Characterizing the “Catchment Area” for Wilmot Cancer Institute

Mustafa Ali and Erika Ramsdale, MD
CSC/DSC 240/440 Final Project Report

INTRODUCTION AND BACKGROUND

The Wilmot Cancer Institute (WCI) is the largest cancer care provider in upstate New York, serving more than 3.4 million people in Monroe county and a 27-county regional catchment area. The clinical infrastructure includes the state-of-the-art Wilmot Cancer Center at the University of Rochester Medical Center, as well as 12 regional clinic sites across the catchment area (Figure 1). In addition to providing clinical cancer care, WCI provides access to cutting-edge clinical trials for patients across the region, and supports well-established and renowned research programs in basic science, translational and clinical research, and cancer control.

In mid-2021, WCI will be submitting a large grant to the National Cancer Institute (NCI), with the aim of becoming an NCI-Designated Cancer Center. Currently, 71 other centers in the United States have this designation. In order to be considered for this designation, a cancer center must demonstrate scientific leadership in clinical and basic science research, train medical researchers, and have significant impacts on the health and well-being of its surrounding communities and the public at large.[1] The demonstration of community outreach and impact is particularly critical, as this is a characteristic that distinguishes WCI from the 7 other NCI Cancer Centers within New York.

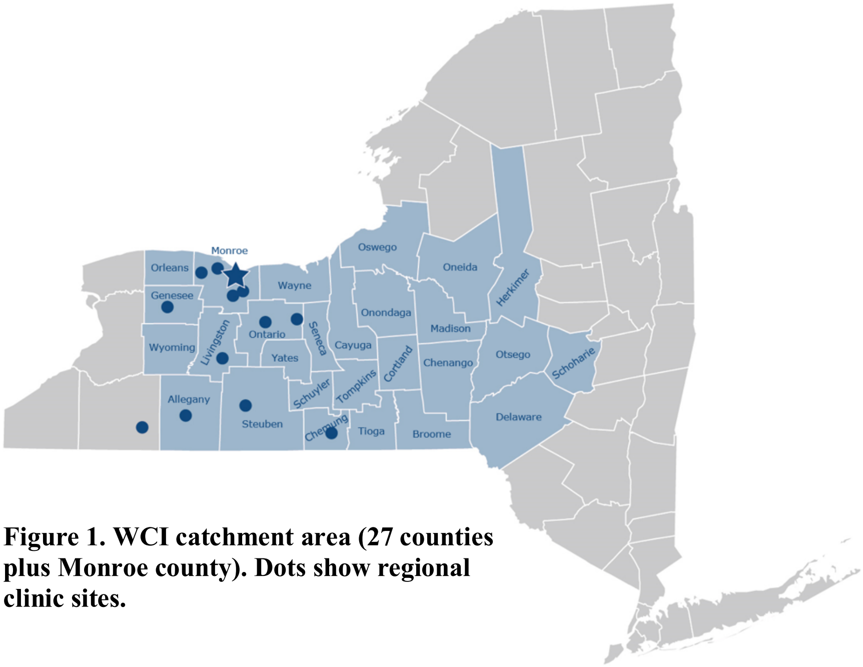


Figure 1. WCI catchment area (27 counties plus Monroe county). Dots show regional clinic sites.

In order to develop conceptual models and interventions for community outreach and impact, a deeper understanding is needed regarding the characteristics of patients living within WCI’s regional catchment area. This has not been closely examined previously due to the lack of organized, relevant datasets (as all data was housed within the electronic medical record [EMR], historically requiring an extensive degree of manual data capture). However, in recent months WCI has hired an informatics team to build and implement a federated database system capturing relevant patient data from the EMRs of all WCI sites, as well as data from multiple other sources (including the Clinical Trials Office [CTO], New York State Cancer Registry, and the United States Census Bureau). Although an infrastructure now exists to extract, merge, organize, and present data, extensive mining and analysis of datasets generated from this infrastructure have not yet been conducted; this project therefore represents a test case for the developing informatics infrastructure, requiring attention to data and process validation, as well as an opportunity to deliver exploratory, hypothesis-generating data to inform the NCI designation grant activities and WCI community outreach.

Based on discussions with WCI leadership, several questions were generated to guide this analysis. The primary questions examined in this project are:

1. Who are the patients coming to us for cancer care from outside Monroe County?
2. Which patients from the regional catchment area are enrolling in clinical trials?
3. For which patients are we missing opportunities for outreach (including cancer screening and clinical trial treatment options)?

METHODS

Datasets and Data Acquisition

This analysis required the merging of several datasets, in collaboration with the WCI informatics team. This new informatics team currently consists of 3 full-time staff; since 2019, a member of this project team (Ramsdale) has led the Data Analytics Working Group within WCI, which supports the development of informatics infrastructure, collaborates closely with the informatics team, and provides clinical content expertise; the current project is an extension of this "real-world" collaboration between clinicians and software developers and programmers. Beginning from a single database, over the last year (and even since the inception of this project in March 2020) the data architecture has grown significantly (Figure 2).

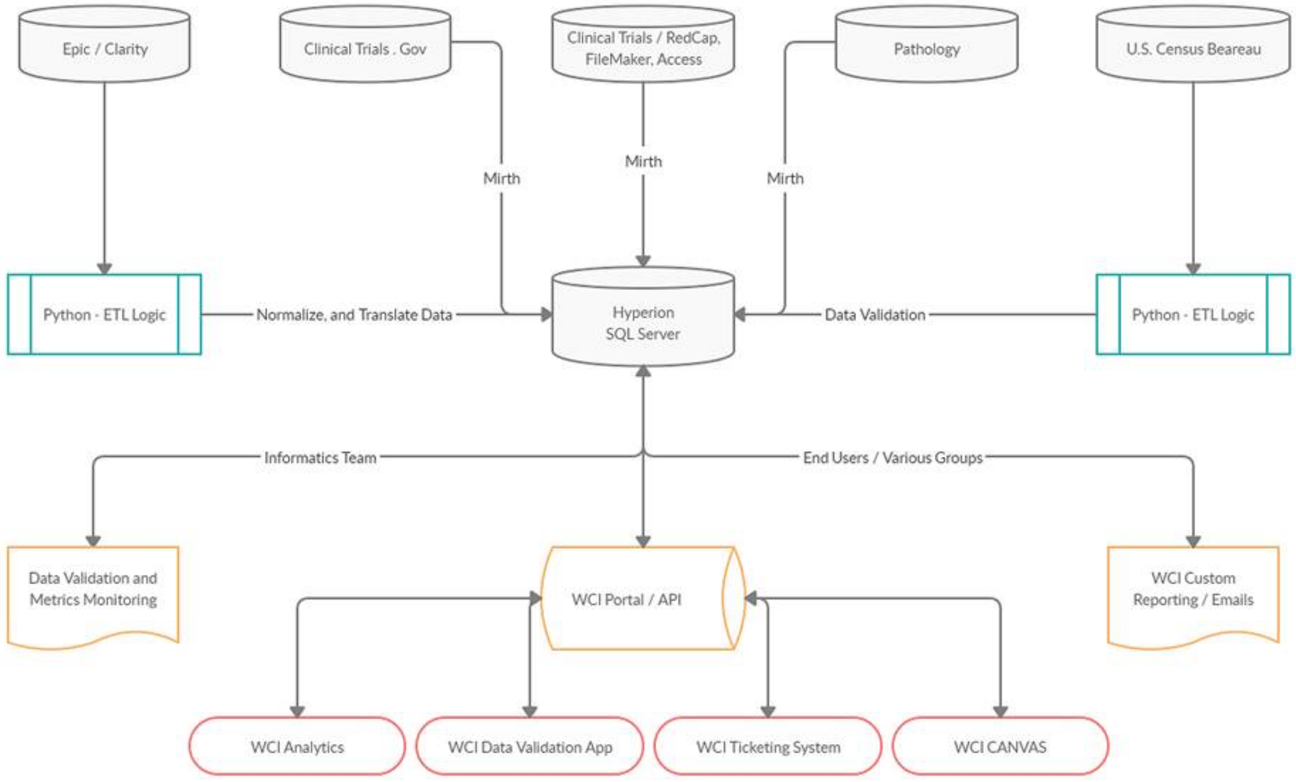


Figure 2. Current iteration of WCI data architecture.

