Preliminary EHR Data Analysis Toward Developing Precision Oncology for ColoRectal Cancer

Vijay L. Badrinarayanan, Zhen Pu, Matthew J. Reilley,
Jianhui Zhou, Timothy W. Clark, Malathi Veeraraghavan

University of Virginia
{mr7db, mv5g}@virginia.edu
Dec. 31, 2019

Abstract

There is well established evidence of inter- and intra-patient variability in pharmacokinetics (PK) when Fluorouracil (5-FU), a key component of chemotherapy treatments for ColoRectal Cancer (CRC), is dosed by body surface area. Toward developing precision oncology methods that leverage Electronic Health Records (EHR) to customize dosing for patients, we started a data-analytics project for CRC patients. This document describes successes and failures encountered in this preliminary effort. We successfully procured a dataset consisting of five files: patient information, medications/doses, labs, vitals and Cancer Registry data. However, the dataset was incomplete and we could only extract sufficient information for 94 patients out of a starting set of 1460 patients. But we made important advances in defining single scores for complex multi-drug chemotherapy regimens, handled the problem of irregular time series, and were able to visualize the efficacy and toxicity of treatment for these 94 patients. We also made preliminary advances in creating subsets of patients, an initial step toward developing predictive models.

1 Introduction

ColoRectal Cancer (CRC) is the third most common and the second most lethal cancer in the United States, with over 145,000 new diagnoses and over 50,000 deaths expected in 2019 [1]. Fluorouracil-based therapies remain the standard treatment for over 60% of patients diagnosed with regional or distant stages of disease [2], and represent a frontline regimen for many other cancers (gastroesophageal, pancreatic, biliary, head and neck, and breast). Despite significant efforts to develop novel therapeutics for CRC, standard chemotherapy has remained essentially unchanged over the past few decades [3, 4]. Strategies to improve fluorouracil efficacy while reducing toxicity would impact the hundreds of thousands of patients who receive this treatment each year. We believe there is both a need and an opportunity to use archival data from patients with CRC to build a model for advanced prediction of toxicity and efficacy. Our long-term goal is the creation and validation of a clinical decision support tool for personalized chemotherapy dosing.

The chemotherapeutic compound 5-fluorouracil (5-FU) is a uracil analogue that after downstream metabolism results in pool imbalances of deoxynucleotides, severely disrupting DNA synthesis and repair, and resulting in DNA damage. 5-FU is processed by multiple different enzymes with activity influence by genomic and environmental differences. There is well established evidence of inter- and intra-patient variability in pharmacokinetics (PK) when 5-FU is dosed by body surface area. As a result, standard dosing can lead to highly different outcomes in patients [5]. Adjusting doses according to PK data show consistent results with significant reduction in toxicity and improved survival [6]. One of the major limitations to better dosing is the fact that measuring human 5-FU levels is technically challenging. Circulating 5-FU is rapidly degraded by an enzyme[1] present in blood cells and analysis of human samples

[1]The enzyme name is DPD (dihydropyrimidine dehydrogenase). There is a 2016 study on the toxicity-predicting value of this enzyme [7].

is performed at only a handful of centers. Therefore, despite decades of evidence supporting precision dosing of 5-FU, it has not been widely adopted [6].

We believe that using EHR in combination with available genomic data, we can build a model to better predict the effect of 5-FU on patients. By studying effects of the FOLFOX (which includes 5-FU) chemotherapy regimen in CRC patients, we are maintaining a narrow focus on a single chemotherapy regimen within relatively homogenous disease state, yet one that is common enough to allow for large datasets. We expect to first develop a predictive model based on pre-treatment and tumor genomic data, then subsequently expand the model to incorporate how patient and therapy changes over different time frames predicts future outcomes. Since FOLFOX is used across multiple gastrointestinal tumor types, we can subsequently adapt this model to other GI cancers. If what we learn from CRC cancer can be directly applied to similar patients in a tumor-agnostic fashion, this would be an incredibly powerful way to connect patients with rare cancers to benefit from a larger collective experience pool. The long-term goal of this work would be to build a clinical-decision support algorithm for personalized chemotherapy dosing of 5-FU that would be validated in a prospective clinical trial to improve outcomes in cancer patients.

This paper describes our work to date toward achieving our long-term goals. We successfully procured a dataset consisting of patient information, medications and doses, labs, vitals and Cancer Registry data for CRC patients treated from 2010-2017. We first ran exploratory data analysis and visualized efficacy and toxicity, as expressed by the impact of chemotherapy on three blood counts. We then tried to create clusters of patients based on age, gender and cancer stage/grade and analyzed the dataset for differences in toxicity and response between these clusters. We recently procured a second Cancer Registry dataset, which included race data for patients. This allowed for a preliminary comparison of cancer types across races.

Our key findings are as follows:

1 Chemotherapy regimens typically use multiple drugs. We figured out a method for combining dosage levels of different drugs in a regimen to create one score per regimen after normalizing doses to the values provided in National Comprehensive Cancer Network (NCCN) guidelines.
2 To deal with the problem of irregular time series and minor perturbations in the day of treatment/labwork, we used the number of weeks since the date of the first treatment as our measure of time.
3 We adopted the use of dose reduction and dose delay as metrics of toxicity based on prior work [8], along with our metrics of certain blood counts.
4 Our visualization of patients' disease response and toxicity to treatment illustrated how clinicians stop/start regimens, and delay or reduce dosage levels based on toxicity. The dilemma between balancing disease progression and toxicity was evident.
5 We learned from our mistakes in cohort-specification, which led to a significant reduction in the size of the dataset; specifically out of the original dataset of 1460 patients who were diagnosed with CRC between 2010-2017, we had useful data for only 94 patients. We have now procured a new dataset with over 1000 patients in all files (medications and doses, labs, and Cancer Registry).
6 The only measure of disease progression in this dataset was a CRC tumor marker called Carcino Embryonic Antigen (CEA), which is measured with blood work. We need imaging data, which is known to provide better indicators of efficacy and/or disease progression.
7 While age, gender, race and cancer stage/grade may be good variables for creating subsets (clusters) of patients, we anticipate needing genomic data.
8 Much larger datasets are required for creating subsets of patients, which is needed for improved prediction models.

Section 2 describes our starting point: acquisition of a CRC-patient dataset from the UVA medical center EPIC system. Section 3 lists a set of issues and problems, which made this dataset far from complete. Section 4 describes three types of analyses with results, though our conclusions are highly limited by the small size of useful data. This document is concluded in Section 5.

2 Dataset

The University of Virginia (UVA) Medical Center uses the EPIC EHR system. Data from the EPIC system was sent periodically to the UVA Clinical Data Repository (CDR) until June 2017. The CDR system offers researchers a GUI to define patient cohorts of interest. We used the CDR to specify our cohort. Since CDR is no longer updated, the TriNetX GUI is now available for researchers to specify cohorts, and the data is extracted directly from Epic. Section 2.1 describes our specified cohort and Section 2.2 describes the files we received from the analytics division that supports EHR-data-driven research.

