MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220224

Using EHR Data to Identify Social Determinants of Health Affecting Disparities in Cancer Survival

Wanting Cui[a], Joseph Finkelstein[a]

[a] *Center for Biomedical and Population Health Informatics, Icahn School of Medicine at Mount Sinai, New York, New York, United States*

Abstract

The aim of this pilot study was to identify social determinants of health (SDH) that affect disparities in cancer survival. A limited dataset was generated by querying electronic medical records (EHR) from an academic medical center in New York City between January 2003 and November 2020. Socio-demographic characteristics that affected survival in 22,096 cancer patients were analyzed using descriptive statistics and logistic regression analyses. Two subsets of adult patients were identified: patients who were deceased less than 1 year after diagnosis and patients who survived over 5 years after diagnosis. Percentage of individuals with short survival in Blacks and Whites was respectively 41.4% and 22.2% for lung cancer, 9.8% and 7.1% for colorectal cancer, 2.9% and 0.7% for breast cancer, 6.8% and 4.0% for multiple myeloma, and 1.4% and 0.8% for prostate cancer. Logistic regression identified SDH factors increasing likelihood of shorter survival that included older age, and being male, Black or Hispanic. We concluded that further analysis of a broader spectrum of SDH factors is warranted.

Keywords:

Cancer survival, health disparities, EHR

Introduction

Cancer screening has become an important topic in the past 20 years. According to the National Cancer Institute, cancer screening is an effective way to detect cancer in its early stage and it significantly reduces the proportion of late stage cancer detection [1]. Tools and methods used to detect cancer has improved drastically in recent years including mammography used to screen for breast cancer [2], prostate-specific biomarkers utilized in prostate cancer screening [3], low-dose CT scans used to detect lung cancer for smokers and older age adults [4]. As a result, the 5 year survival rate has increased significantly for various cancer types. As of 2018, the 5 year cancer survival rate has reach 99.3% for breast cancer, 64.7% for colon and rectum cancer, 97.3% for prostate cancer and 21.7% for lung and bronchus cancer [5].

Cancer screening and treatment should improve survival across all population subgroups. A majority of studies focused on the medical determinants that relate to cancer survivals [6-8]. However, there has been a number of studies on socio-demographic factors affecting disparities in cancer diagnoses and cancer screenings [9–12]. Most of these studies focused on the demographics of a single cancer types or selective demographic groups. Our study aims to provide a comparative overview of socio-demographic differences in 5 major cancer types: breast cancer, colorectal cancer, lung cancer, multiple myeloma and prostate cancer, within one health system in New York City. Using this approach, we aim at initial assessment of a number of social determinants of health as potential drivers of disparities in cancer survival, especially for those who deceased in a short amount of time after diagnosis.

Methods

Dataset:

A limited dataset was generated from electronic medical records from Mount Sinai Health System for five cancer types: breast cancer, colorectal cancer, lung cancer, multiple myeloma and prostate cancer. All cancer diagnoses within these 5 categories were verified for all patients in this dataset using a validated semi-automated pipeline. The dataset contained patients' cancer records from 1980 to November 2020. There were 14094 cases of breast cancer, 7900 cases of colorectal cancer, 5971 cases of lung cancer, 4801 cases of multiple myeloma and 16,165 cases of prostate cancer. Prostate cancer and breast cancer constituted a majority proportion of diagnoses, with 32% and 30 % respectively, followed by 16% of diagnoses of colorectal cancer, 12% of lung cancer and 10% of multiple myeloma.

In the analytic dataset, we extracted 2 subsets from the original dataset: patients who has a short period of time between diagnoses and death, and patients who have survived for over 5 years after diagnoses. We defined patients who survived a short time as the number of days between diagnoses and death are equal or less than 1 year. There were 2 sets of patients constituted to patients who have survived for over 5 years after diagnoses. For patients who are deceased, we include those whose number of days between diagnoses and death are greater than 1825 days (5 years). And for patients who are alive, we included those whose diagnoses date was earlier than 2015/1/1. Since early screening are promoted, we only included patients whose diagnoses are later than 2003/1/1. Furthermore, we excluded children from the analysis, as there were few children who had a short period of time between diagnoses and death. We further excluded patients who had missing age, sex and race information. Although each patient could have been diagnosed with more than one cancer, we treat each diagnosis as independent.

