Bioinformatics Advances
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: day month year
Original Paper

OXFORD

Data and Text Mining

# A cloud-based pipeline for analysis of FHIR and long-read data

**Tim Dunn [1,*], Erdal Cosgun [2]**

[1]Computer Science and Engineering, University of Michigan, Ann Arbor, MI 48109, USA and
[2]Biomedical Platforms and Genomics, Microsoft Research, Redmond, WA 98052, USA.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

## Abstract

**Motivation:** As genome sequencing becomes cheaper and more accurate, it is becoming increasingly viable to merge this data with electronic health information to inform clinical decisions.
**Results:** In this work we demonstrate a full pipeline for working with both PacBio sequencing data and clinical FHIR® data, from initial data to tertiary analysis. The electronic health records are stored in FHIR® – Fast Healthcare Interoperability Resource – format, the current leading standard for health care data exchange. For the genomic data, we perform variant calling on long read PacBio HiFi data using Cromwell on Azure. Both data formats are parsed, processed, and merged in a single scalable pipeline which securely performs tertiary analyses using cloud-based Jupyter notebooks. We include three example applications: exporting patient information to a database, clustering patients, and performing a simple pharmacogenomic study.
**Contact:** timdunn@umich.edu
**Supplementary information:** Supplementary data are available at *Bioinformatics Advances* online, and at `https://github.com/microsoft/genomicsnotebook/tree/main/fhirgenomics`.

## 1 Introduction

When visiting the doctor's office for an annual physical, it's typical to have your vitals taken. Weight, blood pressure, and heart rate are all important measurements that can indicate impending health issues. While waiting in the lobby, it's also common to fill out a short survey regarding sleep, smoking, drug, and other lifestyle habits that may impact your health. Some doctors even recommend blood work – laboratory testing which measures cell counts and micronutrient levels – the results of which could indicate other less visible issues. This smattering of multi-modal information is then used by doctors to make informed decisions about lifestyle and medication changes that may improve your overall health. The more information a trained medical professional has available, the better recommendations they can make towards improving their patient's health.

Soon, genomics data may routinely be used to complement this clinical data. The first "complete" human genome was finished in 2000 at an estimated cost of $300,000,000 [19]. Due to rapid improvements in sequencing technologies, however, this cost has sharply declined over the past two decades. Currently, whole genome sequencing costs around $700 per patient [19] and is usually reserved only for those suffering from cancer or rare genetic diseases. In just a few years, however, the cost will likely be low enough for routine sequencing of ordinary patients.

Not only can genome sequencing lead to earlier and more accurate genetic disease and cancer diagnoses, but it can also be used to predict individualized responses to medications and characterize the body's internal micro-organisms and pathogens. For example, sequencing has been widely used to identify exact SARS-CoV-2 strains [36] and analyzing the gut microbiome can lead to insights regarding overall well-being [26]. Once sequencing costs have lowered, it will be possible to integrate genomic data with existing clinical data to provide a more comprehensive view of each and every patient. This data will, in turn, lead to a better understanding of patient health and disease – particularly with the help of machine learning.

Machine learning has caused immense scientific progress in recent years when applied to new domains such as text recognition, protein folding, and nanopore sequencing basecalling [48, 27, 40]. Unfortunately, there are a number of legal, practical, and ethical concerns preventing the immediate use of machine learning for diagnosing patients [15, 8, 28]. In the inevitable case of false positives and false negatives, how can

 

```
'resource': {'resourceType': 'Patient',
    'id': 'b5f1da11-3826-4821-bb84-dd72294c9a4c',
    'meta': {'versionId': '1',
     'lastUpdated': '2022-07-13T19:23:23.981+00:00',
     'profile': ['http://hl7.org/fhir/us/core/StructureDefinition/us-core-patient']},
    'text': {'status': 'generated',
     'div': '<div xmlns="http://www.w3.org/1999/xhtml">Generated by Synthea</div>'},
    'extension': [{'extension': [{'url': 'ombCategory',
        'valueCoding': {'system': 'urn:oid:2.16.840.1.113883.6.238',
         'code': '2106-3',
         'display': 'White'}},
       {'url': 'text', 'valueString': 'White'}],
      'url': 'http://hl7.org/fhir/us/core/StructureDefinition/us-core-race'},
     {'extension': [{'url': 'ombCategory',
        'valueCoding': {'system': 'urn:oid:2.16.840.1.113883.6.238',
         'code': '2135-2',
         'display': 'Hispanic or Latino'}},
       {'url': 'text', 'valueString': 'Hispanic or Latino'}],
      'url': 'http://hl7.org/fhir/us/core/StructureDefinition/us-core-ethnicity'},
     {'url': 'http://hl7.org/fhir/StructureDefinition/patient-mothersMaidenName',
      'valueString': 'Julia241 Luna60'},
     {'url': 'http://hl7.org/fhir/us/core/StructureDefinition/us-core-birthsex',
      'valueCode': 'M'},
     {'url': 'http://hl7.org/fhir/StructureDefinition/patient-birthPlace',
      'valueAddress': {'city': 'Portsmouth',
       'state': 'Saint John Parish',
       'country': 'DM'}},
       ...
```


