Medicine. GENOSPACE. http://www.genospace.com/ (accessed 15 Nov2015).

71 Gottesman O, Kuivaniemi H, Tromp G, *et al.* The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genet Med* 2013;**15**:761–71. doi:10.1038/gim.2013.72

72 (target_somatic[PROP]) - Tests - GTR - NCBI. http://www.ncbi.nlm.nih.gov/gtr/tests/?term=(target_somatic[PROP])&test_type=Clinical (accessed 24 Feb2016).

73 FoundationOne. http://foundationone.com/order.php (accessed 25 Feb2016).

74 CMI Overview. Caris Life Sciences. http://www.carislifesciences.com/platforms/cmi-overview/ (accessed 25 Feb2016).

75 Quest Diagnostics: : Vendor Resources. http://www.questdiagnostics.com/home/companies/hit-vendors/resources.html (accessed 25 Feb2016).

76 SmartGenomics™ | Pathgroup. http://www.pathgroup.com/oncology/smartgenomics/ (accessed 25 Feb2016).

77 Medical Professionals. Guardant Health. https://www.guardanthealth.com/medical-professionals/ (accessed 25 Feb2016).

78 Oncotype DX Breast, Colon and Prostate Cancer Tests | Genomic Health, Inc. http://www.genomichealth.com/oncotype_iq_products/oncotype_dx (accessed 25 Feb2016).

79 UW Laboratory Medicine - Genetics Diagnostic Testing. http://depts.washington.edu/labweb/Divisions/MolDiag/MolDiagGen/ (accessed 25 Feb2016).

80 Test Directory | BioReference Laboratories. http://www.bioreference.com/test-directory/?type=by_test&test_id=31122 (accessed 25 Feb2016).

81 Genomic Testing in Cancer. http://gps.wustl.edu/cancer#cancer%20order (accessed 25 Feb2016).

82 Tsimberidou AM, Eggermont AMM, Schilsky RL. Precision cancer medicine: the future is now, only better. *Am Soc Clin Oncol Educ Book* 2014;:61–9. doi:10.14694/EdBook_AM.2014.34.61

83 Ronquillo JG, Li C, Lester WT. Genetic testing behavior and reporting patterns in electronic medical records for physicians trained in a primary care specialty or subspecialty. *J Am Med Inform Assoc* 2012;**19**:570–4. doi:10.1136/amiajnl-2011-000621

84 Sara Nasser, Ahmet a. Kurdolgu, Tyler Izatt, *et al.* An integrated framework for reporting clinically relevant biomarkers from paired tumor/normal genomic and transcriptomic sequencing data in support of clinical trials in personalized medicine. In: *Biocomputing 2015*. WORLD SCIENTIFIC 2014. 56–67.http://www.worldscientific.com/doi/abs/10.1142/9789814644730_0007 (accessed 7 Aug2015).

85 Rubin MA. Health: Make precision medicine work for cancer care. *Nature* 2015;**520**:290–1. doi:10.1038/520290a

86 FACT SHEET: President Obama's Precision Medicine Initiative. whitehouse.gov. https://www.whitehouse.gov/the-press-office/2015/01/30/fact-sheet-president-obama-s-precision-medicine-initiative (accessed 7 Feb2015).

87 Toussaint JS, Berry LL. The Promise of Lean in Health Care. *Mayo Clinic Proceedings* 2013;**88**:74–82. doi:10.1016/j.mayocp.2012.07.025

88 Jirjis J, Weiss JB, Giuse D, *et al.* A Framework for Clinical Communication Supporting Healthcare Delivery. *AMIA Annu Symp Proc* 2005;**2005**:375–9.

89 HIPAA Incident | BioReference Laboratories. http://www.bioreference.com/hipaa-incident/ (accessed 25 Feb2016).

90 Mandel JC, Kreda DA, Mandl KD, *et al.* SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. *J Am Med Inform Assoc* Published Online First: 17 February 2016. doi:10.1093/jamia/ocv189

91 Warner JL, Maddux SE, Hughes KS, *et al.* Development, implementation, and initial evaluation of a foundational open interoperability standard for oncology treatment planning and summarization. *J Am Med Inform Assoc* 2015;**22**:577–86. doi:10.1093/jamia/ocu015

92 open.epic :: Ancillary Systems. https://open.epic.com/Ancillary (accessed 14 Oct2015).