Variables

Variables in the dataset includes patients' demographic information, such as age, sex and race, alive indicator, cancer diagnoses date and death date if applicable. We categorized patients' age into 3 levels: young adult, middle-age adult and older adult. We defined young adult as patients who were diagnosed between 18 to 40 years old. Middle-age adult was defined as patients who were diagnosed between 41 and 65 years old. And older adult was defined as patients who were 66 or older

at the time of diagnosis. In addition, we calculated the number of days between patients' diagnoses and death if applicable.

Statistical Method

In the first part of analysis, we will performed exploratory data analysis. We calculated mean, medium and standard deviation for continuous variables and we calculated frequency and proportions for categorical variables. In addition, we plotted histogram, bar plot and trend plots to compare variables with multiple levels.

In logistic regression, we planned to investigate the effect of demographical factors on patients duration of survival after cancer diagnosis. The independent variables are age, sex and race. We defined the dependent variable as whether a patient has a short time between diagnosis and death. Thus, patients who survived over 5 years are labeled as 0 and patients whose number of days between diagnoses and death are equal or less than 1 year are labeled as 1. We performed logistic regression on the 5 cancer types independently, and calculated odds ratio (OR), 95% confidence interval (CI) and p value for each factors accordingly.

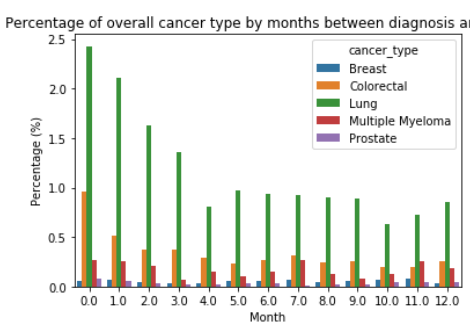
All analyses were performed in Anaconda Jupyter Notebook, using Python 3.7.6.

Results

Exploratory Data Analysis

There are 1280 patients who survived less than 1 year between diagnoses and death in the analytic dataset. 78 (6%) of them had breast cancer, 292 (23%) of them had colorectal cancer, 757 (59%) patients had lung cancer, 92 (7%) patients had multiple myeloma and 61(4%) patients had prostate cancer. In contrast, we identified 20,816 patients who survived over 5 years after their cancer diagnoses. 6390 (31%) of them had breast cancer, 3613(17%) of them had colorectal cancer, 2064 (10%) patients had lung cancer, 2017 (10%) patients had multiple myeloma and 6732 (3%) patients had prostate cancer. Although there are less overall colorectal cancer and lung cancer diagnoses, there are significantly more patients who deceased within a short of period of time, comparing to the other 3 cancer types. And this disparities was true for up to 12 month between diagnoses and death (Figure 1). In addition, according to figure 1, there were more death for the interval between diagnoses and death was less than 3 months. After that the number of death remain constant for other intervals. Thus, it is important to identify factors that contribute to the short timeline for lung cancer and colorectal cancer patients.

Figure 1 Histogram of Months between diagnosis and death



For lung cancer patients, according to table 1, the average age for patients who died within 1 year are 69.84 years old, comparing to 67.23 years old for patients who survived for over 5 years. Most patients who were diagnosed with lung cancer were middle age adults (n= 829, 40.16%) and older adults (n=1204, 58.33%). And most young adult who were diagnosed with lung cancer survived. However, we identified 4 young adult patient who diagnosed with lung cancer who deceased within 1 year of diagnosis. In addition, there were more female patients (n = 1181, 57.22%) who were diagnosed with lung cancer and survived, whereas there were more male patients (n = 883, 55.35%) who were diagnosed with lung cancer and deceased within a short period time. Race was also an important factor. There were higher proportions of non-white and non-Asian patients who were diagnosed with lung cancer and deceased in a short time. Furthermore, figure 2 is the histogram of number of days between cancer diagnosis and death for lung cancer patients who survived a short period of time.