**Fig. 1.** Example synthetic FHIR® data, generated with Synthea.


```
##fileformat=VCFv4.3
##fileDate=20210115
##source=VariantCallerScript
##reference=file:///references/HG38.fasta
##contig=<ID=20,length=62435964,assembly=B36,species="Homo sapiens">
##phasing=partial
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID  REF ALT    QUAL FILTER INFO              FORMAT   NA00001        NA00002
20     14370   rs6 G   A      29   PASS   DP=14;DB;         GT:DP:HQ 0|0:1:51,51    1|0:8:51,51
20     17330   .   T   A      3    q10    DP=11;AF=0.017    GT:DP:HQ 0|0:3:58,50    0|1:5:65,3
20     1110696 rs5 A   G,T    67   PASS   DP=10;AA=T;DB     GT:DP:HQ 1|2:6:23,27    2|1:0:18,2
20     1230237     .   T  .    47   PASS   DP=13;AA=T        GT:DP:HQ 0|0:7:56,60    0|0:4:51,51
20     1234567 ms1 GTC G,GTCT 50   PASS   DP=9;AA=G         GT:DP    0/1:4          0/2:2
```


**Fig. 2.** Example VCF data, including both the file header and data. The first eight tab-separated fields refer to a specific variant, and the remaining fields store information about that variant as it pertains to each sample.

we perform root cause analysis or ensure that the same mis-diagnoses won't happen again? In many cases, we can't. Machine learning can be used, however, to find correlations between genetic alterations and clinical observations, which can be used to guide further scientific research. Used properly, machine learning can be a tool for discovery, accelerating our progress in understanding genetic diseases and even lead to advances in medication and gene therapy. In this work, we present a scalable and secure proof-of-concept pipeline for combining clinical and genomic data in the cloud, and demonstrate several possible use cases.

## 1.1 Clinical Data

### FHIR®: Fast Healthcare Interoperability Resource Format

Clinical data can come in the form of numbers, raw text, images, or even 3D scans. Despite the inherent diversity of this data, it must be stored in a consistent digital format that allows for easy and efficient exchange between hospitals, laboratories, and data centers. The "Fast Healthcare Interoperability Resource" (FHIR®) format is the current leading standard for health care data exchange [5]. Each chunk of FHIR® data is an instance of one of 140 pre-defined resources, represented in XML, JSON, or RDF format. The framework was designed to be broad and extensible, covering clinical healthcare, clinical trials, organization administration, and finances. Data is commonly hosted on a secure server and accessed from a FHIR® RESTful API that ensures secure and efficient querying of patient data.

### Synthea

Synthea is a widely-used open source tool for generating realistic (but synthetic) patient data in FHIR® format [47]. This enables researchers to work with realistic clinical datasets without worrying about any of the legal, ethical, or security concerns that would accompany working with real patient data. Figure 1 shows a snippet of realistic patient data generated with Synthea.

## 1.2 Genomics Data

### Sequencing Technologies

Illumina short read sequencing remains the dominant technology for genome sequencing today. For years it has successfully provided relatively low cost and massively-parallel short read sequencing. Recently, however, newer long read technologies have begun to prove competitive. Illumina short reads of several hundred bases typically achieve around 99.9% accuracy [16]. The newer and less mature long read sequencing technologies such as Pacific Biosciences SMRT [42] and Oxford Nanopore [25] haven't achieved similar accuracy results until very recently. These newer technologies can easily achieve average read lengths of over 10,000 bases [25]. This greatly aids in assembling the human genome in complex or repetitive genomic regions, and PacBio HiFi reads were instrumental in the "Telomere-to-Telomere" consortium completing the first truly complete human genome in 2021 [35].

### Variant Calling

Since two human genomes are 99.9% identical [20], the end goal of most DNA sequencing efforts is to identify the differences between an individual's DNA and a standard reference sequence. This problem is known as "variant calling". These small changes in DNA can be in the form of single nucleotide polymorphisms (abbreviated SNPs, `A`→`G`), insertions (`A`→`ATT`), deletions (`AGC`→`A`), or structural variants (in which large segments of DNA are inserted or deleted). In this work, we focus on small variants, which include SNPs and insertions or deletions shorter than 50 bases. These variants are stored in "Variant Call Format" (VCF), which notes the "reference" and actual ("alternate") observed DNA sequence. Figure 2 shows an example. Databases of known mutations and their functional consequences (if any) on patient health are used to identify important mutations [17].

## 1.3 Cloud Frameworks

### FHIR® Integration

All of the major cloud providers have their own implementation of a FHIR server that can readily be used with other cloud services. Amazon supports "FHIR Works" on Amazon Web Services (AWS), Microsoft supports an "Azure API for FHIR" on Azure, Google supports a "Cloud Healthcare API" on Google Cloud, and IBM supports an "IBM FHIR Server" on the IBM Cloud [29, 33, 10, 18]. There are numerous other open and closed source standalone FHIR server implementations as well, such as HAPI FHIR [2]. For our purposes, any of the FHIR server implementations that integrate easily with Cromwell and a major cloud provider would work. We selected Microsoft Azure and the "Azure API for FHIR" simply because we had reduced-cost access to Azure Cloud computing resources.