93 Cerner | Laboratory. http://www.cerner.com/solutions/Hospitals_and_Health_Systems/Laboratory/ (accessed 14 Oct2015).

94 Standard Interfaces - GE Healthcare. https://www2.gehealthcare.com/us/maintenance_support/maintenance_offerings/ch.standard-interfaces.octonary#COLD Feed (accessed 10 Oct2015).

95 (Chairperson) TJH, Anderson W, Aretz A, *et al.* International network of cancer genome projects. *Nature* 2010;**464**:993–8. doi:10.1038/nature08987

96 Santarpia L, Qi Y, Stemke-Hale K, *et al.* Mutation profiling identifies numerous rare drug targets and distinct mutation patterns in different clinical subtypes of breast cancers. *Breast Cancer Res Treat* 2012;**134**:333–43. doi:10.1007/s10549-012-2035-3

97 Ellis MJ, Perou CM. The Genomic Landscape of Breast Cancer as a Therapeutic Roadmap. *Cancer Discovery* 2013;**3**:27–34. doi:10.1158/2159-8290.CD-12-0462

98 Gerlinger M, Rowan AJ, Horswell S, *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med* 2012;**366**:883–92. doi:10.1056/NEJMoa1113205

99 Dancey JE, Bedard PL, Onetto N, *et al.* The Genetic Basis for Cancer Treatment Decisions. *Cell* 2012;**148**:409–20. doi:10.1016/j.cell.2012.01.014

100 Rioth M, Staggs D, Warner J. Incorporation of externally generated next-generation tumor genotyping into clinical and research workflows: Successes and lessons learned. *Journal of Clinical Oncology* 2014;**32**:abstr 156.

101 Johnson DB, Dahlman KH, Knol J, *et al.* Enabling a Genetically Informed Approach to Cancer Medicine: A Retrospective Evaluation of the Impact of Comprehensive Tumor Profiling Using a Targeted Next-Generation Sequencing Panel. *The Oncologist* 2014;**19**:616–22. doi:10.1634/theoncologist.2014-0011

102 Harris PA, Taylor R, Thielke R, *et al.* Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform* 2009;**42**:377–81. doi:10.1016/j.jbi.2008.08.010

103 Clovis Oncology. A Study to Assess the Safety and Efficacy of the VEGFR-FGFR Inhibitor, Lucitanib, Given to Patients With FGF Aberrant Metastatic Breast Cancer. In: *ClinicalTrials.gov*. Bethesda (MD): : National Library of Medicine

104 Conley BA, Doroshow JH. Molecular Analysis for Therapy Choice: NCI MATCH. *Seminars in Oncology* 2014;**41**:297–9. doi:10.1053/j.seminoncol.2014.05.002

105 NCI-Molecular Analysis for Therapy Choice (NCI-MATCH) Trial. National Cancer Institute. http://www.cancer.gov/about-cancer/treatment/clinical-trials/nci-supported/nci-match (accessed 22 Aug2015).