2.1 Specified Cohort

In Dec. 2018, we used the CDR GUI to specify characteristics of our desired patient cohort:

- The ICD10 diagnosis codes[2] specified were C18 (Malignant Neoplasm of the Colon), with its billable child codes (C18.0-C18.9), C19 (Malignant Neoplasm of Rectosigmoid Junction) and C20 (Malignant Neoplasm of Rectum).
- Clinical labs data specified were CEA, White Blood Cells (WBC), Haemoglobin (HGB), Hematocrit (HCT), platelets (PLT), neutrophils (NETAH and NEUTAC), Blood Urea Nitrogen (BUN), creatine (CREA and CRPOC), potassium (K) and magnesium (MG). In addition, vitals data were requested.
- The following chemotherapy drugs were specified: 5-Fluorouracil, 5-Fluorouracil powd, Adrucil 2.5 gm/50ml IV soln, Adrucil 5 gm/100ml IV soln, Adrucil 50 mg/ml IV soln, Adrucil 500 mg/10ml IV soln, Adrucil IV, Fluorouracil infusion (home health), Fluorouracil infusion (non-chargeable), Fluorouracil IV and Fluorouracil powd.
- Date of diagnosis was specified as the range [2010-2017].

2.2 Received Data files

Data files were received from the MD-DMED Institutional Analysis group at UVA's School of Medicine. Our specified cohort yielded data for up to 1460 patients in the form of 5 files: (i) Patient information, (ii) Labs, (iii) Medicines and Doses, (iv) Vitals, and (v) Cancer Registry. The first four datasets were extracted using EPIC Caboodle in Mar. 2019, while this initial Cancer Registry data was extracted from the CDR in June 2019. We obtained a new Cancer Registry dataset from Emily Couric Cancer Center (ECCC) in Nov. 2019. Below is a description of the data dictionaries for each of these files.

Patient information: This file has 12200 rows of data for 1460 patients. Table 1 describes the columns in this file.

Labs: This file has 51781 rows of data with lab records for 1074 patients. Each row contains one observed lab value. Table 2 describes the columns in this file.

Medications and Doses: This file has 8005 rows of data with records for 171 patients. Each row contains information about the medicines (with doses) administered to a patient in each encounter of a visit (a visit could have multiple encounters). Table 3 describes the columns in this file.

Vitals: This file has 188002 rows with records for 1151 patients. Vitals include height (in inches), weight (in pounds), BMI (calculated from weight and height), blood pressure, temperature (°F), pulse and respiration (count per min). Each row has information for one vital measurement. Table 4 describes the columns in this file.

Cancer Registry: This registry is maintained by the UVA Emily Couric Cancer Center in compliance with CSC National Program for Cancer Registries[3]. Registrars extract data manually from the EPIC EHR system, and enter the

[2] https://www.icd10data.com/ICD10CM/Codes/C00-D49/C15-C26

[3] https://www.cdc.gov/cancer/npcr/index.htm

Column Name	Description
MRN	Medical Record Number of the patient
Gender	Gender of the patient
DOB	Date of birth of the patient
AliveStatus	Status indicating if the patient was alive or dead
AliveStatusDate	Date when the patient's AliveStatus was last updated
VisitID	Unique ID for a patient's visit to the clinic
Admit_Datetime	Date and time of admission to the hospital (or date and time of outpatient clinic visit)
Discharge_Datetime	Date and time of discharge from the hospital (or date and time of outpatient clinic visit)

Table 1: Data Dictionary for the Patient Information File

Column Name	Description
MRN	Medical Record Number of the patient
VisitID	Unique ID for a patient's visit to the clinic
Lab_Test	Name of the lab parameter tested, e.g., WBC
Lab_Value	Observed value of the particular lab
Lab_Obs_Datetime	Date and time when the patient's blood was drawn
Admit_Datetime	Date and time when the patient was admitted to the hospital or time of visit to the clinic

Table 2: Data Dictionary for the Labs File

Column Name	Description
MRN	Medical Record Number of the patient
Gender	Gender of the patient
DOB	Date of birth of the patient
AliveStatus	Status indicating if the patient was alive or dead
AliveStatusDate	Date when the patient's AliveStatus was last updated
VisitID	Unique ID for a patient's visit to the clinic
Admit_Datetime	Date and time of admission to the hospital (or date and time of outpatient clinic visit)
EncounterKey	Unique ID for the encounter in which the medicine was administered
MedicationKey	Categorical key for the medicine listed
Medication_Name	Name of the medicine administered
Medication_Form	Form in which the medicine was taken (solution or tablets)
Medication_Route	Method used to administer the medicine (IV, oral, etc.)
Medication_Dose	Quantity of medication given to the patient
Med_Dose_Unit	Unit for the quantity
Med_Dose_Count	Number of times the dose was given in that particular encounter
Med_Administered_Datetime	Date and time when the medicine was given to the patient

Table 3: Data Dictionary for the Medications and Doses File

Column Name	Description
MRN	Medical Record Number of the patient
Gender	Gender of the patient
DOB	Date of birth of the patient
AliveStatus	Status indicating if the patient was alive or dead
AliveStatusDate	Date when the patient's AliveStatus was last updated
CaseID	Unique ID similar to VisitID
Admit_Datetime	Date and time of admission to the hospital (or date and time of outpatient clinic visit)
EncounterKey	Unique ID for the encounter in which the vitals were measured
VitalsKey	Categorical key for the vital specified in the row (e.g., blood pressure)
Vitals_Name	Name of the vital that was measured
Vitals_Value	Value of the measured vital
Vitals_Abnormal	A flag to indicate if the vital measured was outside the normal limits (an indicator of toxicity)
Vital_1stInEncounter	Value: 1 if it was the first measurement in the encounter for that corresponding vital; 0: otherwise
Vital_LastInEncounter	Value: 1 if it was the last measurement in the encounter for that corresponding vital; 0: otherwise
Vitals_Datetime	Date and time when the vital was measured

Table 4: Data Dictionary for the Vitals File

extracted data into the cancer registry. This program requires the following:"Central cancer registries must collect and submit data for all reportable cancers and benign neoplasms, including at a minimum, primary site, histology, behavior, date of diagnosis, race/ethnicity, age at diagnosis, gender, stage at diagnosis, first course of treatment according to CDC specifications, and other information required by CDC." Therefore, for most patients, information on just their first course of treatment is available in this dataset.

We obtained an initial version of this file in June 2019, and a second version of the file in Nov. 2019. Across both cancer registry data files, we have tumor data for 1727 patients. The June 2019 cancer registry file had 10231 rows with records for 774 patients. Table 5 describes the columns in the June 2019 file.

The Nov. 2019 cancer registry file contained 1426 rows of data for 1389 patients. This new data file had additional fields, relative to the June 2019 extract, as listed in Table 6. Other differences were: (i) Site variable has codes, such as C209 for rectum, in Table 6, while Site has only names like colon cecum in Table 5, and (ii) CC and qualsurv fields in Table 5 are not in this new table. The CC codes in Table 5 were different from the Site codes in Table 6.

3 Incomplete dataset

The number of patients for whom we had data in the received files described above varied considerably. The first six rows of Table 7 provides these numbers. Most of this work was done before Nov. 2019, and hence most of the presented results do not include the 2019 Nov. Cancer Registry data.