### Bioinformatics using Cromwell

Cromwell is an open-source workflow management system designed by the Broad Institute for performing bioinformatics at scale [21]. Cromwell can be configured to run with a Google Cloud backend through the Google Genomics Pipelines API, an AWS backend using AWS Batch, or an Azure backend using the "Cromwell on Azure" project [22, 31]. As mentioned in the previous section, we selected the Microsoft Azure backend . PacBio has released a public workflow for running a human whole genome sequencing (WGS) pipeline using "Cromwell on Azure" [6]. It begins with unaligned or aligned reads (in FASTQ or BAM format, respectively), and determines large scale structural variants using pbsv [49] and hifiasm [9]. Reads are

phased using WhatsHap [37] and small variants are called in VCF format with DeepVariant [44].

### 1.4 Related Work

Previous works in merging clinical and genomic information have primarily focused on extending the FHIR® implementation to include genomic data. Earlier efforts such as "SMART on FHIR Genomics" were influential in envisioning the design of such a standard [3]. A more recent project by the Electronic Medical Records and Genomics (eMERGE) Network developed a new standard format based on HL7® FHIR® to represent clinical genomics results [34].

There has been significant progress within HL7® to adopt a standard format for sharing genomic data as well. In particular, HL7® has organized a Clinical Genomics Work Group to tackle this effort [23]. The current version of the HL7® FHIR® specification includes a "Genomics Implementation Guidance" page, which is currently in the "Trial Use" phase of development, with a Maturity Level of only 1 (on a scale from 0 to 5). Over the past several years, the HL7's recommendations for storing genomic data have shifted and evolved rapidly due to fast-paced technological advancements and learnings from practical experience. Regardless of whether the current standard is to store data in an `Observation`, `Sequence`, `Observation-genetics`, `MolecularSequence`, or `Variant` resource, the format is not yet mature and there are no guarantees of long-term stability with regards to the current format. To avoid this problem, we convert the contents of stable FHIR® resource implementations to a tabular format prior to merging with genetic information.

Although several previous works explore merging clinical and genomic patient data, they primarily focus on the release of large datasets of cancer patients. Project GENIE is the largest of such efforts, resulting in more than 100,000 sequenced patients from cancer centers worldwide [11]. Despite the availability of merged clinical and genomic data – at least for cancer patients – there is no publicly available standard pipeline for merging the two data modalities. This work presents such a pipeline for secure and scalable merging of clinical and genomic data using cloud resources.

## 2 Methods

Figure 3 shows an overview of our data processing pipeline, which was shared for all three of our example applications. This pipeline processes the PacBio and FHIR® data prior to merging them into a single DataFrame for further analyses. In the following two sections, we first describe our treatment of the PacBio genomics data, and then our methods for dealing with the synthetic FHIR® data.

### 2.1 Genomics Data

For our genomics data, the first step was variant calling using Cromwell on Azure. A deployment script available from Microsoft's Cromwell on Azure repository [31] was first downloaded and executed to initialize the workflow environment. The `GCA_000001405.15_GRCh38_no_alt_analysis_set.fasta` reference genome was downloaded from the NCBI [39] database, and a BED file containing tandem repeat regions to exclude was downloaded from the pbsv repository [49]. In order to demonstrate the Cromwell workflow, we used the GIAB ("Genome in a Bottle") PacBio sequencing data for human sample HG002 [50]. We use sample VCF data from a larger dataset for all downstream processing. Additionally, we made several modifications to PacBio's default human WGS workflow configuration [6]. These changes consist of several bugfixes, which have now been merged into their official repository, and removing the tandem-genotypes step for simplicity.

Once the variants were called using Cromwell on Azure, we used BCFTools [12] to normalize the variant representation and split sites with multiple alleles. Normalizing involves left-aligning insertions and deletions (INDELs) and splitting multi-nucleotide polymorphisms (MNPs) in to single-nucleotide polymorphisms (SNPs) [12]. Splitting multi-allelic sites was necessary because in order to convert the VCF to TSV format, there must be a constant number of fields (columns) per entry (genomic position). By splitting entries with multiple possible variants into multiple entries – each with a single variant – we ensured that fields such as allele frequency are represeneted with a fixed number of columns.

Next, we performed linkage disequilibrium (LD) pruning to remove variants calls with high covariance, and ensure that most remaining variants have a fairly high degree of independence from one another. At this point, we used BCFTools to merge VCF files from all patients into a single VCF file, and extracted select fields into a TSV file. Since our input files were VCF, and not GVCF, they did not contain information regarding the quality of reference calls for non-variant positions. To deal with this, we assumed all missing entries to be a reference call of average quality and depth for that sample, and imputed the genotype and phred likelihoods under this assumption. As a final step, the TSV was loaded into a Pandas DataFrame and transposed so that each row corresponds to a patient, and the columns are the relevant genomic information: the genotype, allele frequency, depth, and phred likelihood scores for a select group of variants.