Figure 2 Distribution of days for lung cancer

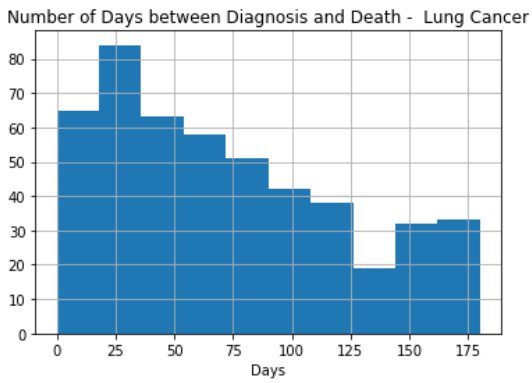
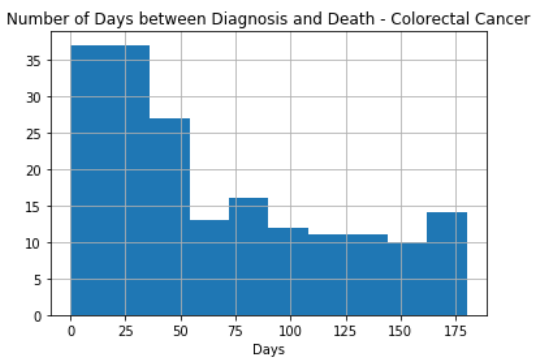


Figure 3 Distribution of days for colorectal cancer



Based on figure 1 and table 1, there were less patients who deceased from colorectal cancer (n = 292) within a short period of time than lung cancer patients (n =757). The average age for patients who survived a short time was 71.74 years old and the average age for patients who survived a long time was 62.41 years old. Similar to lung cancer, most patients who were diagnosed with colorectal cancer were middle age adults and older adults (n= 1502, 41.57%). However, there were more young patients (n = 213, 5.9%) and middle-age patients (n = 1898, 52.53%) with colorectal cancer diagnoses than those with lung cancer diagnoses. There were more male patients with colorectal cancer than female patients in both groups. But the proportion of male patients (n= 162, 55.48%) who deceased in a short time was higher than those (n= 1813, 50.18%) who survived for overall 5 years. The distribution of race of the 2 levels are homogenous. And figure 3 is the distribution of number of days between cancer diagnosis and death for colorectal cancer patients who survived a short period of time.

There were significantly less patients who deceased from breast cancer, multiple myeloma and prostate cancer. The average age of patients who deceased from these 3 cancer groups were larger than the average age of patients who survived a long time. Gender wasn't a contributing factor for breast cancer and prostate cancer, because these cancers were gender specific. For breast cancer and multiple myeloma, there were high proportions of young adults and middle-age adults who survived a long time. In terms of race, there were higher proportion of black patients who deceased within 1 year than the proportion of black patients who survived a long time across all 3 cancer groups. The percentage of individuals with shorter survival was statistically significantly higher in Blacks as compared to Whites across all 5 cancers. Percentage of individuals with short survival in Blacks and Whites was respectively 41.4% and 22.2% for lung cancer, 9.8% and 7.1% for colorectal cancer, 2.9% and 0.7% for breast cancer, 6.8% and 4.0% for multiple myeloma and 1.4% and 0.8% for prostate cancer.

Table 1. Summary statistics of 5 cancer types based on 2 surivival levels.