The following primary data sources were utilized to generate the dataset for analysis.

1. Epic Clarity Database: Clarity is an analytical SQL database containing >12,000 tables and >125,000 columns, updated nightly from the electronic medical record (EMR) transactional database. From Clarity, data specific to patients with cancer are sent to the Hyperion SQL server, housed on WCI servers, which at the time of initiation of this project contained 3-4 tables of demographic, encounter, diagnostic, and treatment data. Restricting the time interval from 2014 – 2019, a total of 35,129 unique patients are included in the database, with 10,300 of these patients enrolled on clinical trials. For the year 2018 – 2019, there are 714,000 clinical encounters captured in the database. The project team had direct access to the Hyperion SQL database as well as the assistance of the informatics team for data queries. The features selected for inclusion in this analysis are:

a. Age/DOB;
b. Gender;
c. Race;
d. Physical location (address and zip code);
e. Primary cancer diagnosis (ICD-10 code);
f. Cancer stage;
g. Site of care;
h. Tobacco use.

The index for this dataset is the patient's medical record number (MRN), a unique patient identifier used throughout the EMR as well as in the Hyperion database.

2. OncoLog database: This database is utilized by WCI to report incident cancer cases to the New York State Cancer Registry. This database was used to supplement diagnosis codes and staging data for our dataset. ICD-10 codes from Clarity (indicating primary cancer diagnosis) are input by the care provider for billing purposes, and accuracy is low. Staging data from Epic, which is input by physicians, are available for <30% of cases. The OncoLog database input is from a trained registrar who reviews charts and manually enters data for reporting to cancer registries.

3. Via Oncology database: This database is linked to a treatment decision tool utilized by WCI oncologists, and has supplemental staging data.

4. Clinical Trials Office (CTO) database: This database yielded an additional feature for analysis, participation in clinical trial (yes/no).

5. New York State Cancer Registry: This is a publicly accessible cancer registry housing county-level cancer-specific incidence data.[2] Cancers of particular interest for this project are those which are common and for which effective screening measures exist:

a. Breast;
b. Colorectal;
c. Lung;
d. Prostate.

6. United States Census Bureau: This data source was used to generate census tract data which was partially cross-walked with patient address data by the WCI informatics team; this crosswalk is not yet complete.

7. United States Department of Agriculture Rural/Urban Commuting Area (RUCA) Codes: These codes "classify U.S. census tracts using measures of population density, urbanization, and daily commuting."[3] The whole number primary codes (range 1-10) were used for this project.

8. Centers for Medicare & Medicaid Services (CMS) 2020 ICD-10-CM. This dataset contains the current International Classification of Diseases, 10th revision (ICD-10) codes used by insurers including CMS.[4] ICD-10 codes, developed and maintained by the World Health Organization (WHO), encode specific information about diagnosis, laterality, chronicity/phase, and severity of disease.

9. Github: Zipcode-level GeoJSON files for New York state were located within a Github repository, found at https://github.com/OpenDataDE/State-zip-code-GeoJSON.

Feature Selection and Data Pre-processing

Feature selection: Feature selection for this project was guided by clinical expertise, and based upon the initial project timeframe and specific questions under consideration (see Introduction and Background). Other features considered for selection were ethnicity and comorbidities (health problems other than cancer). Ethnicity data was not initially captured in the dataset (as separate from race), and comorbidities were deemed too challenging and time-intensive to clean within the project timeline. Both of these features will be considered in future analysis.

General Pre-processing (Figure 3): Data were reported for all patient encounters within WCI or regional sites, for patients living outside Monroe County, from 3/26/2018 to 3/25/2020. From this initial dataset, we excluded patients with out-of-state (including international) zip codes, as well as patients with zip codes inside Monroe County. We then collapsed duplicate patient rows in a multi-step process. First, exact duplicate rows were eliminated. Then, rows containing duplicate MRN's were visually inspected. The "site of care" feature was eliminated as it was frequently discordant between rows and was not expected to yield helpful analytic input. The remaining discordant features between duplicate MRN rows were manually corrected based on reference to patient charts.

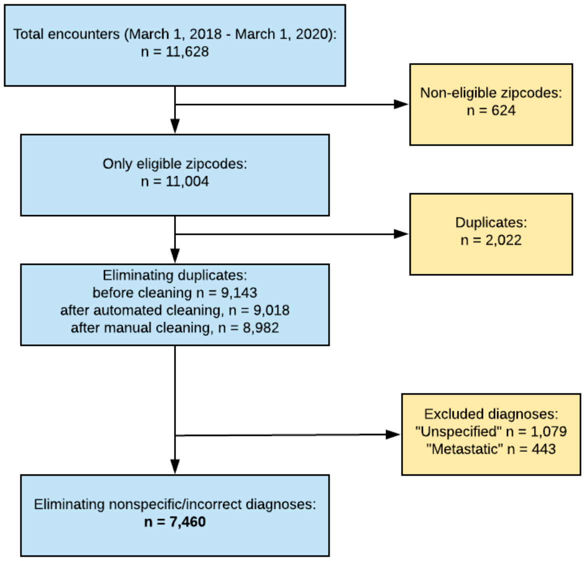


Figure 3. Initial data pre-processing steps, with number of rows at each step.

Next, a “cancer description” column was created by mapping collapsed categories of ICD-10 codes to yield a general diagnostic term (e.g., “breast cancer”, “prostate cancer”). ICD-10 is a hierarchical classification of disease; we created a mapping based on the first three digits of the ICD-10 code representing the level of general diagnosis, without modifiers for laterality, chronicity, or severity (e.g., for “C50.4” representing “malignant neoplasm of upper-outer quadrant of breast,” only “C50” was used to map onto the general diagnosis of “breast cancer”). During this mapping, it was noted that a large number of cases (n = 1,522) mapped onto codes which were not included in ICD-10 (C42) or were nonspecific codes for metastatic cancer (C77). Examining a random sample of charts, these codes appeared to correspond largely to patient cases of hematopoietic malignancies such as leukemia and lymphoma. Given this systematic coding error, prohibiting accurate classification of these cases, we eliminated them from most analyses and proceeded with analysis on only solid tumor (i.e., cancers not arising from the hematopoietic/reticuloendothelial system) cases, which appeared to be accurately classified based on chart validation of a random sample. For cancer stage, a Python script automated conversion from more complex pathologic and clinical staging classifications to a simple 5-level classification (Stage 0-4).