### 2.2 Clinical Data

Using Synthea, we generated a synthetic dataset with the same number of patients as included in our sample VCF files. For our data exportation and patient clustering applications, we simply used Synthea with the default parameters to generate a typical patient population. For the breast cancer study, however, we added a custom module to model the simplified progression of breast cancer in a cohort of affected patients. This module is shown in Figure 4.

Combined with the Synthea flags `-g F -a 30-90`, this module generates an exclusively adult female dataset of breast cancer patients, who are treated with one of two medications: Doxorubicin or Epirubicin. Patient survival rates, assumed to be 50% if left untreated, depend on a combination of medication and presence of a specific variant. For patients without the variant, survival rates can be improved to 90% by prescribing Doxorubicin. For patients with the variant, the same survival rate is instead achieved by prescribing Epirubicin. With all other groups, the survival rate remains unchanged at 50%. This relationship which we've embedded into our synthetic clinical data generation module should be discoverable through downstream analyses, provided our patient cohort is large enough.

In a real clinical setting, patient records will be stored in FHIR® format on a server, where they can be queried by clinicians. To model this setup, we deploy a FHIR® server using the Azure API for FHIR, and transfer our generated Synthea data to the server using the REST API. Using the FHIR bulk import and export functionality would be a viable alternative (and a potential direction for future work). We found that performing individual requests allowed us to transfer fewer FHIR resources in total, and achieved sufficient throughput.

In order to perform data science applications with FHIR® data, we first need to convert the hierarchical JSON data into tabular format. An existing open source tool called the "FHIR to Synapse Sync Agent" solves this problem by downloading each FHIR® resource from a server and converting it to tabular Parquet format [32]. Parquet files store the same information as ordinary CSV ("comma-separated values") or TSV ("tab-separated values") files, but in a more efficient compressed manner. Data is stored in column-major order, which allows compression algorithms such as run length encoding, dictionary encoding, or delta encoding to be applied per column depending upon each column's data format and

