HHS Public Access

Author manuscript

Health Informatics J. Author manuscript; available in PMC 2020 December 03.

Published in final edited form as:

Health Informatics J. 2020 March ; 26(1): 8–20. doi:10.1177/1460458218824692.

Assessing the effect of data integration on predictive ability of cancer survival models

Yi Guo, **Jiang Bian**, **Francois Modave**, **Qian Li**, **Thomas J George**, **Mattia Prosperi**, **Elizabeth Shenkman**

University of Florida, USA

Abstract

Cancer is the second leading cause of death in the United States. To improve cancer prognosis and survival rates, a better understanding of multi-level contributory factors associated with cancer survival is needed. However, prior research on cancer survival has primarily focused on factors from the individual level due to limited availability of integrated datasets. In this study, we sought to examine how data integration impacts the performance of cancer survival prediction models. We linked data from four different sources and evaluated the performance of Cox proportional hazard models for breast, lung, and colorectal cancers under three common data integration scenarios. We showed that adding additional contextual-level predictors to survival models through linking multiple datasets improved model fit and performance. We also showed that different representations of the same variable or concept have differential impacts on model performance. When building statistical models for cancer outcomes, it is important to consider cross-level predictor interactions.

Keywords

cancer survival; data heterogeneities; data integration; interactions; model performance; multi-level data analysis

Introduction

As the second leading cause of death, cancer is responsible for one in every four deaths in the United States.[1] It is estimated that there will be approximately 1.7 million new cancer cases and 600,000 cancer deaths in the United States in 2017.[2] Furthermore, the numbers of new cancer cases and deaths per year are on the rise. By 2020, the number of new cancer cases is expected to reach approximately 2 million per year.[3] Cancer will soon surpass heart disease as the leading cause of death in the United States.[3] To improve cancer survival rates and prognosis, one of the first steps is to improve our understanding of contributory factors

Corresponding author: Jiang Bian, Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, 2004 Mowry Road, Suite 3228, P.O. Box 100219, Gainesville, FL 33610, USA. bianjiang@ufl.edu.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Supplemental material

Supplemental material for this article is available online.

associated with cancer survival. Prior research has suggested that cancer survival is influenced by multiple factors from multiple levels. At the individual level, cancer survival is influenced by not only the cancer type, stage of diagnosis, treatment, patient demographics, and health care access but also risky health behaviors such as smoking, alcohol drinking, obesity, and physical inactivity. At the contextual level, cancer survival is influenced by public policies that affect the health care delivery system which could impact patients' travel distance to the treatment facility and adjuvant cancer treatments.[4]

Prior epidemiologic research on cancer survival in the United States has primarily focused on contributory factors from the individual level due to limited availability of integrated datasets.[5-17] Most of these analyses used data from a single source or two sources at most, such as data from a hospital or a cancer registry. Hospital data used for survival analyses are often data obtained from clinical trials or retrospective chart reviews, whereas cancer registry data are often data from a population-based cancer registry such as the Surveillance, Epidemiology, and End Results (SEER) registry or a state-based cancer registry.[18] For example, many cancer survival studies used data from the SEER program.[5-7,11,15-17] A recent study published in the *Journal of American Medical Association* (*JAMA*) used the SEER data to evaluate racial and ethnic disparities in cancer-specific survival among women diagnosed with invasive breast cancer.[6] Another study used the SEER data to examine prognostic factors for survival among patients with resected early-stage rectal adenocarcinoma.[17] Although cancer survival studies using registry data usually have much larger sample sizes than those using hospital data, these studies often lack information on contextual factors that contribute to cancer survival beyond the individual level.

Our current understanding of contributory factors for cancer survival is mostly gained from disparate information gathered across many studies, each analyzing a subset of predictors. However, it is important to simultaneously examine plausible cancer survival predictors from multiple levels (i.e. top-down approach to model building), so that confounding effects among predictors can be fully understood.[19] In addition, when both individual- and contextual-level factors are available, it is possible to test and identify significant cross-level predictor interactions, which not only improves model fit but also provides important information to inform the design of future interventions. Thus, integrating existing, relevant datasets from heterogeneous sources is required to conduct a sound multi-level data analysis on cancer survival.

On the other hand, researchers are faced with unique challenges when integrating data from different sources since these data can be heterogeneous in terms of syntax, database schema, and semantics. For instance, one of the challenges integrating variables from different sources is that variables measuring the same concept could have different representations. One example is that two measures can be used to define the rurality of a geographic unit (e.g. census tract or county): the rural–urban commuting area (RUCA) codes and the National Center for Health Statistics (NCHS) urban–rural classification scheme.[20,21] Both measures are ordinal variables that can classify a geographic unit along the urban–rural continuum. However, the RUCA classification has 10 categories at the census tract level, whereas NCHS classification has six categories at the county level. Thus, it is necessary to decide whether to integrate RUCA and NCHS codes to create a new rurality indicator, or to

choose one based on their fit with the study's conceptual framework or their performance in data analysis.