106 Ledford H. 'Master protocol' aims to revamp cancer trials. *Nature* 2013;**498**:146–7. doi:10.1038/498146a

APPENDIX 1

PERL Code for enterprise system receiver (from David Staggs)

```
#!/usr/local/bin/perl -w

# Rioth, Matthew John matthew.j.rioth@Vanderbilt.Edu
# Clinical Fellow, Department of Hematology and Oncology
# Vanderbilt Ingram Cancer Center
# Research Post-Doctoral Fellow
# Vanderbilt Department of Biomedical Informatics
#
# Erich R. Haberman
# Senior Interface Developer | Foundation Medicine, Inc.
# 150 Second Street, 1st Floor | Cambridge, MA 02141
# 617.418.2200 x7125 direct | 617.418.2200 main | 617.418.2201 fax
# ehaberman@foundationmedicine.com | www.foundationmedicine.com

use MIME::Base64;
use lib "/usr/star/pm";
use Star;
use Correct;
use Uniq;
use Record;
use Dates;
use Utils;
use Sys::Hostname;
use MIME::Base64;
use XML::Simple;
use Sys::Hostname;

$hostname = hostname;

%blessed = ({REDACTED}
);
print "Content-type: text/html\n\n";
$now = time;
$parser = 'foundation';
$datadir = "/home/localweb/private/data/foundation";
$uploaddir = "/home/localweb/private/data/foundation/upload";
$debugdir = "$datadir/debug";
mkdir $datadir, 0755 unless -d $datadir; # create on the fly
mkdir $datadir, 0755 unless -d $uploaddir;

if (!$blessed{$ENV{REMOTE_ADDR}}) {
 open COMPLAIN, ">> /usr/star/log/foundation.log";
```

```
 print COMPLAIN "foundation/index.cgi ERROR: $ENV{REMOTE_ADDR} is not a
blessed system!\n";
 close COMPLAIN;
 print "<HTML>Hi there, are you from Foundation?</html>";
 exit;
}

#
# Check for debugging
#
if (-e "/home/localweb/private/data/foundation/.debug") {
 $debug = 1;
 mkdir $debugdir, 0755 unless -d $debugdir;
 open DEBUG, ">>$datadir/debug/$now-$$"
 or die "Can't create $datadir/debug/$now-$$:$!\n";
}
#
# Receive and save the raw request
#
mkdir "$datadir/received", 0755 unless -d "$datadir/received";
open OUT,">$datadir/received/$now-$$"
 or die "Can't open $!";

$text = '';
while (<STDIN>) {
 $text .= $_;
}
if (length($text) <10) {
 print "<b>ERROR:</b> request has been corrupted";
 print STDERR "foundation/index.cgi ($now-$$) ERROR: request is corrupt!\n";
 exit;
}
$text = http_decode $text;
$text =~ s!\&\#(\d\d\d?);!pack("C", $1)!ge; # decode HTML decimal chars.
$text =~ s/StarPanel\=//g;
print OUT $text;
close OUT;

# Now for the uploading
$hostname = hostname;

@t = time_s();

# Parse the XML source
$xmlrec = XMLin($text);

$break = "------$t[5]/$t[4]/$t[3] $t[2]:$t[1]:$t[0]-------------------------
--------------\n";
print DEBUG $break if $debug;
```

```
print DEBUG "Incoming: $now-$$\n" if $debug;
if ($xmlrec->{ReportPDF}) {
 print DEBUG "PDF is present\n" if $debug;
 $r{adm} = '01H';
 $r{typ} = 'REP';
 $r{styp} = 'Pathology Report: Foundation One';
 $r{uniq} = make_uniq();
 $r{id} = $xmlrec->{FinalReport}->{PMI}->{MRN};