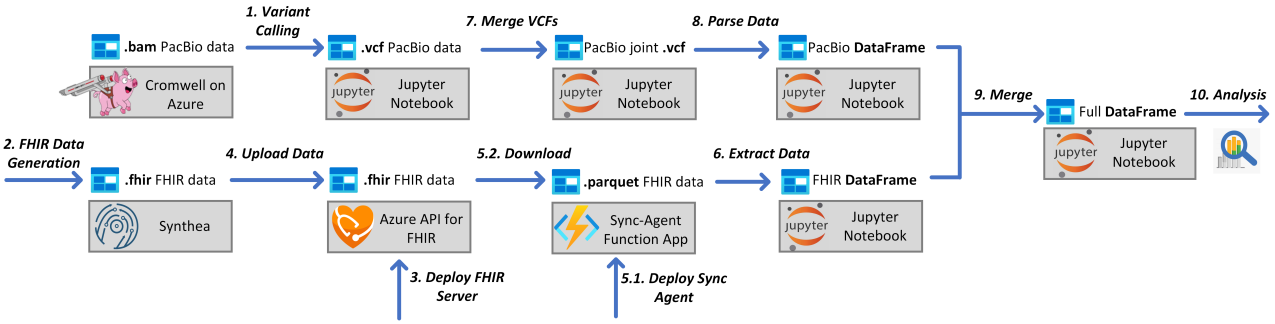


**Fig. 3.** Overview of data processing pipeline, common to all demonstrated applications.

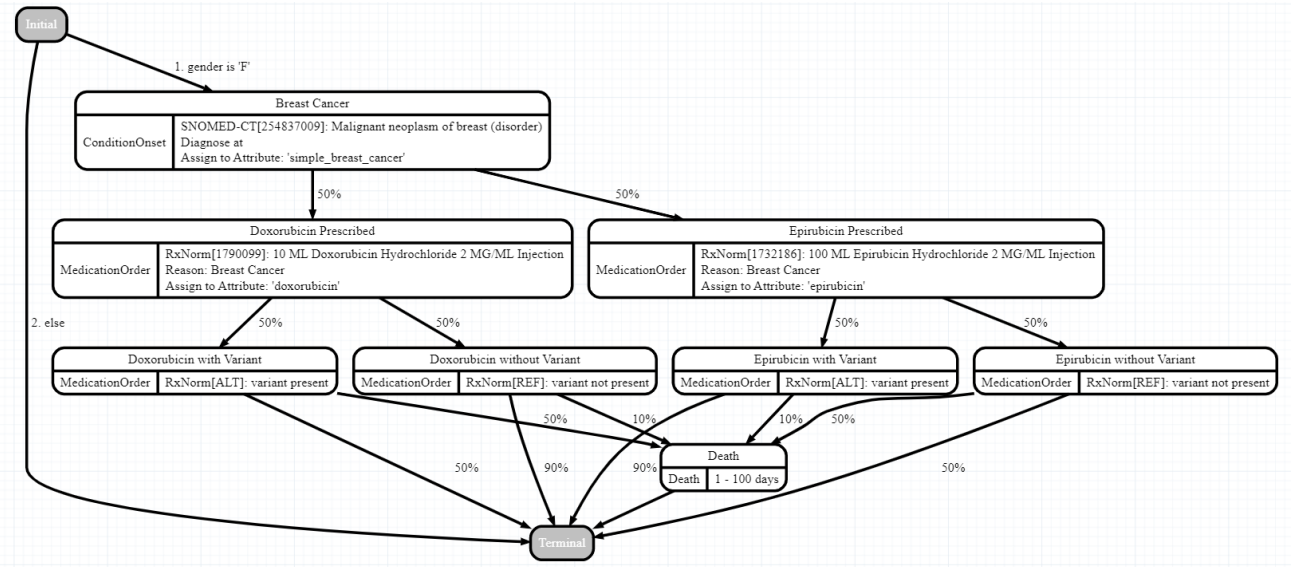


**Fig. 4.** Custom Synthea module for modelling breast cancer treatments and outcomes.

values [24]. It is important to note that although Parquet is a tabular format, it is still able to store unstructured or nested fields present in the original FHIR® resource by encoding them as JSON strings.

Once downloaded, we parse the Parquet FHIR® data to retrieve relevant information and load it into a Pandas DataFrame [41]. Although Parquet is already a tabular format, the converted data contains extraneous information which must be discarded, and any useful information stored within a JSON string must be extracted. Making matters more difficult, FHIR® data pertaining to a single patient is stored in multiple resource types. These files (such as Patient, Medication, or Condition resources) must be parsed separately and the records associated with one another using the patient's unique Medical Record Number (MRN). We are able to map information from multiple resources of the same type to a single patient by adding another column to the DataFrame for each resource instance. In the end, each row stores data for a single patient, and the numerous columns contain all desired information about that patient, extracted from the Parquet files. At this point the FHIR® and PacBio data can be merged together, as shown in Figure 5.

## 3 Results

After merging these two data modalities, we explore three example use cases for such data: exporting the information to a database, clustering patients, and studying a cohort of breast cancer patients.

**Application #1: Export to Database**

The first such application that we explored was the simplest: exporting the FHIR® and PacBio data to a database for further analyses. We selected Azure Synapse Analytics as our database platform because it ensures scalability and reliability. After converting our FHIR® and PacBio DataFrames to Parquet format, we directly import them into Azure Synapse. From there, we can perform joint queries on the two datasets. This application is demonstrated in the supplementary notebook `1-data-export.ipynb`.

**Application #2: Patient Clustering**

The second application we demonstrated was an exploratory clustering of our patient dataset using three different clustering methods. Firstly, K-means++: an iterative approach which updates cluster centroid locations to minimize total inertia and uses repeated random initializations [4]. Inertia is the average squared distance from each data point to the center of its labelled cluster. Secondly, we used DBSCAN: a recursive approach that builds clusters from high-density areas containing "core samples" [14]. Lastly, Spectral Clustering, which performs a low-dimensional embedding of the samples' affinity matrix prior to clustering [45].

Both K-means++ and Spectral Clustering require specifying the final number of clusters a priori. To select a reasonable number of clusters we used the "elbow method", which consists of plotting the number of clusters versus total inertia. As the number of clusters increases, the cluster inertia will always decrease, but once the data has already been clustered fairly well, there will be less benefit in introducing additional clusters. This

| | city | state | country | gender | dead | age | newest_med_code | ... | 49:GT | 49:AF_0 | 49:AF_1 | 49:PL_0/0 | 49:PL_0/1 | 49:PL_1/1 | 49:DP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | Boston | MA | US | female | False | 22.009037 | 748856 | ... | 0/0 | 1 | 0 | 0 | 8 | 8 | 7 |
| **1** | Amesbury | MA | US | male | False | 64.588092 | 310798 | ... | 0/0 | 1 | 0 | 0 | 3 | 3 | 7 |
| **2** | Boston | MA | US | female | False | 17.316361 | 1367439 | ... | 0/1 | 0.285714 | 0.714286 | 29 | 0 | 23 | 7 |
| **3** | Yarmouth | MA | US | male | True | 69.918688 | 896209 | ... | 0/0 | 1 | 0 | 0 | 4 | 4 | 5 |
| **4** | Attleboro | MA | US | female | False | 25.182206 | 751905 | ... | 0/0 | 1 | 0 | 0 | 9 | 9 | 15 |

**Fig. 5.** Final Pandas DataFrame of merged FHIR® and PacBio data.

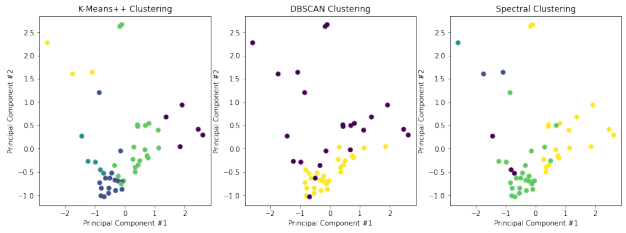


**Fig. 6.** Patient clusters resulting from several different clustering algorithms.

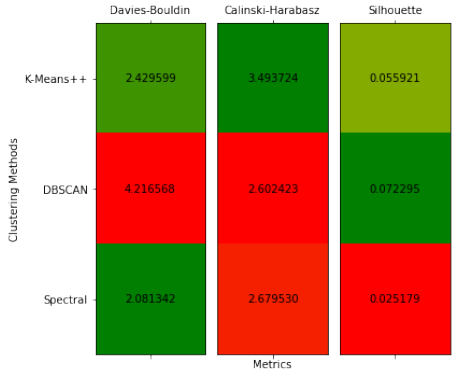


**Fig. 7.** Evaluation of clustering methods using several different metrics.

shows up on an "elbow plot" as a bend towards the horizontal, which we found occurs at $n = 5$. After selecting $n$ for K-means++ and Spectral Clustering, Figure 6 shows the resulting clusterings. Each point represents a single patient, and each cluster of patients is colored uniformly.