Using address data, the WCI informatics team wrote scripts to automate cross-walks between zip code, census tract, and county assignments, which data they provided to us where available. This cross-walk was incomplete during our project timeline and is currently undergoing further development and validation. For those patients with complete data, we created Federal Information Processing Series (FIPS) codes for each patient by concatenating state, county, and census tract codes; FIPS codes were then mapped to primary RUCA codes. RUCA code was used as an additional feature in analyses.

Pre-processing for Classification and Clustering

To prepare the dataset for classification and clustering algorithms, we first eliminated irrelevant features from the dataset, and retained the following features: date of birth (DOB), gender, race, cancer diagnosis, cancer stage, tobacco use, RUCA code, and clinical trial participation. Age was added to the dataset as both a continuous and categorical (18-40, 40-60, 60-80, and 80+) variable. Race was collapsed into 3 categories (White, Nonwhite, and Unknown) given very small numbers of non-white patients.

For classification, clinical trial participation was used as the class label. For all models, only cases with the 12 most common cancer diagnoses were included. Stage, gender and RUCA codes were label encoded, and only rows with non-missing RUCA codes were retained. Tobacco use, cancer diagnosis, and race were one-hot encoded. The dataset was then split into a training set (80%) and test set (20%). Figure 4 shows the significant class imbalance in this dataset. To address this, oversampling was performed on the training set using random sampling with replacement.[5]

For clustering, the same features were used as in classification, except for cancer stage, which was eliminated due to large number of missing values. Age and RUCA were used as continuous variables with min/max normalization. Cancer diagnosis was collapsed into 4 categories: Gastrointestinal (GI), comprising esophageal, gastric, colorectal, and pancreatic cancers; Breast; Respiratory, comprising head and neck and lung cancers, and Genitourinary (GU), comprising prostate and urothelial/bladder cancers. This diagnosis grouping was chosen to mirror the disease specializations of WCI oncologists for common cancers.

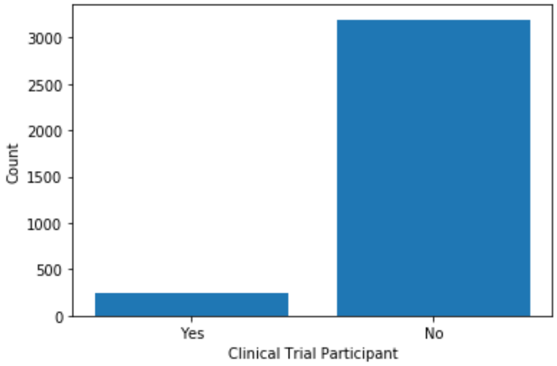


Figure 4. Dataset class imbalance.

Classification Approach

Classification models used features of age, gender, race, cancer stage, tobacco use, cancer diagnosis, and RUCA codes to predict clinical trial participation in the dataset. Logistic regression,[6] random forest,[7] and linear support vector machine (SVM)[8] were compared using confusion matrices and receiver operating characteristic (ROC) curves. Metrics including c-statistic/area under the curve (AUC), accuracy, recall rate, and precision rate were measured for each model. Feature importance plots were generated to visualize relative importance of features in model predictions. All analyses were conducted in Python using the pandas, scikit-learn, and matplotlib libraries.

Clustering Approach

Clustering algorithms were used for data exploration and hypothesis generation, using the features age, gender, race, tobacco use, cancer type, RUCA, and clinical trial participation. Cancer stage was not utilized due to a high percentage (36%) of data missingness not at random, precluding imputation.[9] Clustering algorithms used were k-means using Euclidean distance,[10] agglomerative hierarchical clustering[11] using Ward's method,[12] DBSCAN,[13] OPTICS,[14] and spectral clustering.[15] As ground truth is not available for this application, intrinsic validation measures were used for each method, including the silhouette coefficient,[11] the Calinkski-Harabasz Index,[16] and the Davies-Bouldin Index.[17] For algorithms requiring a prespecified number of clusters such as k-means, the "elbow method"[18] was also used to visually determine the optimal number of clusters. Principal component analysis (PCA)[19] and t-distributed Stochastic Neighbor Embedding (t-SNE)[20] were used as feature reduction techniques in order to plot and visually inspect cluster assignments in 2 dimensions. All analyses were conducted in Python using the scikit-learn, pandas, numpy, and matplotlib libraries, or using the Weka software (downloaded at https://www.cs.waikato.ac.nz/ml/weka/). Maximal itemset mining on clusters was performed using Apriori[21] and FPMax[22] algorithms from the Python mlxtend library.

To examine geospatial distribution, we generated maps using the Python folium library and zipcode-level GeoJSON files downloaded from a github repository. For the most common cancer diagnosis (breast, lung, colorectal, and prostate cancers) we created a geographic heatmap of cancer cases normalized by cancer incidence rates by zipcode. This allows visualization of WCI "market share" for a particular zipcode.

RESULTS

Descriptive Data

In our dataset, the median age was 68 years (mean 66.7 years, range 18-101, Figure 5), 54.1% were female, and 97.3% were white. In terms of tobacco use, 56.2% are former smokers, 41.0% are never smokers, 1% are current smokers, and 1.8% have an unknown tobacco use history. The most common cancer types are given in Figure 6; of note, this figure shows the high proportion of "unspecified" and "metastatic" diagnoses which were excluded from some analyses. Figure 7 shows cancer stage distribution, with >1/3 of patients having missing stage information. Stage 0 represents non-invasive cancer, or carcinoma in situ.

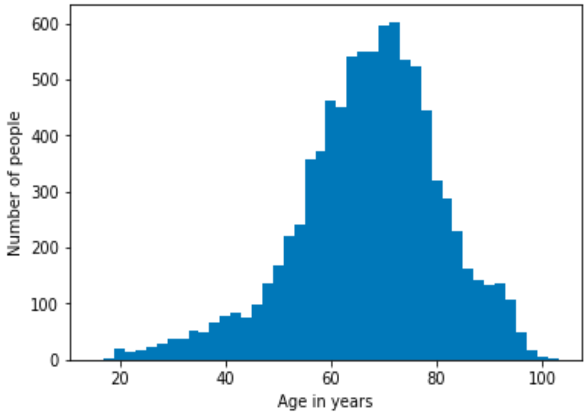


Figure 5. Age histogram.