The goal of the study was to examine how data integration impacts the performance of cancer survival prediction models. We examined the predictive ability of cancer survival models under three common data integration scenarios that researchers often face in data analysis. We linked data from four different sources and evaluated discrimination and reclassification performance of Cox proportional hazard models for breast, lung, and colorectal cancers when (1) additional predictors, especially contextual factors, become available through linking multiple datasets; (2) different forms of the same predictor are available; and (3) different predictors representing the same concept are available.

Methods

Study setting and data sources

This analysis was based on data of the University of Florida (UF) Health Cancer Center Catchment Area (CCCA) from multiple sources. The UF Health CCCA is a region in North Florida that included 20 counties: Alachua, Baker, Bradford, Citrus, Clay, Columbia, Dixie, Gilchrist, Hamilton, Jefferson, Lafayette, Leon, Levy, Madison, Marion, Putnam, Sumter, Suwannee, Taylor, and Union. We obtained data for these 20 counties from six sources: (1) the Florida Cancer Data System (FCDS), a statewide population-based cancer registry; (2) the FLHealthCHARTS,[22] a community health assessment tool created by the Florida Department of Health that has data on health indicators such as chronic diseases and injury from various sources; (3) the 2000 US Census; (4) the Behavioral Risk Factor Surveillance System (BRFSS) of the Centers for Disease Control and Prevention;[23] (5) the RUCA codes; and (6) the NCHS urban–rural classification scheme.

Patients

We identified adult cancer patients (18 years or older at the time of diagnosis) with primary breast, lung, or colorectal cancer from the FCDS. Primary cancer sites were identified using International Classification of Diseases for Oncology–3rd edition (ICD-O-3) codes C50.0 through C50.9 for breast; C34.0 through C34.3, and C34.8 and C34.9 for lung; and C18.0 through C18.9, and C26.0 for colorectal cancer. We excluded patients diagnosed before 1996 because health insurance status was not routinely recorded in FCDS before that year. We also excluded re-occurring cancer cases of the same anatomic site within 5 years. For breast cancer, we only analyzed female breast cancer cases. Data from 50,151 unique cancer patients (18,644 breast; 21,552 lung; and 9955 colorectal) diagnosed between 1996 and 2010 were used for data analysis. We enumerated a few random samples of our data in a supplementary file along with a detailed data dictionary for the variables we used. The data dictionary also indicated the source database of each predictor.

Data integration and variables

The outcome of the study was survival time (in months) obtained from the FCDS. Survival time was defined as the time from the date of cancer diagnosis to the date of death due to cancer or the date of last contact. The predictor variables and their sources were summarized

in Table 1. We also detailed the predictors of all the models we built in Supplemental Appendix A. The FCDS contains individual-level cancer data. The FLHealthCHARTS, BRFSS, and NCHS data are at the county level, which express the characteristics of each county (e.g. average smoking rate of a county from BRFSS). The US Census and RUCA data are at the census tract level. We linked the FCDS data with the FLHealthCHARTS, US Census, and BRFSS data by mapping an individual's residency to a census tract or county (i.e. with a county code or a census tract code). For example, if a patient lives in the Alachua county, we then pulled in contextual factors such as the average smoking rate of the Alachua county from the BRFSS. The individual-level predictors included patients' demographic, tumor, and treatment information obtained from the FCDS. The contextual-level predictors included county-level density of primary care physicians obtained from the FLHealthCHARTS, the four Social Vulnerability Index (SVI) domain scores and census tract rurality status computed from the 2000 US Census data,[24] the county-level smoking and alcohol consumption rates from the BRFSS, and the area rurality status defined by either RUCA or the NCHS urban–rural classification scheme. The four SVI domains included socioeconomic status (SES), household composition and disability, minority status and language, and housing and transportation.[24]

Statistical analysis

We assessed the predictive ability of Cox proportional hazard models on cancer–specific survival at 5 years under three data integration scenarios.

Data integration scenario A (when additional predictors become available through linking multiple datasets).—Under this scenario, we built a series of nested Cox models: models A1 to A3. In model A1, we used the individual-level predictors from the FCDS only. These predictors included sex, race, age at diagnosis, year of diagnosis, stage of diagnosis, treatment received, smoking status, marital status, and health insurance. In model A2, we used the integrated dataset and additionally added the contextual predictors to model A1. These contextual predictors included the four SVI domain scores from the US Census, density of primary care physicians from the FLHealthCHARTS, and county-level smoking rate, alcohol consumption rate, and health status from the BRFSS. In model A3, we additionally added the interactions between race and the other predictors in model A2. Model A3 is the full model, which provided information on the added impact from cross-level interactions on cancer-specific survival.