We evaluated these three clustering methods using the Davies-Bouldin Index [13], the Calinski-Harabsz Index [7], and the Silhouette Coefficient [43]. Unlike the other two metrics, for the Davies-Bouldin Index a lower score is better. These three evaluation metrics were selected because they all do not require knowledge of some ground truth clustering of samples into classes. Since we are simply trying to discover similarities between patients, no ground truth is available. As Figure 7 shows, K-Means++ selected the best clustering overall. This application and evaluation is included in the supplementary notebook `2-clustering.ipynb`.

**Application #3: Breast Cancer Study**

The final application we demonstrated was a basic pharmacogenomic study to determine the effect of two different medication treatments on breast cancer patients' survival rates. As described in Section 2.2, a custom Synthea module was designed which modelled the fact that outcomes were only improved beyond the base 50% survival rate for patients who were prescribed Epirubicin and had a particular genetic variant, and for patients who were prescribed Doxorubicin and did not have the genetic variant. Figure 8 shows a breakdown of patient variant presence, treatment type, and survival.

We performed a one-sided z-test with a p-value of 0.01 to investigate whether patients in each of the four groups (all combinations of with/without variant and Doxorubicin/Epirubicin) had improved survival rates over the expected outcome for untreated patients (a 50% survival rate). As expected, we found a significant increase in survival rates for patients without the variant who were prescribed Doxorubicin ($p < 10^{-8}$), and for patients with the variant who were prescribed Epirubicin ($p < 10^{-5}$). For patients without the variant who were prescribed Epirubicin ($p = 0.109$) and patients with the variant who were prescribed Doxorubicin ($p = 0.5932$), we did not find a significant improvement in outcomes, as expected. This application is included in the supplementary notebook `3-pharmacogenomics-confidential.ipynb`. The custom Synthea module definition is included in supplementary file `3-simple-breast-cancer-module.json`.

**Confidentiality**

When dealing with patient health information, ensuring confidentiality and data integrity is paramount. In order to ensure that all data processing is secure, our pipeline works within a Jupyter notebook hosted on an Azure "Confidential Compute" virtual machine [38]. Results are then made available on a local machine using SSH tunneling. These virtual machines have security features such as secure boot, a virtual trusted platform module (vTPM), boot integrity monitoring, and virtualization-based security [30]. Together, these features protect against persistent or advanced threats such as rootkits or bootkits, and ensure that the virtual machine has booted into a trusted environment as expected. Virtualization provides further security by isolating memory address spaces to remove any possibility of memory cross-contamination. Additionally, depending upon the underlying hardware, compute instances will include either Intel Software Guard Extensions (Intel SGX) or AMD Secure Encrypted Virtualization (SEV-SNP) support.

## 4 Discussion

One of the most unique aspects of this study is the analysis of long-read sequencing data together with clinical data for the first time. The outputs of the study will be an important use-case for researchers who are working on genetic data analysis. It is very important to acquire the necessary knowledge for the creation of decision support systems. We envision that eventually analyses such as our pharmacogenomic study will become commonplace, an automated analysis that occurs in real-time. This would provide clinicians with the ability to recommend prescriptions and treatments which are specifically tailored to the genetic makeup of each patient, based on the responses of similar individuals to each possible treatment. These types of research studies will also form the basis of data science models that are likely to be used in the near future.


**Fig. 8.** Sankey diagram showing distribution of patient treatments and outcomes.

The bulk of our pipeline is responsible for processing genomic and clinical data in a manner that ultimately results in a tabular format containing a row for each patient and a column for each feature. Machine learning on tabular data works well with a restricted set of features, but does not scale well as the number of columns increases due to the "curse of dimensionality" [46]. Moreover, some information is not easily represented in tabular format without loss of information or extreme data sparsity. For example, storing the medications taken by each patient would require either a boolean column for each possible medication, or categorical columns storing the last $n$ medications taken by each patient. Most patients may only take one or two medications, but the most heavily medicated patient may take dozens.

In order for machine learning to make complex clinical decisions, such tools will eventually need to be able aggregate information from data in many different formats. One possible solution would be an ensemble of classifiers which each work with a different data modality such as tables, graphs, images, and natural language. Multimodal machine learning is an active area of current research with great potential in the healthcare field [1].