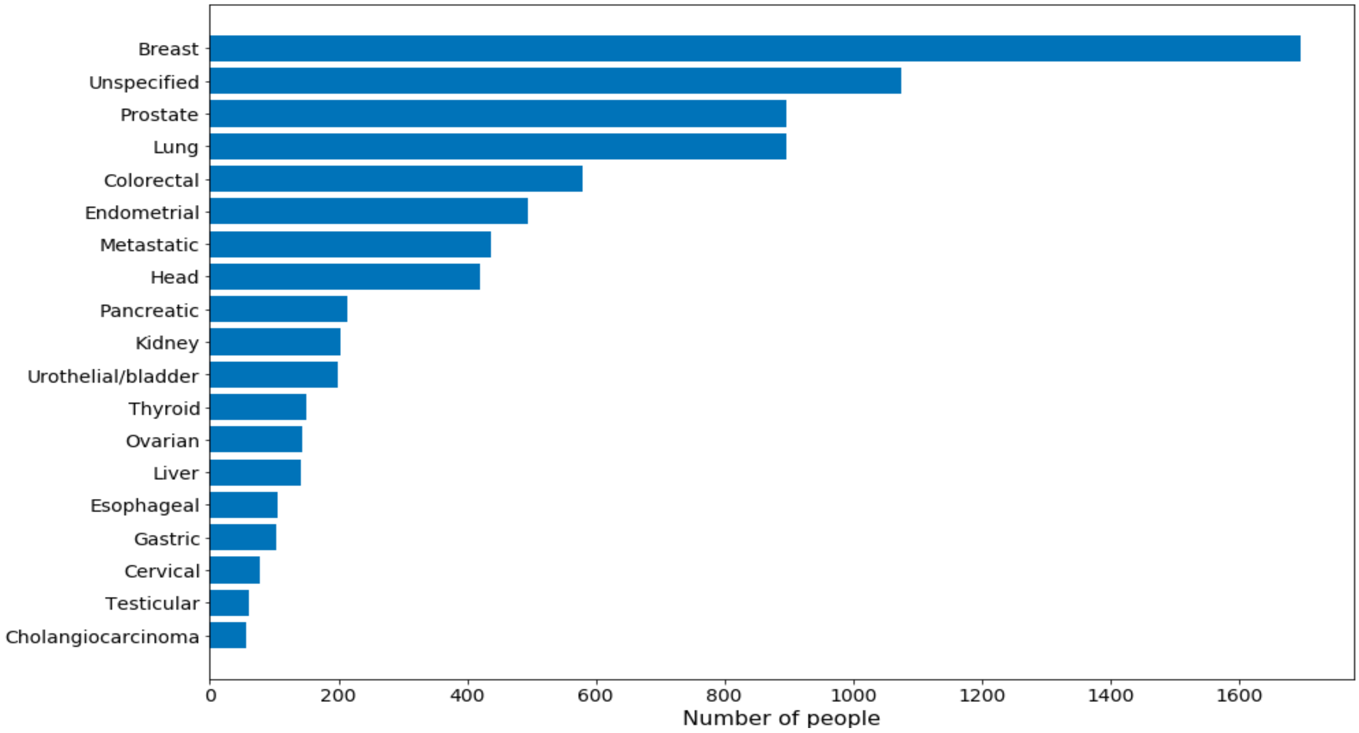


Figure 6. Cancer diagnoses in dataset.

Most patients in the dataset lived in more urban locales, despite being in a more rural part of the state compared to Monroe County (Figure 8). A RUCA code of 1 denotes a “metropolitan area core” with commuting flow primarily within the urban area, whereas 10 represents a rural area with most commuting flow to a census tract outside of an urban area or urban cluster;[3] a higher RUCA code therefore indicates lower population density and, in general, longer commuting times to reach urban areas.

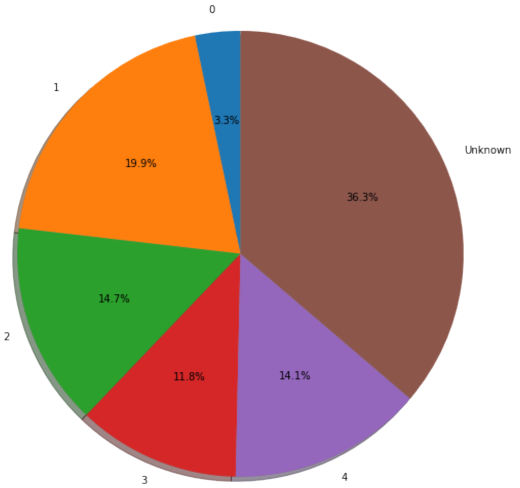


Figure 7. Cancer stage distribution.

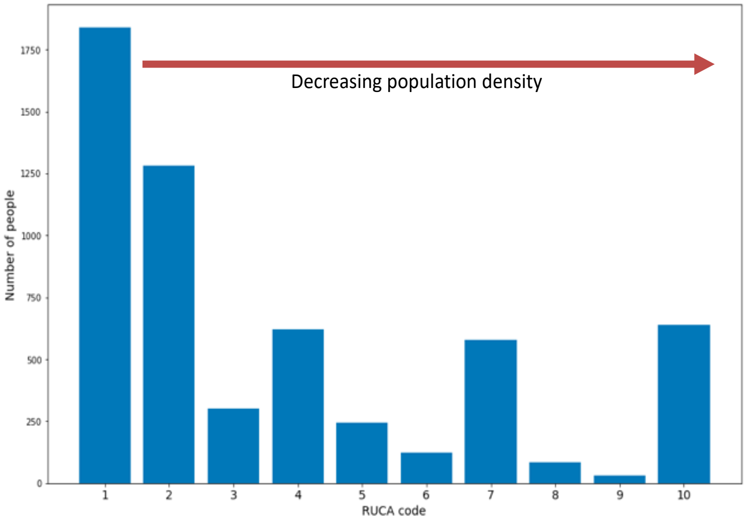


Figure 8. RUCA codes.

Classification

Our initial classification models included the top 12 most common cancer diagnoses in the dataset (see Figure 6). In these models, logistic regression and linear SVM performed equally well, while the random forest classifier had equal precision but lower accuracy (Table 1). The random forest classifier has a slightly higher specificity (0.9 vs 0.84) but a much lower sensitivity/recall (0.41 vs 0.7). Examining the feature importance for the random classifier model (Figure 9, left) as well as the coefficients in the logistic regression model, metastatic and unspecified cancers made significant contributions to the model performance. In the random forest model, age, RUCA code, and cancer stage made the largest contributions to the model. However, in the logistic models, the largest beta coefficients were for the metastatic and unspecified cancer diagnoses.

Table 1. Classifier performance measures.

	With all diagnoses			Excluding unspec. and metastatic		
Measure	Logistic Regression	Random Forest	SVM	Logistic Regression	Random Forest	SVM
Accuracy	0.83	0.83	0.83	0.64	0.88	0.68
Recall	0.7	0.41	0.7	0.55	0.1	0.47
Precision	0.45	0.43	0.45	0.13	0.2	0.13
AUC	0.818	0.746	0.765	0.603	0.608	0.614

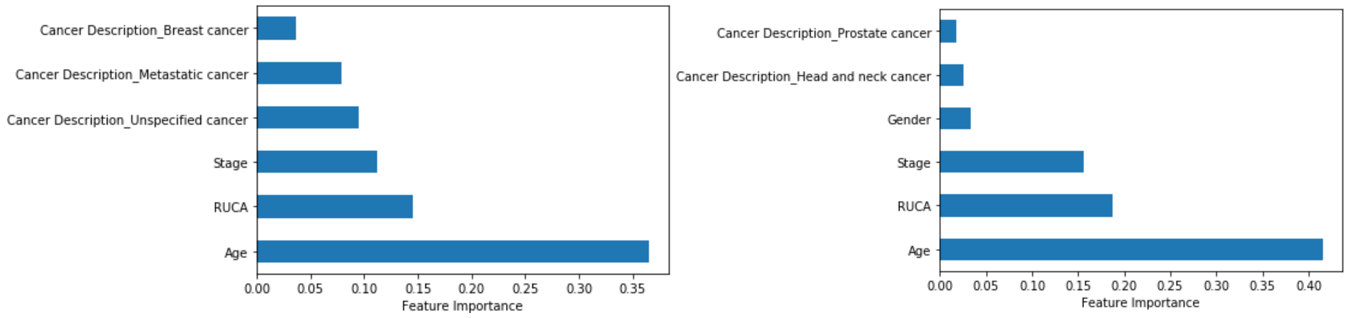


Figure 9. Random forest classifier top 6 most important features: with unspecified and metastatic diagnosis included (left) and without (right).