Data integration scenario B (when different forms of the same variable are available).—Under this scenario, we built four models using four different census tract rurality definitions based on the RUCA codes. RUCA codes use measures of population density, urbanization, and daily commuting to classify US Census tracts into 10 groups. As seen in Table 2, model B1 included a rurality predictor based on the original 10-level RUCA definition. Models B2 to B4 included a rurality predictor based on different scientifically meaningful groupings of the RUCA codes. For all the models, we included all the individual-level factors from the FCDS as other predictors.

Data integration scenario C (when different variables representing the same concept are available).—Under this scenario, we built one model (model C1) using a rurality predictor based on the NCHS urban–rural classification scheme and compared this model to model B1, which included a rurality predictor based on the RUCA rurality definition. According to the NCHS urban–rural classification scheme, the US counties can be classified into six groups: large central metropolitan, large fringe metropolitan, medium metropolitan, small metropolitan, micropolitan, and non-core. All the individual-level factors from the FCDS were also included in model C1 as predictors.

For all prediction models, performance was evaluated using the Akaike information criterion (AIC), c-statistic, integrated discrimination improvement (IDI), relative IDI, and continuous net reclassification improvement (NRI).[25-27] The AIC and c-statistic were computed for all models. The IDI and continuous NRI were computed for comparing models 2 and 3 to model 1. The c-statistic, IDI, and relative IDI are measures of discrimination, which is a model's ability to distinguish between subjects with and without an event (i.e. cancer-related death at 5 years). The c-statistic is the estimated area under the receiver operating characteristic (ROC) curve. The IDI equals the difference in discrimination slopes between the model with additional predictors and the model without, or the difference in the proportion of variance explained by the two different models. The continuous NRI is a measure of improvement in reclassification, defined as the sum of two differences in proportions resulting from the addition of new predictors: (1) proportion of individuals with events who have an increase in predicted risks minus the proportion with a decrease (event NRI), and (2) proportion of individuals without events who have a decrease in predicted risks minus the proportion with an increase (non-event NRI). The use of c-statistic, IDI, and NRI has been extended in the context of survival data. Performance statistics were computed using validated SAS macros available from http://ncook.bwh.harvard.edu/sas-macros.html. All statistical analyses were performed using SAS, version 9.4 (SAS, Cary, NC, USA).

Results

Patients' characteristics

We summarized the patients' characteristics in Table 3. The average age of the patients was 63.2, 68.8, and 70.2 years for breast, lung, and colorectal cancers, respectively. For lung and colorectal cancers, 58.4 and 52.0 percent of the patients were men. In addition, the majority of the patients were non-Hispanic Whites (NHWs) for the three cancers analyzed. The percentage of NHW patients was 87.8 percent for breast cancer, 91.3 percent for lung cancer, and 86.8 percent for colorectal cancer. Regarding stage of diagnosis, most of the breast cancer patients (67.8%) were diagnosed at a localized stage. In contrast, only 16.2 percent and 33.1 percent of the lung and colorectal cancer patients were diagnosed at a localized stage. The rate of late-stage cancer diagnosis (i.e. regional or distant cases) was 27.3, 64.6, and 55.9 percent for breast, lung, and colorectal cancers, respectively. Regarding cancer treatment, most of the breast (93.2%) and colorectal (86.5%) cancer patients had surgery to remove the tumor, whereas only 24.0 percent of the lung cancer patients had cancer-removing surgery.

The majority of the lung cancer patients smoked cigarettes, with 34.6 and 32.5 percent of these patients being former and current smokers. Only 8 percent of the lung cancer patients had never smoked cigarette, compared to 45.5 and 39.8 percent of the breast and colorectal cancer patients being never smokers. A higher percentage of the breast cancer (24.5%) patients had private health insurance, compared to the lung (9.8%) and colorectal (12.2%) cancer patients. The lung and colorectal cancer patients were more likely to have Medicare or Medicare with supplement compared to the breast cancer patients. The 5-year survival rates for the breast, lung, and colorectal cancer patients were 82.6, 17.0, and 58.3 percent, respectively.

Scenario A

We summarized the hazard ratios (HRs) from the Cox models in Supplement A. For breast cancer, we identified a significant race by health insurance interaction (chi-square = 33.6, $p < 0.001$) in the full model. Compared to NHWs, non-Hispanic Blacks (NHBs) had significantly higher hazard when they had Medicaid, private, other type, or no insurance. Hispanics had significantly lower hazard than NHWs when they had no insurance. For lung cancer, we identified three significant interactions: race by stage of diagnosis (chi-square=20.9, $p = 0.002$), race by marital status (chi-square = 18.6, $p = 0.001$), and race by SVI SES (chi-square = 6.31, $p = 0.043$). There was no difference in hazard between NHWs and NHBs when diagnosed with regional or distance lung cancer. However, NHBs had significantly higher hazard than NHWs when the diagnoses were localized. Furthermore, lower SES meant higher hazard among NHWs, but not among NHBs or Hispanics. For colorectal cancer, two interactions were significant: race by stage of diagnosis (chi-square = 20.7, $p = 0.002$) and race by marital status (chi-square = 35.1, $p < 0.001$). NHBs had significantly higher hazard than NHWs when diagnosed with regional colorectal cancer.