The goal for our exploratory data analysis was to understand the impact of FOLFOX and FOLFIRI regimens on a patient's CEA (measure of disease progression), and on three lab values (WBC, PLT, HGB) selected for their use in estimating toxicity. An intersection of patient MRNs from the Medications and Doses file, Labs file and 2019 June Cancer Registry file yielded a total of only 109 patients as shown in Row 7 of Table 7. We then restricted the

Column Name	Description
MRN	Medical Record Number of the patient
Age	Age of the patient
Dxdate	Date of diagnosis
Site	Location of the tumor
CC	Clinical classification code for the diagnosis
TumorSize	Size of the tumor
SizeQual(ifier)	Unit of measure for the tumor size
Histology	Type of cancer cells
CancerType	Broader name for the cancer, e.g., Colon, when the site is colon cecum
Grade	Grade of the cancer
SurgSummary and SurgDate	Surgery details, if applicable
ChemoSummary, ChemoStart, ChemoEnd	Details of chemotherapy, if applicable
HormoneSummary, HormoneStart, HormoneEnd	Details of hormonal therapy, if applicable
RadSummary, RadStart, RadEnd	Details of radiation therapy, if applicable
LastStatus and LastContact	Number of days since the diagnosis of tumor when the status of the patient was obtained (but the LastStatus field does not have a status, instead it has the same value as LastContact)
VitalStatus	Vital status (dead/alive) of patient as of the last status day
CancerStatus	Cancer status (free or not free from disease) of patient as of the last status day
QualSurv	Quality of survival of the patient during the course of the treatment
cT, cN, cM, cGrp	Clinical staging: Size of tumor (T), degree of spread to regional lymph nodes (N), presence of distant metastasis (M), and cancer stage[4]
pT, pN, pM, pGrp	Pathological staging (from tissue biopsy; and hence this value, if available, is given more weight than clinical stage)

Table 5: Data Dictionary for the Cancer Registry June 2019 Extract

Class_of_Case	Whether DX and/or RX were performed here
Date_First_Contact	If DX was done elsewhere, the Dxdate in Table 5 could be different
Chemotherapy, Hormonetherapy, Radiation, Surgery	These fields are in addition to the corresponding Summary, Start and End dates fields listed in Table 5
Immunosummary, Immunotherapy, Immunostart, Immunoend	Columns related to Immunotherapy
Physician_Managing	Name of the patient's physician
Race	Ethnicity of the patient
postal_code	Postal code of the patient's home address

Table 6: Additional Fields in the Cancer Registry Nov. 2019 Extract

Row No.	Data file	Patient count
1	Patients	1460
2	Labs	1074
3	Medications and Doses	171
4	Vitals	1151
5	2019 Mar. Cancer Registry	774
6	2019 Nov. Cancer Registry	1391
7	Patients with Meds/Doses, CEA and Cancer Registry data	109
8	Patients with at least 2 CEA measurements	96
9	Patients among 96 with BSA data	94
10	Patients with measurements for WBC, HGB and PLT	79

Table 7: Patient counts from original and processed data files

dataset to patients who had at least two CEA measurements, which dropped the cohort size to 96. Due to a lack of weight information, we could not compute BSA for two patients, which further dropped the dataset size to 94. Finally, as shown in row 10, only 79 out of the 94 patients had measurements for WBC, HGB and PLT. The others had no measurements for these labs. The analysis presented in the next section thus used these limited datasets.

In addition to the problem of having incomplete datasets, we found other issues:

- In the set of medicines we specified in our cohort definition, we left out Xeloda (Capecitabine), an oral tablet form of Fluorouracil (5FU), which is a key ingredient in both FOLFOX and FOLFIRI regimens. This caused incomplete medications data for few patients.
- By limiting the set of medicines in our cohort definition, we lost the opportunity to control for comorbidities when clustering patients into cohorts.
- Lab measurements for HGB, PLT and WBC when made at the HOPE clinic were post-fixed with “H," i.e., HGBH, WBCH, PLTH. As these keys were not specified in the SQL query applied to the EPIC EHR database, these measurements were not available for certain patients.
- Censoring is another problem in these datasets. A patient was necessarily alive at their[5] last visit. Their survival time past this date is often not available. The AliveStatus in the Patients file rarely showed “Dead." This made a study of survival metrics challenging.
- As noted in Section 2, the Cancer Registry is updated using a painstaking manual process executed by certified registrars. This is a slow process and usually incurs a delay of about 6 months. More importantly, it only records the first course of treatment. Therefore, it is not a good source of information for creating cohorts based on their prior treatments. For example, response to chemotherapy may depend on whether or not the patient previously had surgery, but we cannot easily group patients on this measure using the Cancer Registry data. We need a dataset on procedures executed on patients.
- Some patients’ weight data was missing from the Vitals file. Without weight, we could compute Body Surface Area (BSA), an important measure required to determine whether or not a patient had dose reduction in their chemotherapy treatments.
- While C18-C20 were the ICD10 codes specified for CRC, patients who had Anal cancer (C21 ICD10 code) were included in our dataset. This was an artifact of the word Colon appearing in their records due to metastasis, and the use of this keyword in the SQL query.

[5]Per the Merriam-Webster dictionary, we use “they” and associated terms as gender-neutral pronouns https://www.merriam-webster.com/words-at-play/singular-nonbinary-they

4 Data Analysis

Section 4.1 presents illustrative graphs for 5 patients, showing the impact of treatment on CEA and toxicity indicative lab counts, WBC, PLT and HGB. Section 4.2 shows our preliminary attempts at creating patient clusters. Section 4.3 describes the challenges in quantifying measures of survival. A preliminary look at the impact of race is presented in Section 4.4 using the new Nov. 2019 Cancer Registry dataset. Finally, Section 4.5 describes a summary table that we created with one row for each patient.

4.1 Visualization

Measure of disease progression (CEA): Our first step was to extract data on CEA measurements for each patient. Each measurement includes the date on which the CEA was measured, and the observed CEA value. CEA is known to be insufficient for tracking disease progession. Imaging results are more trustworthy but we are yet to procure the required datasets, and hence we use CEA, our only available metric for disease progression, in this analysis.

Measure of medications administered: Our second step was to extract data on the medications administered to each patient who was treated with the FOLFOX or FOLFIRI chemotherapy regimens. FOLFOX is a combination of three drugs: FOL stands for Folinic Acid (Leucovorin Calcium), F stands for Fluorouracil (5FU, Adrucil), and OX stands for Oxaliplatin. First, Oxaliplatin (*Oxa*) and Leucovorin are administered at an infusion center. Then a bolus shot of 5FU (*5FU_bolus*) is given as an injection. Finally, the patient is sent home with a pump that is attached to a patient's port (or equivalent device) and 5FU is administered over a 46-hour period (*5FU_inf*). FOLFIRI is similar, except Irinotecan (*Irino*) replaces Oxaliplatin. As described in Table 3, the Medications and Doses file contains multiple rows of entries for each patient on each visit, with each row indicating the medication type and dose. Therefore, for each patient, our data-parsing code extracts the date administered and dosage value for five medications: 5FU_bolus, 5FU_inf, Leucovorin, Oxa, and Irino.

To obtain a single score for medications in these complex regimens, we addressed *two questions*:

i. How do we compare doses received by different patients, and by a single patient on different dates?
ii. How do we combine the four component medications for these two combination chemotherapies?