As manual data validation confirmed a systematic coding error for hematopoietic tumors, the tumors classified as “unspecified” and “metastatic” were eliminated from the dataset. Unfortunately, as shown in Table 1, this significantly decreased model metrics across all three classifiers. In these models, the logistic regression model had the highest sensitivity/recall, but significantly worse than previously at 0.55. The ROC curves for the logistic regression models are shown in Figure 10, illustrating the deterioration in model performance. The model with excluded unspecified/metastatic diagnosis did not perform much better than random assignment at predicting clinical trial enrollment in the test set.

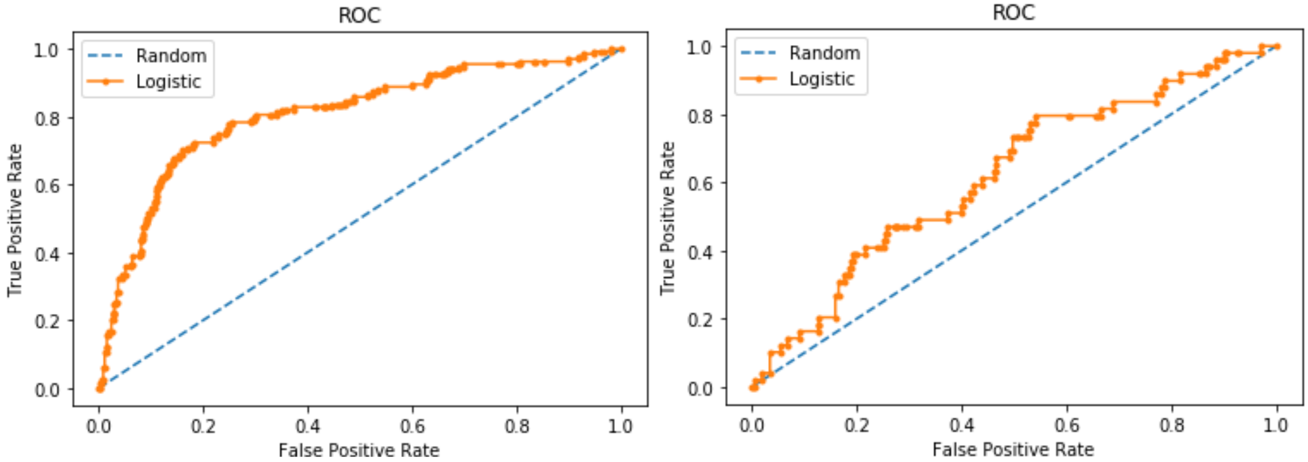


Figure 10. ROC curves: with unspecified and metastatic diagnosis included (left) and without (right).

Clustering and Frequent Itemset Mining

For exploratory analysis of underlying data structure, several clustering algorithms were applied and compared using the following features: age, gender, race, tobacco use, cancer type, RUCA, and clinical trial participation. Table 2 summarizes key features and internal validation measures for each algorithm; ground truth labels are not available for this analysis. For all algorithms requiring parameter specification, number of clusters (or MinPts, in the case of DBSCAN) were varied and internal validation measures plotted for each parameter value. Of note, varying epsilon for DBSCAN did not substantially alter the validation measures. "Optimal" numbers of clusters were not chosen based on best values for the validation measures, but rather by selecting an "elbow" point based on visual inspection of these plots. An example of these plots for the k-means algorithm is shown in Figure 11.

Table 2. Clustering algorithm comparison.

Measure	K-means	DBSCAN	OPTICS	Agglomerative	Spectral
optimal # of clusters	10	12	7	8	10
# of noise points	0	357	779	0	0
Silhouette coefficient	0.584	0.613	0.496	0.472	0.562
Calinski-Harabasz Score	1536.2	1062	940.5	1051.6	1425.9
Davies-Bouldin Score	0.952	1.19	1.23	1.15	0.955

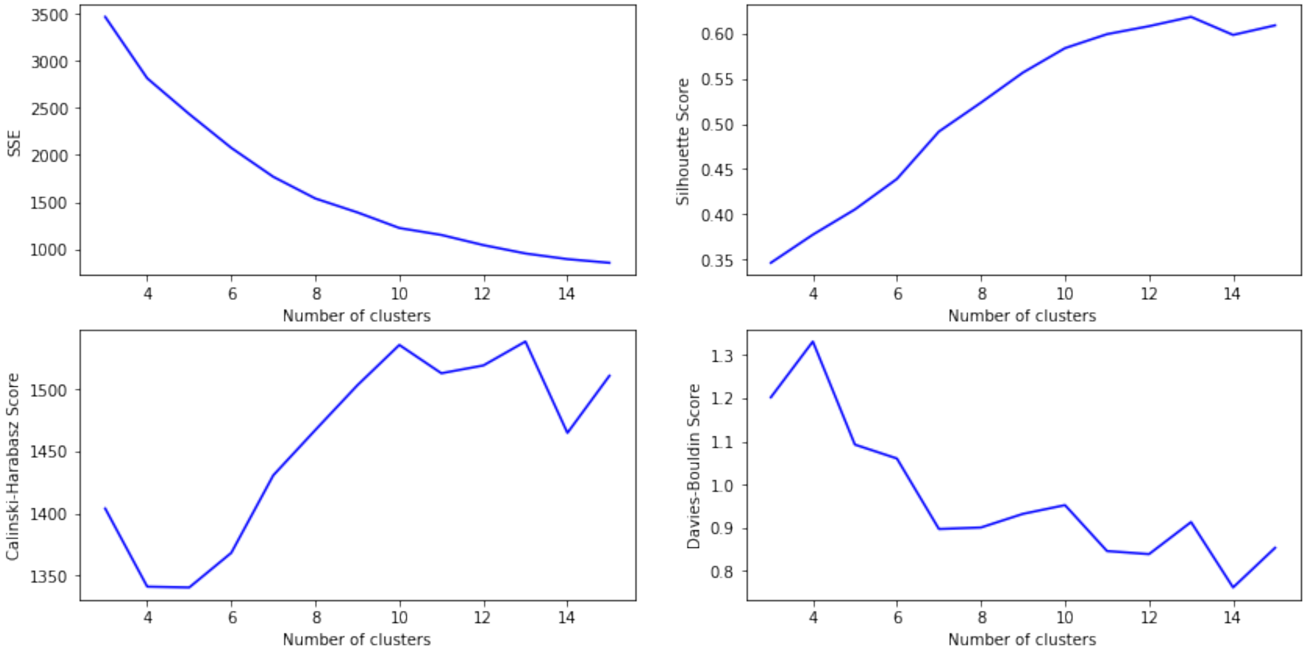
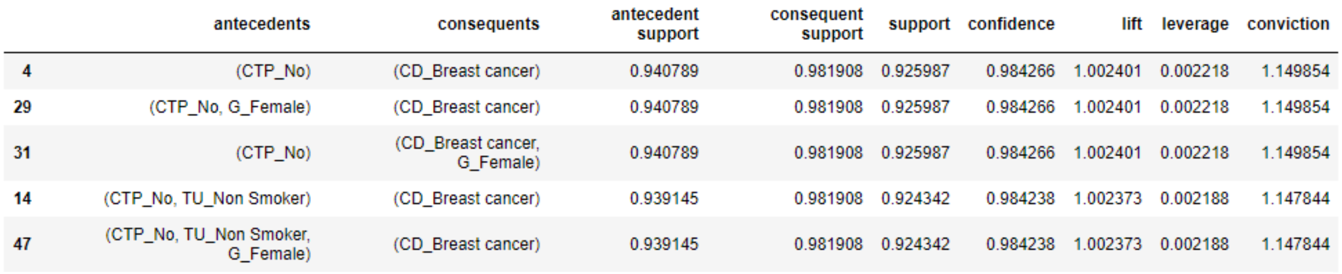