Cancer Type	**Lung**		**Colorectal**		**Breast**		**MM**		**Prostate**	
Survival Period	Short	Long	Short	Long	Short	Long	Short	Long	Short	Long
Count	757	2064	292	3613	78	6390	92	2017	61	6732
Age										
mean	69.84	67.23	71.74	62.41	71.38	56.38	67.93	59.38	73.48	62.27
std	11.16	11.05	15.04	13.84	15.71	12.98	12.34	10.95	10.62	8.40
median	70	68	74	62	70.5	56	67	60	73	62
Age group										
Young Adult	0.53%	1.50%	3.08%	5.90%	0.00%	10.03%	2.17%	4.71%	0.00%	0.30%
Middle Age Adult	33.29%	40.16%	27.40%	52.53%	37.18%	65.49%	42.39%	66.44%	22.95%	64.97%
Older Adult	66.18%	58.33%	69.52%	41.57%	62.82%	24.48%	55.43%	28.85%	77.05%	34.73%
Gender										
Female	44.65%	57.22%	44.52%	49.82%	100.00%	99.33%	40.22%	45.51%	0.00%	0.00%
Male	55.35%	42.78%	55.48%	50.18%	0.00%	0.67%	59.78%	54.49%	100.00%	100.00%
Race										
American Indian	0.26%	0.10%	0.00%	0.08%	0.00%	0.09%	0.00%	0.10%	0.00%	0.07%
Asian	0.66%	1.70%	1.71%	2.82%	0.00%	3.97%	0.00%	1.54%	0.00%	0.86%
Black	23.25%	12.06%	17.47%	13.01%	34.62%	13.97%	27.17%	16.86%	24.59%	15.21%
Islander	4.49%	3.68%	8.22%	7.61%	1.28%	1.61%	3.26%	1.04%	0.00%	0.94%
Other	16.38%	11.68%	14.73%	14.89%	28.21%	20.22%	20.65%	27.17%	16.39%	15.08%
White	54.95%	70.78%	57.88%	61.58%	35.90%	60.13%	48.91%	53.30%	59.02%	67.84%

Logistic Regression

In this part of the study, we want to evaluate the demographical factors described previously on patients' length of survival after cancer diagnoses. We performed logistic regression on all 5 cancer types individually. We defined patients who survived over 5 years after diagnosis as 0 and patients who survived less than 1 year after diagnosis as 1. All independent variables: age, gender and race were categorized into levels.

According to table 2, older age, gender and race were significant factors for lung cancer. Male patients were likely to deceased in a short period of time after diagnoses than female patients. And the odds of black patients and other race patients, which normally constitutes Hispanic patients at Mount Sinai, deceased in a short time is twice that of white patients.

For colorectal cancer, older age, gender and race are significant variables. Compared to young adults, older adults with colorectal cancer were 3 times more likely to survive less than 1 year after diagnoses. In addition, similar to lung cancer patients, male patients were likely to survive a short time compare to female patients. Black patients were more likely to deceased in a short period time compare to that of white patient.

For breast cancer, we tested the influence of age and race. The odds of older patients and non-white patients surviving only a short period of time were significantly higher than that of middle age patients or white patients. In addition, older adults and black patients were at higher risk of death in a short time, compare to young or white patients with multiple myeloma. Similar to breast cancer, we only tested the effect of age and race for prostate cancer patients. And compared to middle age adults, older adults were 6 times more likely to be deceased in a short time after cancer diagnoses.

Discussion

Race was an important factor. There was significantly more proportions of black patients deceased in short period of time than in a long time after diagnosis. This disparity was observed in all five cancer types. When comparing within race, only 12% of black patients survived a long time after lung cancer diagnosis, whereas over 23% of black patients deceased in a short time after lung cancer diagnosis. In colorectal cancer, 13% of black patients survived a long time, compared to 17% of black patients survived a short time. In breast cancer, 14% of black patients survived a long time; however over 34% of black patients deceased in a short time after diagnosis. In multiple myeloma, 17% of black patients survived a long time and 27% of black patients survived a short time. In prostate cancer, 15% of black

patients survived a long time and 25% of black patients survived a short time. When comparing with other races, black patients who were diagnosed with lung cancer or breast cancer were more likely to be deceased in a short time after diagnosis compare to white patients. And hispanic patients (race_other) with lung cancer were also at risk of shorter survival length compared to white patients. Thus, promoting routine cancer screening is important in black and hispanic communities.