The answer to the *first question* is drawn from the NCCN guidelines for physicians, which recommends normalized dosage values that should be multiplied by a patient's Body Surface Area (BSA). The BSA value is computed from a patient's body weight and height as follows:

$$BSA = \sqrt{\frac{\text{height in cm X weight in kg}}{3600}} \tag{1}$$

Since a cancer patient's weight can change considerably, to compare doses administered to even a single patient at different visits, we need to access the latest weight and height data from the Vitals data file as described in Table 4. As height is stored in inches and weight in oz, appropriate conversions to cm and kg, respectively, are needed before applying (1). By dividing the dosage value by the BSA, we find the normalized dosage value, which then allows for a comparison of doses between patients, and also doses administered to a single patient on different dates.

Our answer to the *second question* is to use weights for each of the component medications, and compute a weighted combined score. The NCCN guideline values for per-BSA full-dosage values are: 5FU_bolus: 400 mg/m^2; 5FU_inf: 2400 mg/m^2; Leucovorin: 400 mg/m^2; and OXA: 85 mg/m^2. We used weights of 65% for 5FU_inf, 30% for Oxa, 5% for 5FU_bolus, and 0% for Leucovorin (Leucovorin does not have a major impact on response or toxicity and is hence left out of our formula for FOLFOX score.) Along with the required normalization by BSA, we used these weights to compute a FOLFOX score for each administered treatment for each patient.

$$FOLFOX\_score = 0.65 * 5FU\_inf/(BSA * 2400) + 0.3 * Oxa/(BSA * 85) + 0.05 * 5FU\_bolus/(BSA * 400) \tag{2}$$

In the FOLFIRI regimen, the 5FU_inf and 5FU_bolus full-dosage values are the same as in the FOLFOX regimen, and Irino full-dosage value is 180 mg/m^2. Using the same weights to compute FOLFIRI score as we did in (2), with Irino replacing OXA, we get

$$FOLFIRI\_score = 0.65 * 5FU\_inf/(BSA * 2400) + 0.3 * Irino/(BSA * 180) + 0.05 * 5FU\_bolus/(BSA * 400) \quad (3)$$

A patient who receives full-dosage on all three components used in (2) or (3) gets a FOLFOX or FOLFIRI score of 1, respectively. Dose reduction in any of the component medications results in a score that is less than 1.

To obtain a set of discrete levels for easier dose reduction computation, we approximate the FOLFOX_score and FOLFIRI_score (shown as "score" below) as follows:

$$\text{if } (i/10) \leq score < (i+2)/10 \quad \text{score } = (i+1)/10, \quad \text{where } i = 1, \cdots 15 \quad (4)$$

Some patients received the 5FU treatments without Oxa or Irino. For these treatments, we used a 5FU_score defined as:

$$5FU\_score = 0.65 * 5FU\_inf/(BSA * 2400) + 0.05 * 5FU\_bolus/(BSA * 400) \quad (5)$$

Since the weights do not add to 1, the 5FU_score is always less than 1.

In addition to these two chemotherapy regimens, a third regimen called **CAPOX** is also used, in which the 5FU infusion is substituted with Capecitabine oral tablets. We do not include results for this regimen here due to a lack of data.

Measure of time: Our third step was to address the problem of irregular time series that arises from (i) a single patient's visits being aperiodic (even though these two chemotherapy regimens recommend that a treatment be administered every two weeks, dose delays occur quite frequently), and (ii) different patients visiting on different days. We handled this problem in the following manner. We decided to use "week" as the discrete time unit instead of day. This decision helped resolve minor date discrepancies; for example, a patient had their OXA infusion and 5FU_bolus on one day but returned the next day to start their 5FU_inf because they simply forgot to bring in their pump.

We marked time as week 0 for the week in which a patient had their first FOLFOX or FOLFIRI treatment. Some patients had a CEA or other lab measurement on a week prior to week 0, and hence these lab values would have a negative time value, e.g., a CEA measurement on a day that is two weeks prior to the day of the patient's first treatment would be assigned -2 on the time axis.

Before we concluded that we should set week 0 as the week of the first treatment, we had used number of weeks from the date of diagnosis as our measure of time. The assumption was that this latter definition would allow us to characterize measures of survival. But for reasons explained in a later section, this proved difficult.

Measure of toxicity: Our final step in this visualization exercise was to select three lab measurements that are indicative of toxicity (side effects) and plot these measures along with the medication scores as a function of time. The three labs are WBC, PLT and HGB. These labs are typically taken just before each treatment, and so most commonly, the measurement dates match the treatment dates. But since the same rule is used with lab measurements as with treatments, weeks since the first treatment is used as the time index for the presented lab values. For certain hospitalized patients, there can be multiple lab measurements within a week, in which case we used the first value.

Results: As examples, we randomly selected results for five patients[6] and present two sets of graphs for each. In both sets, the x-axis is the time measure, i.e., weeks from the first treatment. In the first set of graphs, the right y-axis,

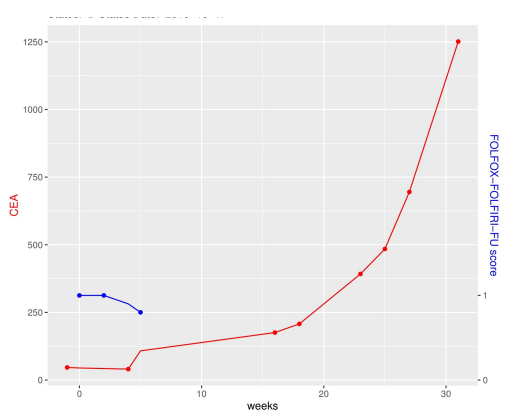
[6]They all just happened to be male, but we have a comprehensive set of graphs for all 94 patients, which includes graphs for both males and females.

Lab	Gender	Minimum	Maximum	Unit
CEA	Both		5.0	ng/ml
HGB	Male	13.2	16.6	grams/dL
HGB	Female	11.6	15	grams/dL
WBC	Both	3.4	9.6	cells/L
PLT	Male	135	317	billion/L
PLT	Female	157	371	billion/L

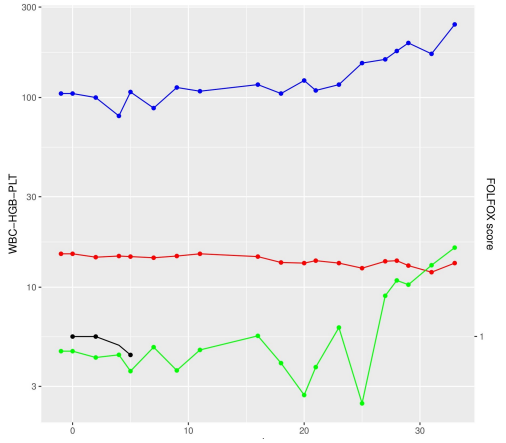
Table 8: Limits for Lab Values

labeled "FOLFOX-FOLFIRI-FU score," shows the FOLFOX-score, FOLFIRI-score, or 5FU_score based on the treatment received by the patient in that week. If a graph for a type of score is missing, e.g., 5FU_score, it means the patient did not receive that type of treatment. The right y-axis in the second set is just FOLFOX-score since the toxicity impact of only FOLFOX was considered in this study. The study will be extended to consider the toxicity effects of FOLFIRI. The left y-axis, in the first set of graphs, shows the CEA values, while in the second set of graphs, it shows the WBC, PLT or HGB values. A de-identified numbering system was used for patients, and the patient number is listed in the figure captions. We have such graphs for each of the 94 patients in the set described in Section 3.