Although we demonstrate variant calling using Cromwell on Azure for a public single patient dataset, the current inputs to our pipeline are a larger set of sample VCFs resulting from PacBio's human whole genome sequencing workflow. We do not currently have access to the original reads used to generate these VCFs. In future work, we would like to run a modified Cromwell workflow which outputs results in GVCF format. Unlike VCF files, GVCF files contain variant calling information about all (including reference/non-variant) positions on the genome, grouped into blocks of configurable size. VCF files only report confidently called variants. In this work, we impute reference quality and depth scores using the genome-wide average, akin to a GVCF with an exceptionally large block size. Transforming our pipeline to use GVCF inputs would simultaneously simplify post-processing of variants and improve the quality of all data regarding reference calls.

## Acknowledgements

The authors would like to thank Venkat Malladi and Olesya Melnichenko for their support throughout this project.

## Funding

Microsoft's Biomedical Platforms and Genomics Team supported this work by hosting Tim Dunn through the 2022 Microsoft Research Intern Program. Tim Dunn was also partially supported by the National Science Foundation under NSF Graduate Research Fellowship 1841052.

## Conflicts of Interest

Erdal Cosgun is a Senior Data and Applied Scientist at Microsoft Research on the Biomedical Platforms and Genomics Team. Tim Dunn was a Summer Research Intern on the same team during Summer 2022. The Biomedical Platforms and Genomics team was responsible for developing "Cromwell on Azure", which is included in the proposed pipeline. We also used Microsoft's Azure Cloud as the cloud provider for this pipeline.

## References

[1]Julián N Acosta, Guido J Falcone, Pranav Rajpurkar, and Eric J Topol. Multimodal biomedical ai. *Nature Medicine*, 28(9):1773–1784, 2022.

[2]James Agnew. Hapi fhir. https://github.com/hapifhir/hapi-fhir.

[3]Gil Alterovitz, Jeremy Warner, Peijin Zhang, Yishen Chen, Mollie Ullman-Cullere, David Kreda, and Isaac S Kohane. Smart on fhir genomics: facilitating standardized clinico-genomic apps. *Journal of the American Medical Informatics Association*, 22(6):1173–1178, 2015.

[4]David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Technical report, Stanford, 2006.

[5]Duane Bender and Kamran Sartipi. Hl7 fhir: An agile and restful approach to healthcare information exchange. In *Proceedings of the 26th IEEE international symposium on computer-based medical systems*, pp. 326–331. IEEE, 2013.

[6]Pacific Biosciences. Pacbio human whole genome sequencing workflow wdl. https://github.com/PacificBiosciences/pb-human-wgs-workflow-wdl.

[7]Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.

[8]Danton S Char, Michael D Abràmoff, and Chris Feudtner. Identifying ethical considerations for machine learning healthcare applications. *The American Journal of Bioethics*, 20(11):7–17, 2020.

[9]Haoyu Cheng, Gregory T Concepcion, Xiaowen Feng, Haowen Zhang, and Heng Li. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature methods*, 18(2):170–175, 2021.

[10]Google Cloud. Overview of the cloud healthcare api. https://cloud.google.com/healthcare-api/docs/concepts/introduction.

[11]AACR Project Genie Consortium, AACR Project GENIE Consortium, Fabrice André, Monica Arnedos, Alexander S Baras, José Baselga, Philippe L Bedard, Michael F Berger, Mariska Bierkens, Fabien Calvo, et al. Aacr project genie: powering precision medicine through an international consortium. *Cancer discovery*, 7(8):818–831, 2017.

[12]Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, et al. Twelve years of samtools and bcftools. *Gigascience*, 10(2):giab008, 2021.

[13]David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227, 1979.

[14]Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pp. 226–231, 1996.

[15]Food, Drug Administration, et al. Proposed regulatory framework for modifications to artificial intelligence/machine learning (ai/ml)-based software as a medical device (samd). 2019.

[16]Edward J Fox, Kate S Reid-Bayliss, Mary J Emond, and Lawrence A Loeb. Accuracy of next generation sequencing platforms. *Next generation, sequencing & applications*, 1, 2014.

[17]Carolyn Hutter and Jean Claude Zenklusen. The cancer genome atlas: creating lasting value beyond its data. *Cell*, 173(2):283–285, 2018.

[18]IBM. What is ibm fhir server? https://www.ibm.com/products/fhir-server.

[19]NIH: National Human Genome Research Institute. The cost of sequencing a human genome. https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost.

[20]NIH: National Human Genome Research Institute. Genetics vs. genomics fact sheet. https://www.genome.gov/about-genomics/fact-sheets/Genetics-vs-Genomics.

[21]The Broad Institute. Cromwell. https://github.com/broadinstitute/cromwell.

[22]The Broad Institute. Cromwell backends. https://cromwell.readthedocs.io/en/stable/backends/Backends/.

[23]Health Level Seven International. Welcome to fhir: Release 4b. https://www.hl7.org/fhir/.

[24]Todor Ivanov and Matteo Pergolesi. The impact of columnar file formats on sql-on-hadoop engine performance: A study on orc and parquet. *Concurrency and Computation: Practice and Experience*, 32(5):e5523, 2020.

[25]Miten Jain, Hugh E Olsen, Benedict Paten, and Mark Akeson. The oxford nanopore minion: delivery of nanopore sequencing to the genomics community. *Genome biology*, 17(1):1–11, 2016.