The performance statistics from Cox models A1 to A3 were summarized in Table 4. The AIC decreased across models A1 to A3 for all three cancers, indicating increasingly better model fit when adding contextual predictors and interactions to the models. In moving from model A1 to model A2, there was almost no change in the c-statistic for all three cancers. However, we observed a small yet statistically significant increase in the IDI and relative IDI, suggesting a slight increase in the models' distinguishing ability. The continuous NRI was statistically significant for all three cancers. The event NRI was significant for the lung cancer model only (event NRI = 0.03; 95% confidence interval (CI): 0.02–0.05), indicating a significant increase in the model's ability to correctly classify individuals who died of cancer after 5 years when contextual predictors were added to the model. The non-event NRI was statistically significant for all three cancers, suggesting a significant increase in the models' ability to correctly classify individuals who did not die from cancer after 5 years when adding contextual predictors. In moving from model A1 to model A3, we observed similar trends in the performance statistics. The IDI, relative IDI, and continuous NRIs were all larger than those obtained from comparing model A1 to model A2.

Scenario B

The performance statistics from models B1 to B4 were summarized in Table 5. For all three cancers, the c-statistic did not differ much across the models. For breast cancer, models B2

to B4 had significantly lower IDI, relative IDI, and NRI than model B1, indicating lower discrimination and reclassification performance. On the other hand, IDI, relative IDI, and NRI stayed unchanged across models B1 to B4 for lung cancer, suggesting using different RUCA rurality definitions did not affect model performance. For colorectal cancer, models B2 and B3 had significantly lower IDI and relative IDI than model B1, whereas model B4 did not differ from model B1 in discrimination performance. The continuous NRI did not change across the models for colorectal cancer.

Scenario C

The performance statistics comparing models B1 and C1 were also summarized in Table 5. Models B1 and C1 had identical c-statistic for all three cancers. For breast cancer, model C1 had significantly lower IDI, relative IDI, and NRI than model B1, suggesting lower discrimination and reclassification performance. However, model performance did not differ between the two models for lung and colorectal cancers.

Discussion

In this study, we examined the predictive ability of Cox proportional hazard models under three different data integration scenarios for breast, lung, and colorectal cancers. We showed that adding additional contextual-level predictors to survival prediction models through linking multiple datasets improved model fit and performance (scenario A). We also showed that different representations of the same variable (scenario B) or concept (scenario C) have differential impacts on the performance of cancer survival models.

Including plausible predictors from multiple levels in statistical models can avoid omitting significant determinants of cancer outcomes and therefore maximize validity and predictive power.[19] In practice, the top-down approach to model building, in which one starts with a maximal or full model, including plausible interactions, is favored by many. Although overfitting might be of concern, it produces unbiased estimates for expected values (i.e. fixed effects).[19] Sample size could be an issue when fitting larger statistical models, but cancer registry data usually have sufficient sample sizes to support large models. On the other hand, underfitting can create severe bias, which cannot be reduced or eliminated by simply increasing the sample size.[19] In this study, the identification of additional contextual-level predictors proved the importance of adopting the full model approach in modeling cancer survival. In practice, one need to choose a conceptual framework to guide the selection of plausible predictors. For instance, based on the social-ecological model (SEM), cancer survival is determined by a complex interplay of contributory factors from multiple levels: individual, interpersonal, organizational, community, and policy levels.[28] A multi-level survival analysis is only possible through integrating predictors from these five levels.

Furthermore, our results showed that interactions are highly likely to present when modeling cancer survival. Considering only race by other predictor interactions, we observed numerous significant interactions, including cross-level predictor interactions, in the largest model for all three cancers. It is not surprising that these factors interact with each other to impact cancer survival. In fact, we expect many of the social determinants of health (SDH) variables, such as SES, to have differential impacts on cancer survival across different race–

ethnic groups. These interactions, especially cross-levels interactions, can have many clinical and public health implications. For instance, when designing interventions targeting access to care on cancer outcomes, researchers may need to take into consideration the differential effects of access to care across different participant groups (e.g. males vs. females or African Americans vs. Whites). Only through testing interactions in statistical models can we quantify these differential impacts. Although previous studies have identified many determinants of cancer survival, most of these studies did not address the multiple levels of plausible determinants, and almost none of them tested interactions among determinants. Integrating data from various heterogeneous sources not only improves survival model performance but also sets up the stage for building defendable statistical models.