Figure 11. Internal validation measures for k-means clusterer, by number of clusters.

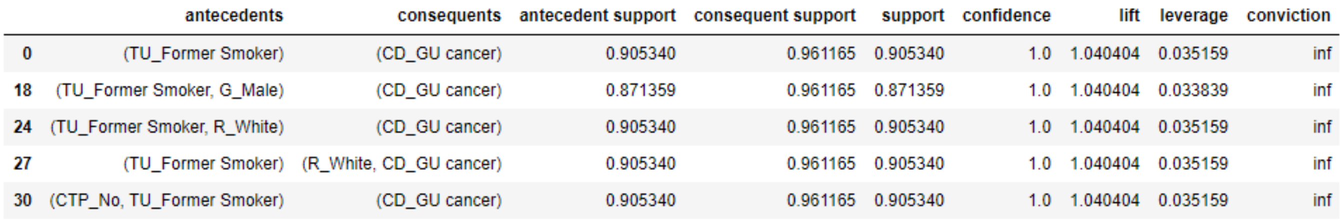
Inspection of the cluster characteristics was hampered by high number of clusters, but some interesting patterns could be gleaned from the data. In the k-means output, for example, 3 of the 4 clusters with the highest clinical trial participation had a significantly higher proportions of non-white patients, and predominantly GI cancers. In 2 respiratory cancer clusters, one represented 76 male patients with a lower than average age, markedly higher than average current smoking rate, high clinical trial participation, and more rural than average address; the other represented 291 female patients at higher age from more urban than average areas, lower rates of current smoking, and low rates of clinical trial participation. There was not a significant qualitative difference in insight gleaned from algorithms defining spheroid clusters (such as k-means) and those which can accommodate non-spherical cluster shapes (such as DBSCAN, OPTICS, and spectral clustering).

In order to try to systematically investigate clusters, frequent itemset mining was applied to clusters using Apriori and FPMax algorithms to yield closed and maximal itemsets. Examples of output are given as Figures 12 and 13. Overall, these confirmed but did not add to the insights gleaned from inspection and comparison of cluster proportions.



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
4	(CTP_No)	(CD_Breast cancer)	0.940789	0.981908	0.925987	0.984266	1.002401	0.002218	1.149854
29	(CTP_No, G_Female)	(CD_Breast cancer)	0.940789	0.981908	0.925987	0.984266	1.002401	0.002218	1.149854
31	(CTP_No)	(CD_Breast cancer, G_Female)	0.940789	0.981908	0.925987	0.984266	1.002401	0.002218	1.149854
14	(CTP_No, TU_Non Smoker)	(CD_Breast cancer)	0.939145	0.981908	0.924342	0.984238	1.002373	0.002188	1.147844
47	(CTP_No, TU_Non Smoker, G_Female)	(CD_Breast cancer)	0.939145	0.981908	0.924342	0.984238	1.002373	0.002188	1.147844

Figure 12. K-Means (Cluster #2).



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(TU_Former Smoker)	(CD_GU cancer)	0.905340	0.961165	0.905340	1.0	1.040404	0.035159	inf
18	(TU_Former Smoker, G_Male)	(CD_GU cancer)	0.871359	0.961165	0.871359	1.0	1.040404	0.033839	inf
24	(TU_Former Smoker, R_White)	(CD_GU cancer)	0.905340	0.961165	0.905340	1.0	1.040404	0.035159	inf
27	(TU_Former Smoker)	(R_White, CD_GU cancer)	0.905340	0.961165	0.905340	1.0	1.040404	0.035159	inf
30	(CTP_No, TU_Former Smoker)	(CD_GU cancer)	0.905340	0.961165	0.905340	1.0	1.040404	0.035159	inf

Figure 13. Agglomerative (Cluster #6).

In order to further evaluate and visualize clusters, feature reduction was performed and plotted using principal component analysis (PCA), with two features, and t-distributed Stochastic Neighbor Embedding (t-SNE). For PCA, the first 2 principal components explain 59% of the variance, and the first 3 explain 70%. Mapping cluster labels onto a 2-dimensional plot of PCA components shows reasonable clustering for most algorithms. Figure 14 shows this plot for DBSCAN with minPts=30, showing good clustering but many noise points (labeled "-1"). Figure 15 shows a t-SNE plot for k-means, using a perplexity of 40 and 300 iterations.

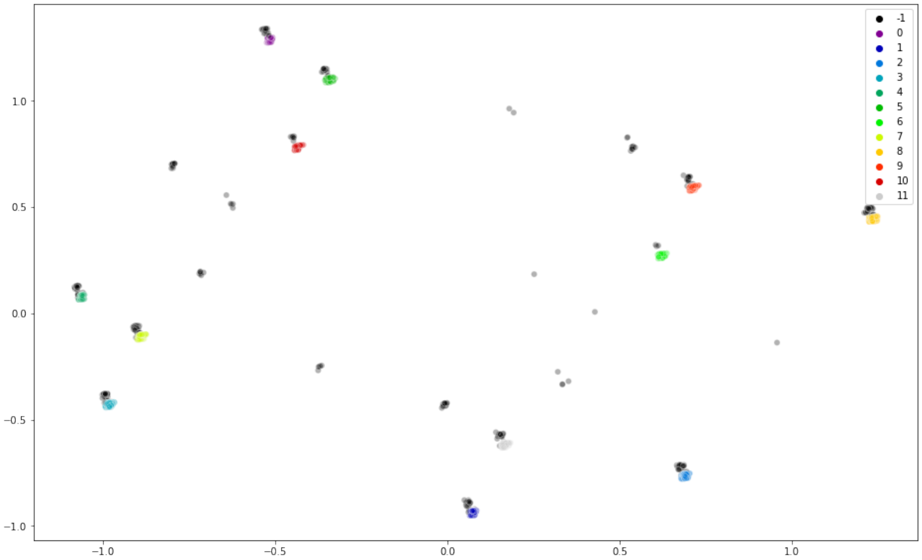


Figure 14. Two-dimensional PCA decomposition showing DBSCAN clusters.