 $r{na} = $xmlrec->{FinalReport}->{PMI}->{LastName} . ', ' . $xmlrec-
>{FinalReport}->{PMI}->{FirstName};
 $xmlrec->{FinalReport}->{PMI}->{CollDate} =~ s/-/\//g;
 $r{dat} = $xmlrec->{FinalReport}->{PMI}->{CollDate};
 $r{odoc} = $xmlrec->{FinalReport}->{PMI}->{OrderingMD};
 $blockID = $xmlrec->{FinalReport}->{Sample}->{BlockId};
 print DEBUG "blockID: $xmlrec->{FinalReport}->{Sample}->{BlockId}\n";

 # r{odoc} must match what is in NPI or it fails
 # providers.txt provided by Matthew Rioth
 open IN, "/home/localweb/htdocs/cgi-bin/foundation/providers.txt";
 while (<IN>) {
 @flds = split '\|', $_;
 $name = "$flds[1], $flds[2]";
 $ProvID{$name} = $flds[0];
 }

 $srcname = uc($r{odoc});
 if ($ProvID{$srcname}) {
 print DEBUG "I found $string with code $ProvID{$srcname}\n" if $debug;
 $r{rno} = $ProvID{$srcname};
 } else {
 print DEBUG "No match for $srcname\n" if $debug;
 }

 $xmlrec->{ReportPDF} =~ s/ /\+/g;
 if ($decode = decode_base64($xmlrec->{ReportPDF})) {
 open OF, ">/tmp/$now-$$.pdf";
 binmode OF;
 print OF $decode;
 close OF;
 }
 $r{pq} = "$parser $now-$$.pdf";
 $string = record2string(\%r);
 # Checks
 $result = correct_pt_info \%r;
 print DEBUG "EPI:$result\n" if $debug;
 if ($result) {
 print DEBUG "EPI check out\n" if $debug;
 } else {
```

```
 $errors .= "EPI Mismatch ($r{id}/$r{na})\n";
 }
 $errors .= "No Patient ID Found\n" if !$r{id};
 $errors .= "No Patient Name Found\n" if !$r{na};
 $errors .= "No Date of Services Found ($r{dat})\n" if !$r{dat};
 $errors .= "No Reffering Physcian Found ($r{odoc})\n" if !$r{odoc};
 $errors .= "No Document Title Found ($r{styp})\n" if !$r{styp};
 print DEBUG "$string\n" if $debug;

 if ($errors) { # We failed
 print DEBUG "REJECT:$string\n" if $debug;
 print DEBUG $errors if $debug;
 open sm_MAILTO, "| /usr/ucb/mail -s \"Error:$r{pq} (web:$parser)\"
dstaggs\@vumclib.mc.Vanderbilt.Edu matthew.j.rioth\@Vanderbilt.Edu";
 print sm_MAILTO "Name:$r{na}\n";
 print sm_MAILTO $errors;
 print sm_MAILTO "\n\n**********************************************\n".
   "-----PLEASE DO NOT RESPOND TO THIS MESSAGE-----\n".
   "**********************************************\n";
 close sm_MAILTO;

 # Record what we did in the parser rejection log.
 open TRANS_LOG, ">> /home/localweb/private/data/$parser/.reject";
 my $stamp = "$t[5]$t[4]$t[3] $t[2]$t[1]";
 print TRANS_LOG $stamp;
 print DEBUG $stamp if $debug;
 foreach (qw (uniq typ styp id na dat tim odoc pq )) {
 print TRANS_LOG '|', $r{$_} || '';
 print DEBUG '|', $r{$_} || '' if $debug;
 }
 print TRANS_LOG '|', $blockID || '';
 print DEBUG '|', $blockID || '' if $debug;
 print TRANS_LOG "\n";
 print DEBUG "\n" if $debug;
 close TRANS_LOG;

 print "ERROR:$hostname:$errors:ERROR\n"; # Tell source of error
 print DEBUG "Sending to client:ERROR:$hostname:$now-$$:$errors:ERROR\n" if
$debug;

 print DEBUG "Source File: $now-$$\n" if $debug;