To intrepret results, we need the minimum and maximum limits for each lab. Table 8 shows these limits.



(a) Measure of response with treatments: Red: CEA values; Blue: FOLFOX_score



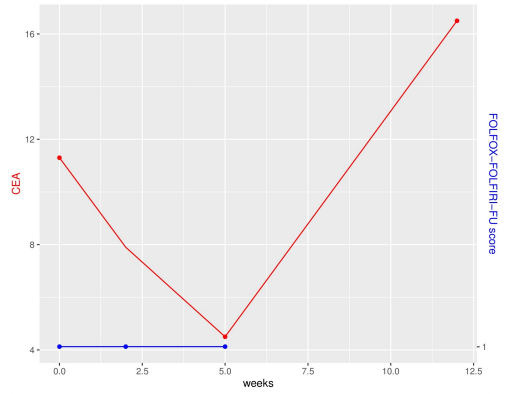
(b) Measure of toxicity with treatments: Red: HGB, Blue: PLT, Green: WBC; Black: FOLFOX_score

Figure 1: Patient 2 (male): Impact of chemotherapy treatments

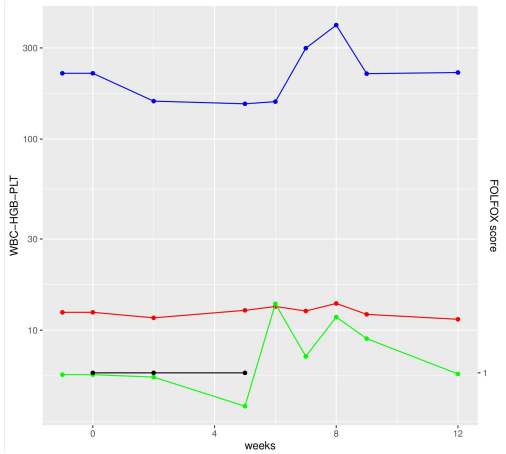
Fig. 1 shows the two sets of graphs for Patient 2. This patient had three cycles (treatments) of FOLFOX, but during the week of his second cycle, the patient's WBC and PLT started dropping, and therefore, there was dose reduction in the third cycle (see drop in FOLFOX_score graph). Toxicity did not subside on reducing the dose. This is likely one of the reasons that the FOLFOX treatment was stopped. We can also see that around week 16, the patient's condition worsened and CEA values increased. When we looked at the patient's complete history file, we found that he was given a different regimen during that time, but that did not help with his disease as the CEA values kept increasing.

Patient 7 graphs in Fig. 2 show that after the third treatment of FOLFOX, his WBC count increased dramatically. This could indicate that the patient was getting sick, i.e., the body's immune system was pumping out white blood cells to fight disease. We looked up the patient's complete history file and found that the patient was admitted to the hospital due to certain complications.

Patient 21 graphs in Fig. 3 show an elaborate story of a patient who had three types of regimens due to toxicity.

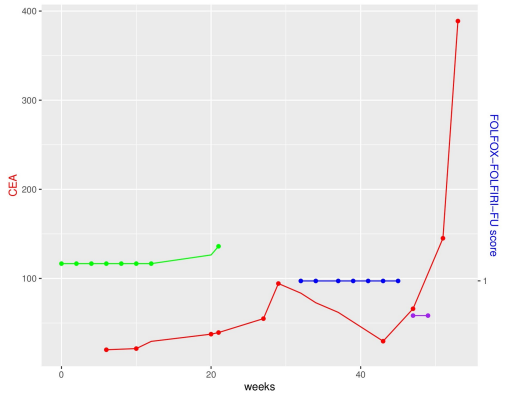


(a) Measure of response with treatments: Red: CEA values; Blue: FOLFOX_score

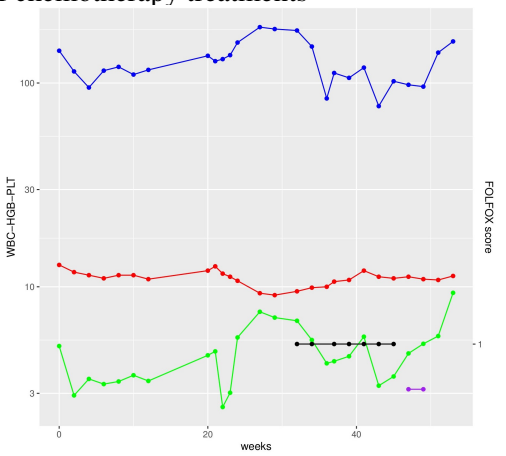


(b) Measure of toxicity with treatments: Red: HGB, Blue: PLT, Green: WBC; Black: FOLFOX_score

Figure 2: Patient 7 (male): Impact of chemotherapy treatments



(a) Measure of response with treatments: Red: CEA values; Blue: FOLFOX_score; Green: FOLFIRI_score; Purple: 5FU_score

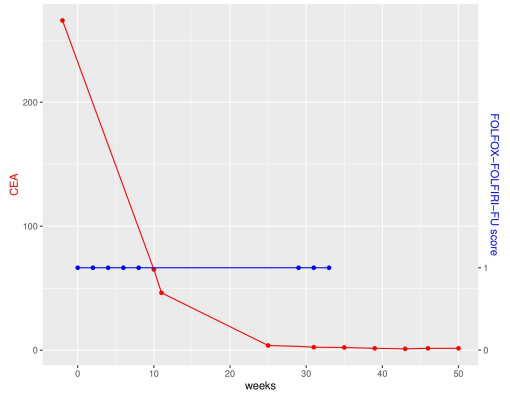


(b) Measure of toxicity with treatments: Red: HGB, Blue: PLT, Green: WBC; Black: FOLFOX_score; Purple: 5FU_score

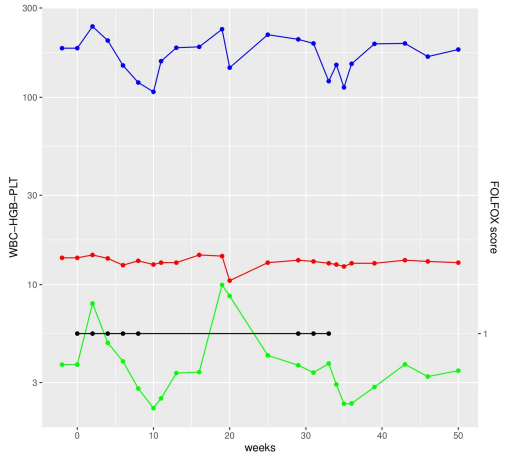
Figure 3: Patient 21 (male): Impact of chemotherapy treatments

The patient started his treatment with FOLFIRI and his CEA level remained low. He was administered a higher dose of FOLFIRI than normal levels. But around week 20, his WBC count started to plummet and FOLFIRI had to be stopped. This change coincided with progression of disease (as evidenced from the CEA increase) and the patient was switched to FOLFOX after a break of several weeks. FOLFOX was effective in bringing his CEA down, but also led to decreases in his WBC and PLT counts. This toxicity was mostly an effect of Oxaliplatin, and therefore, oxaliplatin was cut from his regimen and just 5FU was administered. While his WBC and PLT counts recovered, his disease started progressing as seen in the CEA graph.