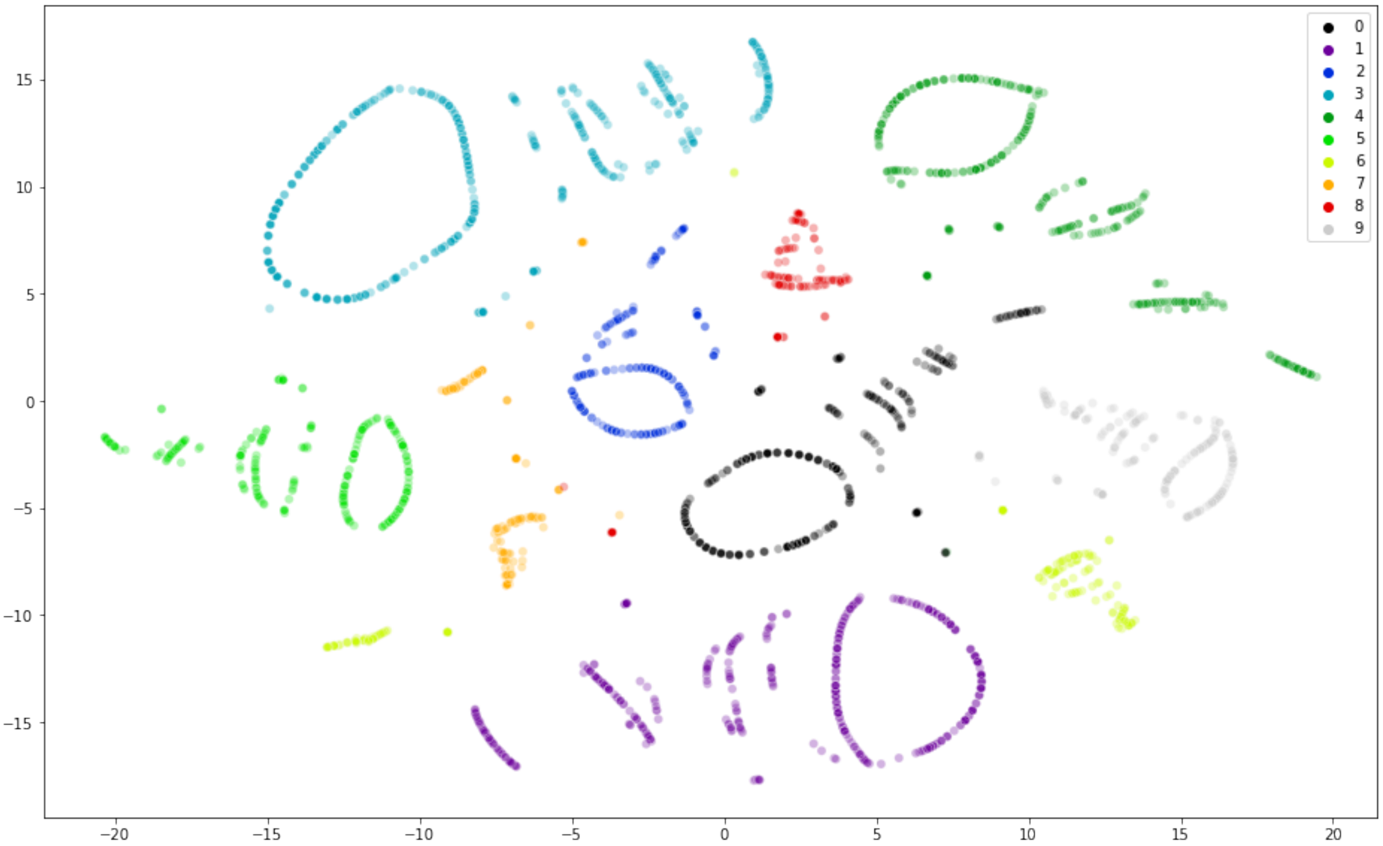


Figure 15. t-SNE plot for k-means clusterer with 10 clusters.

Geographic Clustering and Visualization

For breast, colorectal, lung, prostate, and urinary (comprising kidney, bladder, and urethral) cancers, New York state maps showed market share – defined as number of cases divided by 2 times the annual incidence rate (as our data encompass two years). This overestimates the market share as we could not accurately determine which cases in our dataset were new diagnoses; date of diagnosis will be added as a feature in future analyses. However, these maps show a normalized case rate, which accounts for a surrogate measure of population density. As the zipcode-to-county crosswalk was not completed prior to our analysis, the map shows many grayed-out zipcodes indicating no incidence data available. This crosswalk will be completed within the next several weeks. The geographic heatmap for colorectal cancer is shown in Figure 16. Additional heatmaps are available in the project code.

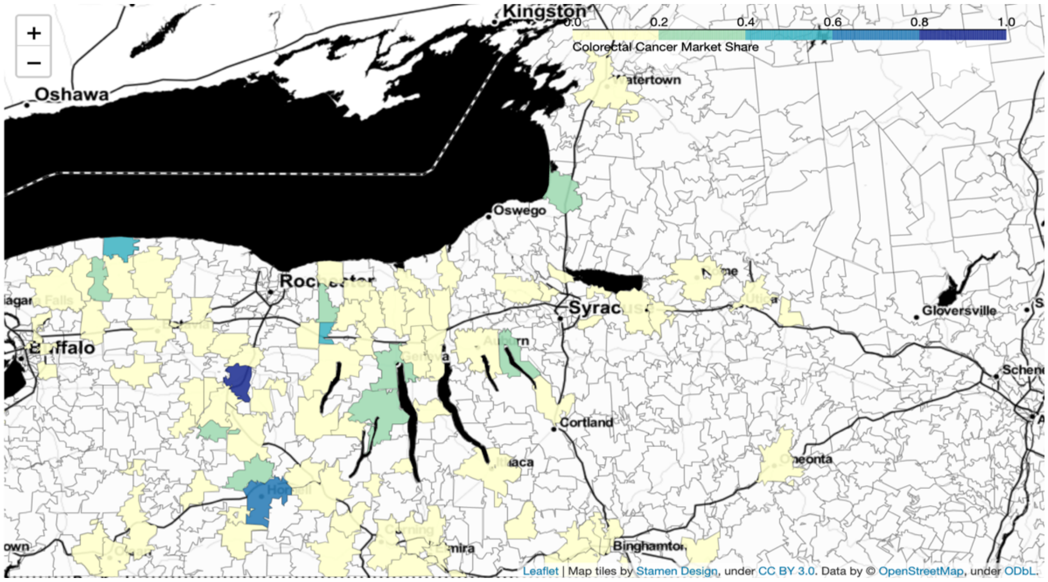


Figure 16. Colorectal cancer market share by zipcode (excluding Monroe County).

DISCUSSION

The recent, rapid progress in the availability and sophistication of clinical informatics architecture within WCI permits complex data queries which were previously untenable. The initial focus of queries, supported by this new architecture, is in support of NCI cancer center designation activities, with a full grant application due in early to mid 2021. This report describes a preliminary analysis of patient data from WCI's "catchment area," encompassing a broad swath of upstate New York outside of Monroe County; a deeper understanding of this population is needed to propose and implement interventions to improve cancer-related outcomes, which is a critical piece of the WCI's mission as well as a crucial component of the NCI designation grant.