Nevertheless, data integration is a daunting task given the idiosyncrasies of how different source data are collected. Researchers are faced with varying conceptual frameworks and abundant modeling choices that may require a number of integration strategies, each having distinctive challenges. The effort required to integrate data from different sources is substantial due to heterogeneities in syntax (e.g. file formats, access protocols), schema (e.g. data structures), and, perhaps more importantly, semantics (e.g. meanings or interpretations) of the data elements. For example, it required significant effort to download, process, extract, and transform the raw data from heterogeneous sources into integrated analytical datasets. Furthermore, a number of data assumptions have to be made as the different datasets were collected at different time periods and on different populations. For example, our FCDS data include cancer patients from 1996 to 2010, while the US Census data we used were collected from the general population in 2010. Thus, we made assumptions that the area-level characteristics derived from the US Census data were applicable across different time periods. Researchers generally do not have a clear picture of these data integration nuances, and thus, innovative informatics methods and tools are needed to address these data integration challenges in an automated fashion. One particularly promising data integration method is the semantic data integration approach. With the approach, a universal conceptual representation of “information” including data and their relationships, via common controlled vocabulary or “ontologies,” is generated to bridge the syntactic, schematic, and semantic heterogeneities across different sources. Ontologies can be used to encapsulate the knowledge of the source data and associate them with the general terms in the corresponding domain. The use of ontologies can facilitate data integration in many ways, including metadata representation, automatic data verification, and global conceptualization.[29]

Conclusion

Including additional contextual predictors through integrating data from various sources improves the performance of cancer survival models. When building statistical models for cancer outcomes, it is important to consider predictor–predictor interactions, especially cross-level predictor interactions. Furthermore, different representations of the same variable or concept have differential impacts on survival model performance. When integrating heterogeneous variables from different sources, model performance is one criterion that can guide the data integration process. To facilitate such data-driven analytical approaches, novel

informatics methodologies, such as semantic data integration, are needed to bridge heterogeneities in data across different sources.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health (NIH) or Patient-Centered Outcomes Research Institute (PCORI).

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported in part by National Institutes of Health (NIH) grant UL1TR001427, the OneFlorida Cancer Control Alliance (funded by James and Esther King Biomedical Research Program, Florida Department of Health, grant number 4KB16), and the OneFlorida Clinical Research Consortium (CDRN-1501-26692) funded by the Patient-Centered Outcomes Research Institute (PCORI).

References

1. CDC—Statistics for different kinds of cancer, 27 6 2017, https://www.cdc.gov/cancer/dcpc/data/types.htm (accessed 26 September 2017).
2. American Cancer Society. Cancer facts & figures 2017, https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2017/cancer-facts-and-figures-2017.pdf (2017, accessed 11 February 2017).
3. Cancer: at a glance reports. Chronic Disease Prevention and Health Promotion (CDC), https://www.cdc.gov/chronicdisease/resources/publications/aag/dcpc.htm (accessed 26 September 2017).
4. Vetterlein MW, Löppenberg B, Karabon P, et al. Impact of travel distance to the treatment facility on overall mortality in US patients with prostate cancer. Cancer 2017; 123(17): 3241–3252. [PubMed: 28472547]
5. Yu XQ. Socioeconomic disparities in breast cancer survival: relation to stage at diagnosis, treatment and race. BMC Cancer 2009; 9: 364. [PubMed: 19828016]
6. Iqbal J, Ginsburg O, Rochon PA, et al. Differences in breast cancer stage at diagnosis and cancer-specific survival by race and ethnicity in the United States. JAMA 2015; 313(2): 165–173. [PubMed: 25585328]
7. Eng LG, Dawood S, Sopik V, et al. Ten-year survival in women with primary stage IV breast cancer. Breast Cancer Res Treat 2016; 160(1): 145–152. [PubMed: 27628191]
8. Islam KMM, Jiang X, Anggondowati T, et al. Comorbidity and survival in lung cancer patients. Cancer Epidemiol Biomarkers Prev 2015; 24(7): 1079–1085. [PubMed: 26065838]
9. Zullig LL, Carpenter WR, Provenzale DT, et al. The association of race with timeliness of care and survival among Veterans Affairs health care system patients with late-stage non-small cell lung cancer. Cancer Manag Res 2013; 5: 157–163. [PubMed: 23900515]
10. Wang HM, Liao ZX, Komaki R, et al. Improved survival outcomes with the incidental use of beta-blockers among patients with non-small-cell lung cancer treated with definitive radiation therapy. Ann Oncol 2013; 24(5): 1312–1319. [PubMed: 23300016]
11. Smith CB, Bonomi M, Packer S, et al. Disparities in lung cancer stage, treatment and survival among American Indians and Alaskan Natives. Lung Cancer 2011; 72(2): 160–164. [PubMed: 20889227]
12. Tapan U, Lee SY, Weinberg J, et al. Racial differences in colorectal cancer survival at a safety net hospital. Cancer Epidemiol 2017; 49: 30–37. [PubMed: 28538169]
13. Wallace K, Sterba KR, Gore E, et al. Prognostic factors in relation to racial disparity in advanced colorectal cancer survival. Clin Colorectal Cancer 2013; 12(4): 287–293. [PubMed: 24188687]