Patient 24 graphs in Fig. 4 show a patient with a successful recovery from the disease with administration of FOLFOX. Based on these graphs, it appears that this patient was treated with adjuvant chemotherapy. A likely sequence was: 1) chemotherapy, 2) surgery (which would explain the 20-week break), and 3) further post-operative chemotherapy. An important point illustrated by this example is that this kind of time delay discrepancy between treatment and observed effect is difficult for humans to discern, but graphs show that such discrepancies can occur and are important to measure.

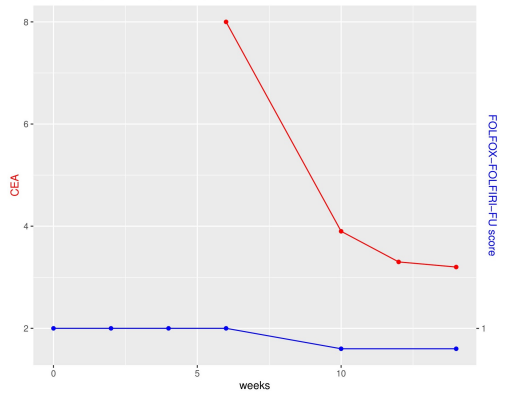


(a) Measure of response with treatments: Red: CEA values; Blue: FOLFOX_score

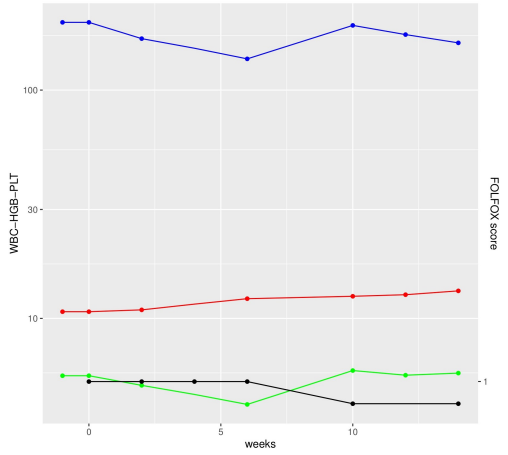


(b) Measure of toxicity with treatments: Red: HGB, Blue: PLT, Green: WBC; Black: FOLFOX_score

Figure 4: Patient 24 (male): Impact of chemotherapy treatments



(a) Measure of response with treatments: Red: CEA values; Blue: FOLFOX_score



(b) Measure of toxicity with treatments: Red: HGB, Blue: PLT, Green: WBC; Black: FOLFOX_score

Figure 5: Patient 30 (male): Impact of chemotherapy treatments

Patient 30 graphs in Fig. 5 show that his four initial FOLFOX cycles were beneficial in dropping his CEA, but his WBC count also dropped. The dose reduction in week 10 was likely due to this toxicity. His CEA stayed low in spite of the dose reduction.

In sum, these sample graphs illustrate how clinicians stop/start regimens, and delay or reduce dosage levels based on toxicity indicators such as WBC and PLT. Further, CEA seems to improve for most patients with the administration of these regimens especially at full dosage levels per NCCN guidelines. The dilemma between balancing disease progression and toxicity is evident.

4.2 Toxicity and Response Analyses

Our sample set, as noted before was small (just 94 patients), but in this set of analyses, we try to determine if there is any evidence of the effects of age, gender, cancer stage and grade at diagnosis, on efficacy (as measured by CEA) and toxicity. Only the FOLFOX regimen was considered in this analysis.

We created a set of bargraphs that capture the effects of 4 independent variables: age, gender, cancer stage (group),

Predictor	Cohort	Size	Cohort	Size	Cohort	Size	Cohort	Size
Gender	Male	37	Female	18				
Age at diagnosis (years)	19-50	21	51-60	22	61-70	10	71 and above	3

Table 9: Total size of cohort for dose-delay based toxicity analysis: 55 patients

and cancer grade, on toxicity and on response. The cancer stage (group) value used is a max(cGrp, pGrp) (see Table 5). The dependent variables for toxicity and response are defined in the subsections below.

Our ultimate goal is to provide clinicians a tool to customize dosage levels for patients, at least at a cluster level. The goal of this set of analyses is to understand the importance of each independent variable (feature) in creating clusters.

Toxicity: With our current datasets, we postulated three measures of toxicity: (i) WBC-PLT-HGB-impact, (ii) Dose-Delay and (iii) Dose-Reduction [8]. The first measure WBC-PLT-HGB-impact was set to 1 if any of the corresponding lab measurements during administration of the FOLFOX regimen were outside the limits specified in Table 8. Dose-Delay was set to 1 if a FOLFOX cycle was delayed by more than 5 days relative to the next expected date (based on the two-week prescribed intervals). This definition is based on an assumption that a patient may delay a dose for a few days due to travel or other considerations, but a delay of 5 or more days is an indication of toxicity. Dose-Reduction was set to 1 if the FOLFOX_score was dropped by at least 20% during the regimen.

We hypothesize that just age at diagnosis and gender are likely to effect toxicity, not the cancer group (stage) or grade. For age, we grouped patients into four categories as shown in our graphs.

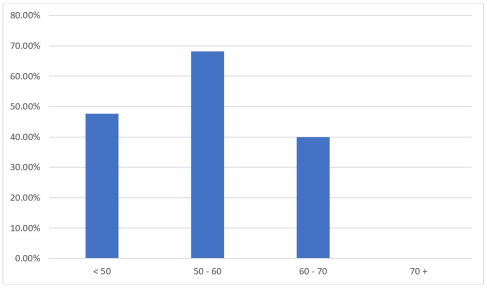
To compute each of these metrics, we needed a corresponding cohort of patients whose lab measurements and treatments fit our definitions. It was challenging to find these cohorts in our limited dataset with 94 patients. For example, there were some patients who started out with a FOLFOX_score less than 1. This occurred because, as mentioned in Section 3, some patients received (oral) Xeloda tablets instead of the 5FU infusion and Xeloda was not included in our medications list. The initial FOLFOX_score for these patients was 0.4 or 0.2 based on whether they received a full starting dose of OXA or reduced dose, respectively. We dropped such patients from this analysis since we did not have an accurate picture of their dosage levels. Due to other such problems, we could only retain the Dose-Delay metric, and hence offer results for just this metric.

Table 9 shows that out of 94 patients, only 55 patients started with a full-dosage level, and had a sufficient number of treatments for us to identify dose delays of 5 or more days with certainty. For example, 15 patients were not administered the FOLFOX regimen, and 3 patients did not have a sufficient number of FOLFOX cycles.

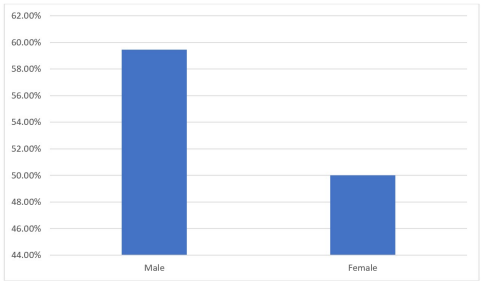
Fig. 6 shows, for each cohort, the percentages of patients who required dose delays. No clear conclusion is possible from these charts due to the limited sizes of the cohorts. No bar is provided for patients 71 and up because we set a minimum threshold of 5 patients for computing percentages and this cohort had only 3 patients as seen in Table 9.