[26]Boyang Ji and Jens Nielsen. From next-generation sequencing to systematic modeling of the gut microbiome. *Frontiers in genetics*, 6:219, 2015.

[27]John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[28]Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, 17(1):1–9, 2019.

[29]AWS Labs. Fhir works on aws deployment. https://github.com/awslabs/fhir-works-on-aws-deployment.

[30]Microsoft. Confidential computing on azure. https://docs.microsoft.com/en-us/azure/confidential-computing/overview-azure-products.

[31]Microsoft. Cromwell on azure. https://github.com/microsoft/CromwellOnAzure.

[32]Microsoft. Fhir to synapse sync agent. https://github.com/microsoft/FHIR-Analytics-Pipelines/blob/main/FhirToDataLake/docs/Deployment.md.

[33]Microsoft. What is azure api for fhir? https://learn.microsoft.com/en-us/azure/healthcare-apis/azure-api-for-fhir/overview.

[34]Mullai Murugan, Lawrence J Babb, Casey Overby Taylor, Luke V Rasmussen, Robert R Freimuth, Eric Venner, Fei Yan, Victoria Yi, Stephen J Granite, Hana Zouk, et al. Genomic considerations for fhir®; emerge implementation lessons. *Journal of biomedical informatics*, 118:103795, 2021.

[35]Sergey Nurk, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V Bzikadze, Alla Mikheenko, Mitchell R Vollger, Nicolas Altemose, Lev Uralsky, Ariel Gershman, et al. The complete sequence of a human genome. *Science*, 376(6588):44–53, 2022.

[36]Adrian A Pater, Michael S Bosmeny, Adam A White, Rourke J Sylvain, Seth B Eddington, Mansi Parasrampuria, Katy N Ovington, Paige E Metz, Abadat O Yinusa, Christopher L Barkau, et al. High throughput nanopore sequencing of sars-cov-2 viral genomes from patient samples. *Journal of biological methods*, 8(COVID 19 Spec Iss), 2021.

[37]Murray Patterson, Tobias Marschall, Nadia Pisanti, Leo Van Iersel, Leen Stougie, Gunnar W Klau, and Alexander Schönhuth. Whatshap: weighted haplotype assembly for future-generation sequencing reads. *Journal of Computational Biology*, 22(6):498–509, 2015.

[38]Fernando Perez and Brian E Granger. Project jupyter: Computational narratives as the engine of collaborative data science. *Retrieved September*, 11(207):108, 2015.

[39]Kim D Pruitt, Tatiana Tatusova, and Donna R Maglott. Ncbi reference sequence (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*, 33(suppl_1):D501–D504, 2005.

[40]Franka J Rang, Wigard P Kloosterman, and Jeroen de Ridder. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome biology*, 19(1):1–11, 2018.

[41]Jeff Reback, Wes McKinney, Joris Van Den Bossche, Tom Augspurger, Phillip Cloud, Adam Klein, Simon Hawkins, Matthew Roeschke, Jeff Tratner, Chang She, et al. pandas-dev/pandas: Pandas 1.0. 5. *Zenodo*, 2020.

[42]Anthony Rhoads and Kin Fai Au. Pacbio sequencing and its applications. *Genomics, proteomics & bioinformatics*, 13(5):278–289, 2015.

[43]Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[44]Kishwar Shafin, Trevor Pesout, Pi-Chuan Chang, Maria Nattestad, Alexey Kolesnikov, Sidharth Goel, Gunjan Baid, Mikhail Kolmogorov, Jordan M Eizenga, Karen H Miga, et al. Haplotype-aware variant calling with pepper-margin-deepvariant enables high accuracy in nanopore long-reads. *Nature methods*, 18(11):1322–1332, 2021.

[45]X Yu Stella and Jianbo Shi. Multiclass spectral clustering. In *Computer Vision, IEEE International Conference on*, volume 2, pp. 313–313. IEEE Computer Society, 2003.

[46]Michel Verleysen and Damien François. The curse of dimensionality in data mining and time series prediction. In *International work-conference on artificial neural networks*, pp. 758–770. Springer, 2005.

[47]Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, and Scott McLachlan. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, 25(3):230–238, 2018.

[48]Tao Wang, David J Wu, Adam Coates, and Andrew Y Ng. End-to-end text recognition with convolutional neural networks. In *Proceedings of the 21st international conference on pattern recognition (ICPR2012)*, pp. 3304–3308. IEEE, 2012.

[49]Aaron M Wenger, Paul Peluso, William J Rowell, Pi-Chuan Chang, Richard J Hall, Gregory T Concepcion, Jana Ebler, Arkarachai Fungtammasan, Alexey Kolesnikov, Nathan D Olson, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature biotechnology*, 37(10):1155–1162, 2019.

[50]Justin M Zook, David Catoe, Jennifer McDaniel, Lindsay Vang, Noah Spies, Arend Sidow, Ziming Weng, Yuling Liu, Christopher E Mason, Noah Alexander, et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific data*, 3(1):1–26, 2016. https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG002_NA24385_son/PacBio_SequelII_CCS_11kb/HG002_GRCh38/HG002_GRCh38.haplotag.10x.bam.