14. Wassira LN, Pinheiro PS, Symanowski J, et al. Racial-ethnic colorectal cancer survival disparities in the mountain west region: the case of Blacks compared to Whites. Ethn Dis 2013; 23(1): 103–109. [PubMed: 23495630]
15. Shaukat A, Salfiti NI, Virnig DJ, et al. Is ulcerative colitis associated with survival among older persons with colorectal cancer in the US? A population-based case-control study. Dig Dis Sci 2012; 57(6): 1647–1651. [PubMed: 22113428]
16. Doubeni CA, Field TS, Buist DSM, et al. Racial differences in tumor stage and survival for colorectal cancer in an insured population. Cancer 2007; 109(3): 612–620. [PubMed: 17186529]
17. Lee K-C, Chung K-C, Chen H-H, et al. Prognostic factors of overall survival and cancer-specific survival in patients with resected early-stage rectal adenocarcinoma: a SEER-based study. J Investig Med 2017; 65(8): 1148–1154.
18. About the SEER Program—SEER, https://seer.cancer.gov/about/overview.html (accessed 26 September 2017).
19. Cheng J, Edwards LJ, Maldonado-Molina MM, et al. Real longitudinal data analysis for real people: building a good enough mixed model. Stat Med 2010; 29(4): 504–520. [PubMed: 20013937]
20. USDA ERS. Rural-urban commuting area codes, https://www.ers.usda.gov/data-products/rural-urban-commuting-area-codes/ (accessed 8 October 2017).
21. Data access—urban rural classification scheme for counties, 5 9 2017, https://www.cdc.gov/nchs/data_access/urban_rural.htm (accessed 8 October 2017).
22. FLHealthCHARTS. County and state profile reports, http://www.flhealthcharts.com/charts/QASpecial.aspx (accessed 8 October 2017).
23. CDC—BRFSS, 24 8 2017, https://www.cdc.gov/brfss/index.html (accessed 8 October 2017).
24. ATSDR. The Social Vulnerability Index (SVI), https://svi.cdc.gov/ (accessed 8 October 2017).
25. Pencina MJ, D'Agostino RB Sr and Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Stat Med 2011; 30(1): 11–21. [PubMed: 21204120]
26. Uno H, Cai T, Pencina MJ, et al. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med 2011; 30(10): 1105–1117. [PubMed: 21484848]
27. Pencina MJ, D'Agostino RB Sr and Demler OV. Novel metrics for evaluating improvement in discrimination: net reclassification and integrated discrimination improvement for normal variables and nested models. Stat Med 2012; 31(2): 101–113. [PubMed: 22147389]
28. Golden SD and Earp JAL. Social ecological approaches to individuals and their contexts: twenty years of health education & behavior health promotion interventions. Health Educ Behav 2012; 39(3): 364–372. [PubMed: 22267868]
29. Wache H, Vögele T, Visser U, et al. Ontology-based integration of information—a survey of existing approaches In: Proceedings of IJCAI-01 workshop: ontologies and information sharing. Seattle, WA, 2001, pp. 108–117, http://ceur-ws.org/Vol-47/wache.pdf (accessed 12 January 2019).

Table 1.

Data sources and variables.

Possible predictors		Data source
Individual level	Sex	FCDS
	Race	
	Age at diagnosis	
	Year of diagnosis	
	Stage of diagnosis	
	Treatment	
	Tobacco use	
	Marital status	
	Health insurance	
Contextual level	SVI socioeconomic status	US Census
	SVI household composition and disability	US Census
	SVI minority status and language	US Census
	SVI housing and transportation	US Census
	Area rurality status[a]	RUCA, NCHS
	County smoking rate	BRFSS
	County alcohol consumption rate	BRFSS
	County health status	BRFSS
	County density of primary care physicians[b]	FLHealthCHARTS

FCDS: Florida Cancer Data System; SVI: Social Vulnerability Index; RUCA: rural–urban commuting area; NCHS: National Center for Health Statistics; BRFSS: Behavioral Risk Factor Surveillance System.

[a]Defined based on RUCA codes and NCHS urban–rural classification scheme.

[b]Number of primary care physicians per 1000 people.

Table 2.