Response to treatment: Two types of responses were considered:

- *Initial-Response* A value of 1 (considered to have a response) was assigned to a patient if a (i) their CEA was above 5, and the value dropped from the first measured value (taken within a two week period before, or in the week of the first treatment) to the measured value after the third treatment, or (ii) the CEA was below 5 at or before the first treatment and stayed below 5 after the third treatment.
- *Response-After-Dose-Reduction* The CEA measurement taken before a dosage drop was compared to the CEA measurement after two lowered-dose treatments. If the first CEA was already below 5, and the value stayed below 5 after two lowered-dose treatments, or if the original CEA was above 5 and a drop in CEA was observed even after two lowered-dose treatments, the patient was assigned a value of 1 (considered to have a response) for this metric.



(a) Dependence on Age



(b) Dependence on Gender

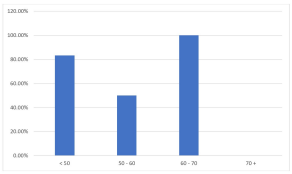
Figure 6: Percentage of patients in each cohort who had dose delays in the FOLFOX regimens

Predictor	Cohort	Size	Cohort	Size	Cohort	Size	Cohort	Size	Cohort	Size
Gender	Male	20	Female	9						
Age at diagnosis (years)	19-50	12	51-60	10	61-70	6	71 and above	1		
Cancer stage	1	1	2	4	3	7	4	14	Unk	3

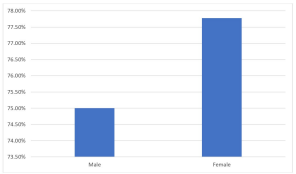
Table 10: Total size of cohort for initial response analysis: 29 patients

For the above two dependent variables, three independent variables of interest are age, gender and cancer stage.

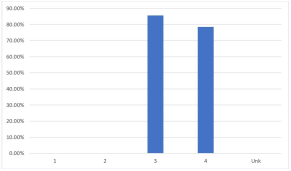
To compute each of these metrics, we needed a corresponding cohort of patients whose lab values and treatments fit our definitions. An analysis of our limited dataset resulted in many patients not fitting the requirements of these definitions, and therefore, we are able to present results for just Initial-Response. Table 10 shows that out of 94 patients, only 29 patients qualified per our definitions for this analysis. For example, 15 patients were not administered the FOLFOX regimen, 27 did not have CEA measured before or on the day of their first treatments, and for 2 patients, neither of their two CEA measures (required to qualify for the set of 94) were taken within a two-week period prior to their first treatments.



(a) Dependence on Age



(b) Dependence on Gender



(c) Dependence on Cancer Group (Stage)

Figure 7: Percentages of patients who had initial response to FOLFOX

The cohort sizes are too small to make definitive conclusions from the percentages of patients who had initial responses to FOLFOX as seen in Fig. 7.

4.3 Survival Analysis

We considered two metrics: overall survival and survival rate. These definitions are taken from https://www.cancer.gov/publications/dictionaries/cancer-terms/. Overall survival, commonly referred to as OS, is defined as the length of time from either the date of diagnosis or the start of treatment that patients are still alive. Survival rate, also called overall survival rate, is defined as the percentage of people in a study or treatment group who are still alive for a certain period of time after they were diagnosed with or started treatment for a disease, such as cancer. The survival rate is often stated as a five-year survival rate, which is the percentage of people in a study or treatment group who are alive five years after their diagnosis or the start of treatment.

As described in the previous section, we first used number of weeks from the date of diagnosis as our measure of time. The assumption was that such a measure could then be used to characterize survival. While the Cancer Registry file had VitalStatus as a metric, i.e., whether the patient was alive or dead, this field was clearly not updated. We learned that the Cancer Registry primarily records the first course of treatment as mentioned in Section 2. Further, the Patient information file has an AliveStatus field, but even this field appeared to have not been updated. Data about patients are entered into EPIC by clinicians primarily, and there appears to be no systematic method for updating this field for patients after death. Data for each patient is naturally censored up to the date of their last visit. Therefore, computing both overall survival and survival rate was challenging with our small dataset.

4.4 New Cancer Registry DataSet Analysis

As noted in Section 2, we received a new cancer registry dataset in Nov. 2019. This file had 1426 rows of data pertaining to 1389 patients. We identified a set of hypotheses/questions that could be tested with an analysis of just this dataset.

- What is the distribution of patients across age groups? What about the stage of cancer at diagnosis? If younger people get tested less frequently, is the percentage of patients diagnosed with stage 3/4 disease dependent on age?
- Is there a dependence of the site of cancer on age, gender or race?
- Does gender or race have an impact on stage at diagnosis?
- Is there a difference in the stage-at-diagnosis for patients from different zip codes (assumes wealth distribution is not uniform across zip codes)?
- Is there a dependence on prescribed first course of treatment (chemotherapy, surgery, radiation, etc.) on stage of cancer or other variables?
- Check the assumption that, per NPCR requirements, data for most patients are only available for the first and/or second course of treatments.

Here we present results with partial answers for just the first two of the above-listed questions, but will soon answer the rest of the questions (and potentially raise and answer new questions) in our forthcoming work.

Figure 8 shows the distribution of patient's age at diagnosis.

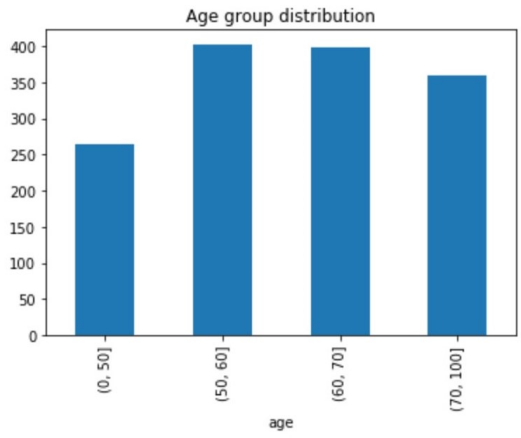


Figure 8: Patient's age distribution

Figure 9 shows the number of patients for each type of CRC tumor site broken down by races. The total number of patients in the three categories are: (i) white: 1191, (ii) black: 171, and (iii) other: 64. No early conclusions are evident from these graphs.