About 9,000 individual patients were seen from the WCI catchment area over 2 years (not including patients from Monroe County or outside New York state). The mean age of this population (68 years) is similar to the mean age of patients diagnose with cancer nationally (66 years).[23] The left-skewed age distribution is also typical of national trends. This population is much more rural than the average U.S. population: nationally, about 84% of the population lives in urban areas (RUCA codes 1 and 2, approximately);[24] in this dataset only 54% do, and 20.4% live in the most rural areas (RUCA code 10). The most common cancers in this dataset are similar to national trends, with the notable over-representation of lung cancer and under-representation of colorectal cancer in our dataset, possibly reflecting referral patterns and higher access to colorectal specialists in the community (thus diverting these patients from WCI). Unfortunately, we were not able to accurately describe the prevalence and distribution of hematopoietic malignancies, such as leukemia and lymphoma, due to systematic coding errors. Communications with the New York State Cancer Registry ultimately revealed that the codes for these diagnoses, labeled "ICD-10" in database tables, were actually encoded using a different hierarchical classification scheme, the International Classification of Diseases – Oncology (ICD-O).[25] Additional data were requested from the registry in order to reconstruct accurate diagnostic categories for the dataset. Solid tumors (such as breast cancer, prostate cancer, etc.) are overlapping between ICD-O and ICD-10 and were unaffected.

Understanding clinical trial participation is an important objective for an academic cancer center, and a key component of NCI designation activities. Within the 2-year catchment area data, only about 14% of the patients participated in clinical trials, and this percentage dropped to 7% when considering only solid tumor diagnoses. We evaluated classification models to predict clinical trial participation, but these did not perform well, particularly when only solid tumors were considered. Although age, gender, and RUCA codes were important predictive features, these were outweighed by diagnostic categories, with the highest weight in categories for which there are more trials, which initially seems tautological but possibly supports the notion that patients are willing to travel if trials are available for their disease. However, we did not draw firm conclusions from this classifier, and further work will be needed to address class imbalance in addition to data accuracy. Techniques such as asymmetric bagging[26] or SMOTE[27] could be evaluated to address the significant class imbalance.

Clustering and maximal itemset mining on the clusters were used to develop insight into the overall structure of the dataset. In bioinformatics, clustering has most often been applied to genomic or molecular data.[28] Few studies exist applying clustering algorithms to complex clinical data. This preliminary analysis shows that application of clustering algorithms is feasible, with similar performance between the algorithms tested in terms of internal validation measures. Further work will be done to more directly compare the cluster labels, to see if the algorithms identify similar clusters. Subspace clustering methods may also be useful to apply to this dataset. Insight gleaned from clustering models will be considered hypothesis-generating for the purposes of further analysis or intervention development. Inclusion of geographical features (such as physical distance from WCI) in place of RUCA code could be considered, and will be feasible once geographic crosswalks are completed as discussed in the report.

CONCLUSIONS AND FUTURE DIRECTIONS

Efforts are ongoing to complete and validate the WCI catchment area dataset, including completing geographic crosswalks to permit multi-level (address, census tract, zipcode, county) analysis and visualization, correcting classifications of tumor diagnoses, and supplementing stage data. The primary deliverable of this first phase of the project is a set of Jupyter notebooks to pre-process the data, apply and measure the performance of supervised and unsupervised learning methods, and generate data visualizations. Though the project team developed familiarity with and significant insight into the dataset, results reported here should be considered exploratory to inform future work.

Further work on this project will proceed under the direction of Dr. Paula Cupertino, Associate Director of Community Outreach, Engagement and Disparities for WCI, and Dr. Charles Kamen, who leads the cancer control and community engagement effort for the NCI designation grant. Biweekly meetings are planned in collaboration with the WCI informatics team to further build on the work reported here, with the anticipation that Mustafa will be hired for an internship to continue his contributions toward this effort.

REFERENCES

1. “NCI Designated Cancer Centers.” Found at: https://www.cancer.gov/research/nci-role/cancer-centers. Accessed May 5, 2020.
2. “About New York State Public Access Cancer Epidemiology Data (NYSPACED).” Found at: https://www.health.ny.gov/statistics/cancer/registry/nyspaced/faq.htm. Accessed May 5, 2020.
3. “Rural-Urban Commuting Area Codes.” Found at: https://www.ers.usda.gov/data-products/rural-urban-commuting-area-codes/. Accessed May 4, 2020.
4. “2020 ICD-10-CM.” Found at: https://www.cms.gov/Medicare/Coding/ICD10/2020-ICD-10-CM. Accessed May 5, 2020.
5. Kotsiantis S, Kanellopoulos D, Pintelas P. Handling imbalanced datasets: A review GESTS International Transactions on Computer Science and Engineering 2006;30.
6. Hosmer DW, Lemeshow S, Sturdivant RX. Applied logistic regression2013.
7. Breiman L. Random Forests. Mach Learn 2001;45:5-32.
8. Hsu C-w, Chang C-c, Lin C-J. A Practical Guide to Support Vector Classification Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. 2003.
9. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009;338:b2393.
10. Forgy EW. Cluster analysis of multivariate data : efficiency versus interpretability of classifications. Biometrics 1965;21:768-9.
11. Kaufman L, Rousseeuw PJ. Finding Groups in Data: An Introduction to Cluster Analysis: Wiley; 2009.
12. Ward JH. Hierarchical Grouping to Optimize an Objective Function. Journal of the American Statistical Association 1963;58:236-44.
13. Ester M, Kriegel H-P, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. Portland, Oregon: AAAI Press; 1996:226–31.
14. Ankerst M, Breunig MM, Kriegel H-P, Sander J. OPTICS: ordering points to identify the clustering structure. SIGMOD Rec 1999;28:49–60.
15. Yu SX, Shi J. Multiclass Spectral Clustering. Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2: IEEE Computer Society; 2003:313.
16. Caliński T, Harabasz J. A dendrite method for cluster analysis. Communications in Statistics 1974;3:1-27.
17. Davies DL, Bouldin DW. A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence 1979;PAMI-1:224-7.
18. Han J, Pei J, Kamber M. Data Mining: Concepts and Techniques: Elsevier Science; 2011.
19. Abdi H, Williams LJ. Principal component analysis. WIREs Comput Stat 2010;2:433–59.
20. van der Maaten LJP, Hinton GE. Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research 2008;9:2579 - 605.
21. Agrawal R, Srikant R. Fast Algorithms for Mining Association Rules in Large Databases. Proceedings of the 20th International Conference on Very Large Data Bases: Morgan Kaufmann Publishers Inc.; 1994:487–99.
22. Grahne G, Zhu J. High Performance Mining of Maximal Frequent Itemsets. 6th SIAM Int’l Workshop on High Performance Data Mining 2003.
23. “Age and Cancer Risk.” Found at: https://www.cancer.gov/about-cancer/causes-prevention/risk/age. Accessed May 5, 2020.
24. “U.S. Cities.” Found at: http://css.umich.edu/sites/default/files/US%20Cities_CSS09-06_e2019.pdf. Accessed May 5, 2020.
25. “International Classification of Diseases for Oncology, 3rd ed.” Found at: https://apps.who.int/iris/bitstream/handle/10665/96612/9789241548496_eng.pdf. Accessed on May 5, 2020.
26. Blagus R, Lusa L. Class prediction for high-dimensional class-imbalanced data. BMC Bioinformatics 2010;11:523.
27. Blagus R, Lusa L. SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics 2013;14:106.
28. Sorlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A 2001;98:10869-74.