 } else { # We are a go
 print DEBUG "We are a go!\n" if $debug;
 print DEBUG "Qsending: /usr/star/par/$parser/curr/$t[5]/$t[4]/$t[3]/$now-
$$.pdf|$now-$$.pdf\n" if $debug;
   #dest      # source
 queue_file ("/usr/star/par/$parser/curr/$t[5]/$t[4]/$t[3]/$now-$$.pdf",
"/tmp/$now-$$.pdf",get_words ('archive-image'));
```

```
 $genes = $xmlrec->{FinalReport}->{Genes}->{Gene};
 $FM_ID = $xmlrec->{FinalReport}->{Sample}->{FM_Id};
 $reptxt .= "<B>Sample:</B> ".
  "<B>BlockID:</B> $xmlrec->{FinalReport}->{Sample}->{BlockId}<BR>".
  "<B>Collection Date:</B> $xmlrec->{FinalReport}->{PMI}->{CollDate}<BR>".
  "<B>Received Date:</B> $xmlrec->{FinalReport}->{Sample}-
>{ReceivedDate}<BR>".
  "<B>Specimen Format:</B> $xmlrec->{FinalReport}->{Sample}->{SpecFormat} ".
  "<B>FM_Id:</B> $xmlrec->{FinalReport}->{Sample}->{FM_Id}<BR>";

 $reptxt .= "<B>Specimen Site:</B> $xmlrec->{FinalReport}->{PMI}->{SpecSite}
";
 $reptxt .= "<B>Submitted Diagnosis:</B> $xmlrec->{FinalReport}->{PMI}-
>{SubmittedDiagnosis}<BR>";

 $reptxt .= "<B>ReportID:</B> $xmlrec->{FinalReport}->{ReportId}<BR><BR>";

 $reptxt .= "<B>Genes:</B><BR>";
 $reptxt .= "<TABLE BORDER COLS=4>".
  "<TR BGCOLOR=yellow><TH>Gene</TH>".
  "<TH>Alteration</TH>";
 foreach $g (@{$genes}) {
 $reptxt .= "<TR><TH BGCOLOR=cyan>${$g}{Name}</TH>";
 $reptxt .= "<TH>${$g}{Alterations}->{Alteration}->{Name}</TH>";
 }
 $reptxt .= "</TABLE><BR>\n";

 $reptxt .= "<B>Signature:</B> $xmlrec->{FinalReport}->{Signatures}-
>{Signature}->{OpName}".
  " <B>/</B> $xmlrec->{FinalReport}->{Signatures}->{Signature}-
>{ServerTime}";

 $r{hp} = "Foundation One

$reptxt

The table above is of 'actionable' mutations as defined by Foundation
Medicine. To view additional information, potential clinical trials, as
well as 'variants of unknown significance' please open the PDF of the
full report below.

Clicking on the link will start up the Adobe Acrobat Reader in a separate
window.

<a href='archive.cgi?$parser:$t[5]/$t[4]/$t[3]/$now-$$.pdf'>

 Click to view the report</a>
";
 $string = record2string(\%r);
```

```
 $r{na} = uc($r{na});
 print DEBUG "UPLOAD:$string\n" if $debug;
 send2star(\%r, 'vumc', 1, 1);

 open TRANS_LOG, ">> /home/localweb/private/data/$parser/.log" or die "Cannot
create .log file \n";
 my $stamp = "$t[5]$t[4]$t[3] $t[2]$t[1]";
 print TRANS_LOG $stamp;
 print DEBUG $stamp if $debug;
 foreach (qw (uniq typ styp id na dat tim odoc pq)) {
 print TRANS_LOG '|', $r{$_} || '';
 print DEBUG '|', $r{$_} || '' if $debug;
 }
 print TRANS_LOG '|', $blockID || '';
 print DEBUG '|', $blockID || '' if $debug;
 print TRANS_LOG "|$FM_ID\n";
 print DEBUG "|$FM_ID\n" if $debug;
 close TRANS_LOG;
 }
 open OF, ">$uploaddir/$now-$$.xml";
 print OF $text;
 close OF;

} else {
 print DEBUG "ERROR:Where is the PDF?\n" if $debug;
}
print DEBUG $break if $debug;
close DEBUG;
```