Table 2. Logistic regression results.

	OR	Confid. Interval		P Value
Lung				
Age_young	1.00			
Age_middle	2.42	0.84	7.00	0.104
Age_old	3.69	1.28	10.65	0.016
Gender_Female	1.00			
Gender_Male	1.72	1.45	2.05	0
Race_White	1.00			
Race_Asian	0.54	0.21	1.40	0.205
Race_Black	2.70	2.16	3.39	0
Race_Islander	1.58	1.03	2.42	0.035
Race_Other	1.89	1.48	2.42	0
Colorectal				
Age_young	1.00			
Age_middle	1.00	0.49	2.03	0.998
Age_old	3.33	1.68	6.60	0.001
Gender_Female	1.00			
Gender_Male	1.33	1.04	1.70	0.021
Race_White	1.00			
Race_Asian	0.75	0.30	1.89	0.548
Race_Black	1.64	1.18	2.30	0.004
Race_Islander	1.14	0.72	1.79	0.576
Race_Other	1.22	0.85	1.73	0.28
Breast				
Age_middle	1.00			
Age_old	5.41	3.40	8.61	0
Race_White	1.00			
Race_Black	4.61	2.70	7.85	0
Race_Other	2.56	1.46	4.49	0.001
Multiple Myeloma				
Age_young	1.00			
Age_middle	1.39	0.33	5.84	0.657
Age_old	4.28	1.02	17.94	0.047
Gender_Female	1.00			
Gender_Male	1.37	0.88	2.11	0.159
Race_White	1.00			
Race_Black	1.90	1.14	3.17	0.013
Race_Other	0.85	0.49	1.46	0.549
Prostate				
Age_middle	1.00			
Age_old	6.52	3.58	11.88	0
Race_White	1.00			
Race_Black	2.16	1.17	3.97	0.014
Race_Other	1.31	0.65	2.66	0.453

There were significantly more lung cancer patients and colorectal cancer patients who deceased in a short time after diagnosis compared to breast cancer, multiple myeloma and prostate cancer patients. Although the average age for patients who survived a short time were higher than the average age for patients who survived a long time across all 5 cancer types, older age, especially for patients who are over 65 years old, was the only significant age factor for all cancer types. For these 3 cancer groups, early and effective screening could contribute to patients' long survival length. There were high proportions of younger adults diagnosed with breast cancer and colorectal cancer. In contrast, lung cancer was harder to detect in early stages.

Gender was a significant variable for lung cancer and colorectal cancer. Since breast cancer and prostate cancer were gender specific cancer, we didn't include this variable in the evaluation. Although there were male patients who were diagnosed with breast cancer, they represented a small amount (less than 1%) and no male patients were deceased in 1 year after diagnoses. For lung cancer, there was a higher proportion of female patients who survived a long time than male patients, yet there were more male patients who deceased a short time after diagnoses than female patients. Furthmore, there were more male patients who were diagnosed with colorectal cancer in both levels. Thus, males are more susceptible to colorectal cancer. In addition, according to past studies, female patients were more proactive in voicing their concerns and seeking assistance for any discomforts. They were also more proactive in participating cancer screening.

However, these analysis are preliminary. In future studies, we would like to add more variables into our analysis. We are in the progress of adding cancer stages for all diagnosed patients. And we also plan to include patients' socio-enconomic status: such as education level, income level and employment status to our dataset. Furthermore, we will look into patients' medical history and identify chronic diseases and calculate the Charlson Comorbidity Index for each patient. In machine learning, we plan to apply unsupervised machine learning method to discover latent clusters within in each cancer types and evaluate each groups characteristics and survival rates [13-14].