Rurality definition based on RUCA codes.

Code	Model B1	Model B2	Model B3	Model B4
1	Metropolitan area: core			Metropolitan
2	Metropolitan area: high commuting	Metropolitan	Metropolitan	
3	Metropolitan area: low commuting			
4	Micropolitan area: core			
5	Micropolitan area: high commuting	Micropolitan		
6	Micropolitan area: low commuting			Non-metropolitan
7	Small town: core		Non-metropolitan	
8	Small town: high commuting	Small town		
9	Small town: low commuting			
10	Rural areas	Rural		

RUCA: rural–urban commuting area; FCDS: Florida Cancer Data System.

Other predictors included all the individual-level factors from the FCDS.

Table 3.

Patients' characteristics by cancer site.

Characteristics	Breast N = 18,644	Lung N = 21,552	Colorectal N = 9,955
Age at diagnosis (year)	63.2 (13.4)	68.8 (10.9)	70.2 (12.6)
Sex			
Women	100%	41.6%	48.0%
Men	–	58.4%	52.0%
Race			
Non-Hispanic Whites	87.8%	91.3%	86.8%
Non-Hispanic Blacks	10.1%	7.5%	11.2%
Hispanics	2.1%	1.3%	2.1%
Year of diagnosis			
1996–2000	29.5%	29.1%	33.3%
2001–2005	26.4%	26.3%	27.1%
2006–2010	44.1%	44.6%	39.7%
Stage of diagnosis			
Localized	67.8%	16.2%	33.1%
Regional	23.2%	23.1%	38.4%
Distant	4.1%	41.5%	17.5%
Unknown	5.0%	19.3%	11.1%
Treatment			
Surgery			
Yes	93.2%	24.0%	86.5%
No	6.8%	76.0%	13.5%
Radiation			
Yes	28.5%	33.4%	2.5%
No	71.5%	66.6%	97.5%
Chemotherapy			
Yes	21.3%	29.1%	20.0%
No	78.7%	70.9%	80.0%
Hormone therapy			
Yes	14.4%	0.4%	0.2%
No	85.6%	99.6%	99.8%
Smoking status			
Never	45.5%	8.0%	39.8%
Former	18.8%	34.6%	25.9%
Current	11.9%	32.5%	10.7%
Unknown	23.7%	24.8%	23.6%
Marital status			
Married	58.1%	54.4%	57.4%
Single	39.0%	42.1%	39.4%
Unknown	3.0%	3.5%	3.2%

Characteristics	Breast N = 18,644	Lung N = 21,552	Colorectal N = 9,955
Health insurance			
Uninsured	2.6%	2.8%	2.4%
Private	24.5%	9.8%	12.2%
Medicaid	3.8%	5.1%	3.2%
Medicare	14.7%	19.9%	20.4%
Medicare with supplement	29.6%	35.4%	39.6%
Other	16.0%	10.3%	9.5%
Unknown	8.8%	16.8%	12.9%
Five-year survival rate	82.6%	17.0%	58.3%

Table 4.

Model performance statistics when linking multiple datasets.

Cancer	Performance statistics	Model A1	Model A2	Model A3
Breast	AIC	34,409.25	34,400.09	34,381.607
	c-Statistic	0.800	0.801	0.801
	IDI	–	0.002 (0.000 to 0.003)	0.005 (0.001 to 0.013)
	Relative IDI	–	0.006 (0.002 to 0.010)	0.019 (0.003 to 0.048)
	Continuous NRI	–	0.10 (0.05 to 0.14)	0.12 (0.06 to 0.20)
	Event NRI	–	0.02 (−0.03 to 0.05)	0.01 (−0.03 to 0.05)
	Non-event NRI	–	0.09 (0.07 to 0.10)	0.11 (0.07 to 0.18)
Lung	AIC	277,780.96	277,708.46	277,701.67
	c-Statistic	0.734	0.734	0.734
	IDI	–	0.001 (0.000 to 0.002)	0.002 (0.001 to 0.004)
	Relative IDI	–	0.004 (0.002 to 0.006)	0.006 (0.002 to 0.013)
	Continuous NRI	–	0.09 (0.05 to 0.14)	0.16 (0.06 to 0.23)
	Event NRI	–	0.03 (0.02 to 0.05)	0.11 (0.02 to 0.13)
	Non-event NRI	–	0.06 (0.02 to 0.11)	0.05 (0.02 to 0.11)
Colorectal	AIC	51,526.92	51,475.53	51,448.46
	c-Statistic	0.782	0.784	0.785
	IDI	–	0.003 (0.002 to 0.005)	0.009 (0.002 to 0.013)
	Relative IDI	–	0.009 (0.005 to 0.013)	0.024 (0.007 to 0.035)
	Continuous NRI	–	0.09 (0.04 to 0.16)	0.25 (0.05 to 0.31)
	Event NRI	–	−0.02 (−0.05 to 0.03)	0.10 (−0.04 to 0.14)
	Non-event NRI	–	0.11 (0.08 to 0.14)	0.15 (0.09 to 0.18)

AIC: Akaike information criterion; IDI: integrated discrimination improvement; NRI: net reclassification improvement. Sex was not included as a predictor in the breast models. Model A1: Basic dataset with individual-level predictors only. Model A2: Integrated dataset with individual- and contextual-level predictors. Model A3: Integrated dataset with individual- and contextual-level predictors, and cross-level interactions.