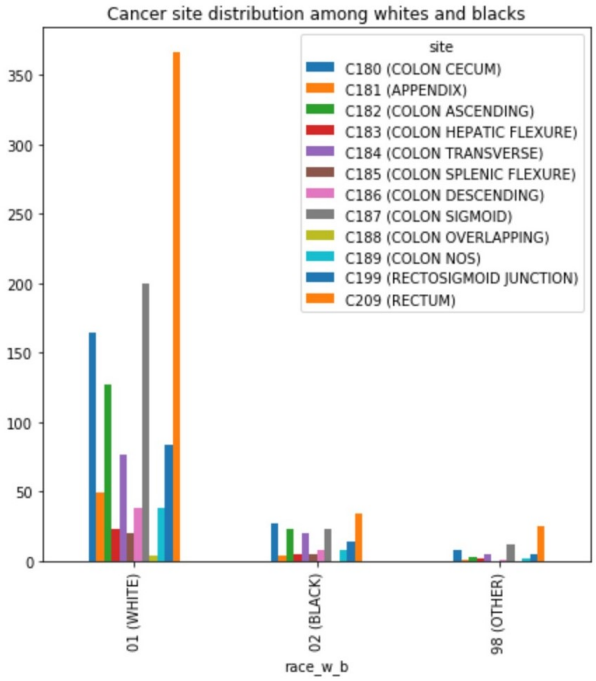


Figure 9: Distribution of cancer's sites for different races

4.5 Per-Patient Data Summarization

For each of the 94 patients in the cohort described in Section 3, we created a row of data in a summary spreadsheet. Effectively, each row tells the patient's story at a high level. Table 11 describes the columns in our spreadsheet.

Input for the first 7 entries were obtained from the Cancer Registry file. Patient gender was obtained from the Patient information file. The next set of rows in Table 11 characterize the chemotherapy treatments received by the patient, and the lab observations available for the patient for CEA and the three critical blood counts, WBC, PLT and HGB.

The last five rows describe the binary values computed to support the plotting of bar charts in our toxicity and response study presented in Section 4.2.

5 Conclusions

Based on this preliminary effort, we conclude that EHR data is definitely worth exploring to characterize inter- and intra-patient variability in pharmacokinetics (PK) of chemotherapy drugs. Clear trends are observed in this data of both patient toxicity and response to different dosage levels. Several lessons were learned. Given the highly manual nature of data extraction from EHR, a better partnership is needed between researchers and EHR data providers to reduce cycle time (data request, response, analysis to verify completeness of dataset, which invariably leads to another follow-up request). Our current dataset is missing imaging data, clinicians' progress notes, pathology and microbiology reports, procedures and diagnosis. We now have data for procedures and diagnosis from a recent Dec. 2019 extract, but data for the first four types are still unavailable. Genomic data will surely be needed for creating patient subsets for improving predictions. Finally, since the number of CRC patients at UVA is not large, we need to procure datasets from other health centers.

Column Name	Description
MRN	Medical Record Number of the patient
Age	Age of the patient
Dxdate	Date of diagnosis
CancerType	The broader name for the cancer
Grade	Grade of the cancer
cGrp	Clinical staging of the cancer's spread y
pGrp	Pathological staging of the cancer's spread
Gender	Gender of the patient
number_of_FOLFOX	Number of FOLFOX cycles administered to the patient
number_of_FOLFIRI	Number of FOLFIRI cycles administered to the patient
number_of_FU	Number of cycles in which the patient had just 5-FU infusions
date and week for first and last treatment	These fields have values for the first and last dates (and first and last weeks) for any type of chemotherapy, and then specifically for FOLFOX, FOLFIRI, just 5-FU, and Bevacizumab. Weeks were computed from the date of diagnosis
number_of_CEA	Number of CEA measurements available for the patient
date_first_CEA	Date of the first available CEA measurement
date_last_CEA	Date of the last available CEA measurement
CEA_end_value	Final CEA value recorded for the patient from the available data
CEA_end_high	Binary value: 1: if the final CEA was higher than the initial CEA; 0: otherwise
min and max values for WBC, HGB and PLT	The minimum and maximum observed value for each of these labs available for the patient during treatments. These values are used to determine if the patient had toxicity
dose_modifying_toxicity	Binary value: 1: if dose was lowered within the first 5 cycles due to toxicity in labs; 0: otherwise
initial_response	Binary value: 1: if there was initial response as defined in Section 4.2; 0: otherwise
redn_response	Binary value: 1: if there was response after dose reduction as defined in Section 4.2; 0: otherwise
dose_reduction	Binary value: 1: if the patient had dose reduction per the definition in Section 4.2 within the first 5 treatment cycles
dose_delay	Binary value: 1: if the patient had dose delay per the definition in Section 4.2 within the first 5 treatment cycles

Table 11: Summary information for each patient

6 Acknowledgment

We thank the UVA Engineering-in-Medicine (EIM) program for funding this work. We thank Ron Grider, Steve Patterson, Glenn Wasson and Joyce Miller for helping us procure the required datasets.

References

[1] A. Noone, N. Howlader, M. Krapcho, D. Miller, A. Brest, M. Yu, J. Ruhl, Z. Tatalovich, A. Mariotto, D. Lewis, H. Chen, E. Feuer, and K. e. Cronin, *SEER Cancer Statistics Review, 1975-2015, National Cancer Institute. Bethesda, MD, based on November 2017 SEER data submission.* April 2018; available at https://seer.cancer.gov/archive/csr/1975_2015/results_merged/sect_01_overview.pdf.

[2] National Comprehensive Cancer Network, *Colon Cancer (Version 4.2018).* available at http://www.nccn.org/professionals/physician_gls/pdf/colon.pdf.

[3] T. Andre, C. Boni, L. Mounedji-Boudiaf, M. Navarro, J. Tabernero, T. Hickish, C. Topham, M. Zaninelli, P. Clingan, J. Bridgewater, I. Tabah-Fisch, and A. de Gramont, "Oxaliplatin, fluorouracil, and leucovorin as adjuvant treatment for colon cancer," *N. Engl. J. Med.*, vol. 350, pp. 2343–2351, Jun 2004.

[4] A. de Gramont, A. Figer, M. Seymour, M. Homerin, A. Hmissi, J. Cassidy, C. Boni, H. Cortes-Funes, A. Cervantes, G. Freyer, D. Papamichael, N. Le Bail, C. Louvet, D. Hendler, F. de Braud, C. Wilson, F. Morvan, and A. Bonetti, "Leucovorin and fluorouracil with or without oxaliplatin as first-line treatment in advanced colorectal cancer," *J. Clin. Oncol.*, vol. 18, pp. 2938–2947, Aug 2000.

[5] M. W. Saif, A. Choma, S. J. Salamone, and E. Chu, "Pharmacokinetically guided dose adjustment of 5-fluorouracil: a rational approach to improving therapeutic outcomes," *J. Natl. Cancer Inst.*, vol. 101, pp. 1543–1552, Nov 2009.

[6] J. J. Lee, J. H. Beumer, and E. Chu, "Therapeutic drug monitoring of 5-fluorouracil," *Cancer Chemother. Pharmacol.*, vol. 78, pp. 447–464, 09 2016.

[7] C. Onesti, A. Botticelli, M. la Torre, M. Borro, G. Gentile, A. Romiti, L. Lionetto, A. Petremolo, M. Occhipinti, M. Roberto, R. Falcone, M. Simmaco, P. Marchetti, and F. Mazzuca, "5-fluorouracil degradation rate could predict toxicity in stages ii–iii colorectal cancer patients undergoing adjuvant folfox," *Anti-Cancer Drugs*, 11 2016.

[8] B. Glimelius, H. Garmo, A. Berglund, L. A. Fredriksson, M. Berglund, H. Kohnke, P. Bystrom, H. Sorbye, and M. Wadelius, "Prediction of Irinotecan and 5-Fluorouracil toxicity and response in patients with advanced colorectal cancer," *Pharmacogenomics J.*, vol. 11, pp. 61–71, Feb 2011.