Conclusion

In this study, we produced descriptive statistics of patients' demographics of lung cancer, colorectal cancer, breast cancer, multiple myeloma and prostate cancer. After that we identified demographical factors that affect patients' length of survival after cancer diagnosis for 5 cancer types by performing logistic regression analyses.

Age, gender, and race were significant factors for lung cancer and colorectal cancer. Age and race were important factors for breast cancer. Age and race were the significant factors for prostate cancer, multiple myeloma and prostate cancer. Older adults, male and patients in black and hispanic communities were the people most susceptible to shorter length of survival after cancer diagnoses. Thus, promoting cancer screening for these demographical groups were crucial to help identify cancer in early stages.

We plan to incorporate more variables, such as: cancer stages, socio-economical status and comorbidity score in our study. Thus, future analysis is warranted.

References

[1] Advisory Committee on Cancer Prevention, Recommendations on cancer screening in the European Union. *Eur J Cancer* **36** (2000), 1473-8.

[2] U.S. Preventive Services Task Force, Screening for breast cancer: U.S. Preventive Services Task Force recommendation statement, *Annals of Internal Medicine* **151(10)** (2009), 716–726.

[3] Olleik G, Kassouf W, Aprikian A, et al., Evaluation of new tests and interventions for prostate cancer management: A systematic review, *J Natl Compr Canc Netw.* **16(11)** (2018), 1340-1351.

[4] American cancer society, Lung Cancer Early Detection, Diagnosis, and Staging, American Cancer Society, 2020 https://www.cancer.org/cancer/lung-cancer/detection-diagnosis-staging/detection.html

[5] National Cancer institute, Cancer statistics: Reports on Cancer SEER Cancer Stat Facts https://seer.cancer.gov/statfacts/

[6] J. Sheridan, P. Walsh, D. Kevans, T. Cooney, S. O'Hanlon, B. Nolan, et al., Determinants of short- and long-term survival from colorectal cancer in very elderly patients, *J Geriatr Oncol* **5(4)** (2014), 376-383.

[7] Merglen A, Schmidlin F, Fioretta G, et al., Short- and Long-term Mortality With Localized Prostate Cancer, *Arch Intern Med* **167(18)** (2007), 1944–1950.

[8] Soerjomataram, I., Louwman, M.W.J., Ribot, J.G. et al., An overview of prognostic factors for long-term survivors of breast cancer, *Breast Cancer Res Treat* **107** (2008), 309–330.

[9] Mostafa, G., Matthews, B. D., Norton, H. J., Kercher, K. W., & al, e., Influence of demographics on colorectal cancer, *The American Surgeon 70***(3)** (2004), 259-64.

[10] Shuman AG, Entezami P, Chernin AS, Wallace NE, Taylor JMG, Hogikyan ND, Demographics and efficacy of head and neck cancer screening, *Otolaryngology–Head and Neck Surgery* **143(3)** (2010), 353-360.

[11] Ghanouni, A., Renzi, C. & Waller, J, A cross-sectional survey assessing factors associated with reading cancer screening information: previous screening behaviour, demographics and decision-making style, *BMC Public Health* 17 **327** (2017).

[12] Robert A. Hiatt, Nancy Breen, The Social Determinants of Cancer: A Challenge for Transdisciplinary Science, *American Journal of Preventive Medicine* **35(2)** (2008) S141-S150.

[13] Cui W, Robins D, Finkelstein J, Unsupervised Machine Learning for the Discovery of Latent Clusters in COVID-19 Patients Using Electronic Health Records, *Stud Health Technol Inform* **272** (2020), 1-4.

[14] Cui W, Cabrera M, Finkelstein J. Latent COVID-19 Clusters in Patients with Chronic Respiratory Conditions, *Stud Health Technol Inform* **275** (2020), 32-36.

Address for correspondence

Wanting Cui, wanting.cui@mssm.edu