Note: Values in parentheses are lower and upper bounds of 95% confidence intervals.

Health Informatics J. Author manuscript; available in PMC 2020 December 03.

Table 5.

Model performance statistics when using different rurality definitions.

Cancer	Performance statistics	Model B1	Model B2	Model B3	Model B4	Model C1
Breast	AIC	34,396.10	34,403.29	34,404.73	34,41 1.01	34,410.68
	c-Statistic	0.800	0.800	0.799	0.800	0.800
	IDI	–	−0.001 (−0.003 to −0.001)	−0.002 (−0.003 to −0.001)	−0.002 (−0.003 to −0.001)	−0.002 (−0.003 to −0.001)
	Relative IDI	–	−0.005 (−0.009 to −0.002)	−0.007 (−0.010 to −0.002)	−0.008 (−0.012 to −0.002)	−0.006 (−0.011 to −0.002)
	Continuous NRI	–	−0.07 (−0.12 to −0.03)	−0.05 (−0.11 to −0.01)	−0.05 (−0.10 to −0.01)	−0.04 (−0.10 to −0.01)
	Event NRI	–	−0.23 (−0.27 to −0.20)	−0.31 (−0.34 to −0.20)	−0.04 (−0.33 to −0.01)	−0.35 (−0.38 to −0.01)
	Non-event NRI	–	0.16 (0.14 to 0.18)	0.26 (0.14 to 0.27)	−0.01 (−0.03 to 0.27)	0.30 (−0.03 to 0.32)
Lung	AIC	277,760.61	277,762.77	277,758.83	277,777.22	277,736.12
	c-Statistic	0.734	0.734	0.734	0.734	0.734
	IDI	–	−0.000 (−0.000 to 0.000)	−0.000 (−0.000 to 0.000)	−0.000 (−0.000 to 0.000)	0.000 (−0.000 to 0.001)
	Relative IDI	–	−0.000 (−0.001 to 0.001)	−0.000 (−0.001 to 0.001)	−0.001 (−0.002 to 0.001)	0.001 (−0.001 to 0.002)
	Continuous NRI	–	0.03 (−0.01 to 0.08)	0.05 (−0.00 to 0.08)	−0.04 (−0.07 to 0.08)	0.03 (−0.07 to 0.08)
	Event NRI	–	−0.05 (−0.06 to −0.03)	−0.02 (−0.06 to −0.00)	−0.15 (−0.16 to −0.01)	−0.37 (−0.38 to −0.01)
	Non-event NRI	–	0.08 (0.05 to 0.12)	0.07 (0.03 to 0.1 1)	0.11 (0.03 to 0.14)	0.40 (0.03 to 0.42)
Colorectal	AIC	51,526.19	51,527.41	51,526.81	51,528.13	51,508.62
	c-Statistic	0.783	0.782	0.782	0.782	0.783
	IDI	–	−0.001 (−0.001 to −0.000)	−0.001 (−0.001 to −0.002)	−0.000 (−0.001 to 0.000)	−0.000 (−0.001 to 0.001)
	Relative IDI	–	−0.002 (−0.004 to −0.001)	−0.003 (−0.004 to −0.001)	−0.001 (−0.004 to 0.000)	−0.000 (−0.004 to 0.002)
	Continuous NRI	–	−0.04 (−0.09 to 0.01)	−0.03 (−0.09 to 0.02)	−0.01 (−0.08 to 0.03)	−0.01 (−0.08 to 0.03)
	Event NRI	–	−0.27 (−0.3 1 to −0.23)	−0.17 (−0.30 to −0.14)	−0.04 (−0.30 to −0.02)	0.04 (−0.29 to 0.06)
	Non-event NRI	–	0.23 (0.21 to 0.26)	0.15 (0.12 to 0.25)	0.03 (0.01 to 0.25)	−0.04 (−0.06 to 0.25)

AIC: Akaike information criterion; IDI: integrated discrimination improvement; NRI: net reclassification improvement; FCDS: Florida Cancer Data System. Other predictors included all the individual-level factors from the FCDS. Sex was not included as a predictor in the breast models.

Note: Values in parentheses are lower and upper bounds of 95% confidence intervals.