APPENDIX 2

PERL code for file parsers

```
#!/usr/bin/perl

use XML::Simple;

#use warnings;

opendir(DIR, ".") or die "cannot open directory";
open (ALLFILE, ">>", "demogr_metadata.csv") or die $!;
printf ALLFILE
"file,FMI_ID,MRN,Name,DOB,Received_date,path_ID,Histology,Sex,OrderingMD,Spec
_Date,Spec_Site\n";
open (VARFILE, ">>", "var_data.csv") or die $!;
@docs = <*>;
foreach $file (@docs) {
 open (REPORT, $file) or die "could not open $file\n";
 $file =~ s{\.[^.]+$}{};
 next if (/\.pl$/|/\.csv$/);
 next if (exists($duplicates{$file}));
 printf "%s\n", $file;#to terminal
 #open (PRINTFILE, ">$file.csv");#opens output file named same as input file
 $xmlrec = XMLin($file, ForceArray=>['Gene']);
 $text = $file;
 $trf = $xmlrec->{FinalReport}->{Sample}->{FM_Id};
 $received_date = $xmlrec->{FinalReport}->{Sample}->{ReceivedDate};
 $sample_type= $xmlrec->{FinalReport}->{Sample}->{SpecFormat};
 $path_id = $xmlrec->{FinalReport}->{Sample}->{BlockId};
 $mrn = $xmlrec->{FinalReport}->{PMI}->{MRN};
 $pt_lastname = $xmlrec->{FinalReport}->{PMI}->{LastName};
 $dx = $xmlrec->{FinalReport}->{PMI}->{SubmittedDiagnosis};
 $dx =~ s/,//s;
 $gender = $xmlrec->{FinalReport}->{PMI}->{Gender};
 $ordering_md = $xmlrec->{FinalReport}->{PMI}->{OrderingMD};
 $ordering_md =~ s/,.*//s; #removes MD first name
 $specimen_site = $xmlrec->{FinalReport}->{PMI}->{SpecSite};
 $specimen_site =~ s/,//s;
 $specimen_date = $xmlrec->{FinalReport}->{PMI}->{CollDate};
 $dob = $xmlrec->{FinalReport}->{PMI}->{DOB};
 $genes = $xmlrec->{FinalReport}->{Genes}->{Gene};
 #printf "%s\n", scalar $genes;
 $text .=
",".$trf.",".$mrn.",".$pt_lastname.",".$dob.",".$received_date.",".$path_id."
,".$dx.",".$gender.",".$ordering_md.",".$specimen_date.",".$specimen_site;
 printf ALLFILE "%s\n",$text;
 foreach $gene (@{$genes}){
  $name = $gene->{Name};
  $alteration = $gene->{Alterations}->{Alteration}->{Name};
```

```
  if (index($alteration, 'exon')==-1) {
  @alterations = split(/,/, $alteration);
  }
  $alteration = {FinalReport}->{Genes}->{Gene}->{Alterations}->{Alteration}-
>{Name};
  foreach $variant (@alterations){
  $variant =~ s/^\s+//; #removes leading white spaces
  printf VARFILE "%s,%s,%s,%s,%s\n", $file, $trf, $dx, $name, $variant;
  printf PRINTFILE "%s,%s,%s,%s,%s\n", $file, $trf, $dx, $name, $variant;
  }

 }
}

opendir(DIR, ".") or die "cannot open directory";
open ALLFILE, ">>", "allvus.csv" or die $!;
@docs = grep(/\.txt$/|/\.p2t$/,readdir(DIR));#only open text or pdftotext
files

foreach $file (@docs) {
 open (REPORT, $file) or die "could not open $file\n";
 $file=~ s{\.[^.]+$}{};
 next if (exists($duplicates{$file}));
 printf "%s\n\n", $file;#to terminal
 open (PRINTFILE, ">~/csvs/$file.csv");#opens output file named same as input
file
 while(<REPORT>){

 if(/meaningful in the future./../Electronically/){#regex to find VUS block
  next if /^\n/;#skips blank lines
  next if /the future./||/Electronically/;#skips anchors
  next if /^For more comprehensive information/||/^For additional
information/;
  $_ =~ s/\s/,/;#adds commas to variant files
  @tokens = split(/,/, $_);#splits off multiple variants from the same gene
  my $gene = shift@tokens;#first split is gene name
  my $last = pop@tokens;#removes last for csv formatting
  foreach my $token (@tokens){
  printf PRINTFILE "%s,%s,%s\n",$file, $gene, $token;
  printf ALLFILE "%s,%s,%s\n",$file, $gene, $token;
  }
  printf PRINTFILE "%s,%s,%s",$file, $gene, $last;
  printf ALLFILE "%s,%s,%s",$file, $gene, $last;
 }
 }
 close $file;
 close PRINTFILE;
}
```

APPENDIX 3

PERL code for error log parser

```
#!/usr/bin/perl

opendir(DIR, ".") or die "cannot open directory";
@docs = readdir DIR;

open ALLFILE, ">", "errorlog.csv" or die $!;
printf ALLFILE "file,error,trf,mrn,received,in-chart,epi,provider\n";
open TEXTFILE, ">>", "all_log.txt";

#print @docs;

foreach $file (@docs) {
 next if ($file =~ m/^\./);
 open (REPORT, $file) or die "could not open $file\n";

 $file=~ s{\.[^.]+$}{};
 printf "%s\n", $file;#to terminal
 printf ALLFILE "$file,";
 while(<REPORT>){
 if(m/We are a go/){
  printf ALLFILE "no error";
  $error=1};
 if (m/------(.*):/){$rec_date=$1};
 $ord_date=/DAT:201{4|5}\/(\d\d\/\d\d)/im;
 if(m/ERROR:star40:$file:(.*) /){$reason=$1};
 if (/EPI Mismatch (.*)/im){$epi=$1};
 if (/No match for (.*)/im){$pname=$1};
 if (/TRF(\d*)/im){$trf=$1};
 if (/:ID:(\d*)/im){$mrn=$1};
 if (/'actionable'/){$chart=1};
 #printf TEXTFILE "$_"; #compiles all files into one
 }
 printf ALLFILE ",TRF%s,%s,%s,%s,%s,%s,%s\n",
$trf,$mrn,$rec_date,$chart,$reason,$epi,$pname;
 $reason=$epi=$pname=$trf=$mrn=$chart="";
 close $file;
 }
 close ALLFILE;
```