ORIGINAL RESEARCH

Creation and Evaluation of the Illinois Cancer Risk Index as a Predictor of Four Common Cancers

Lei Guo, MA, MS[1]; Margaret E. Wright, PhD[2]; Meredith C. Osias, BA[3];

Mahdi Vaezi, PhD[4]; M. Courtney Hughes, PhD, MS[3]

PEER REVIEWED

Summary

What is already known on this topic?

Information is needed on how social and environmental determinants of health affect outcomes of common cancers to create a measure and identify geographic locations that are most in need of public health intervention.

What is added by this report?

A risk index representing 25 predictors of death from the 4 most common cancers was created by using population-based, county-level data in Illinois. We correlated the index with mortality rates from the 4 most common cancers, and both exhibited similar geospatial distribution across the state.

What are the implications for public health practice?

Professionals in many health fields can adapt our framework to construct indexes and inform public health resource allocations.

Abstract

Introduction

Nearly half of all cancer deaths in the US are attributed to 4 common cancers: lung, colorectal, breast, and prostate. Illinois residents experience higher rates of cancer death from all 4 cancers compared with the US overall. We developed the Illinois Cancer Risk Index (ICRI), which incorporates many predictors of these cancers into a single summary measure, to identify Illinois counties that would benefit most from public health intervention.

Methods

We identified 90 county-level predictors of 4 common cancers, used multicollinearity testing to reduce this number to 61, and applied factor analysis to extract and analyze 4 factors representing 25 variables. Next, we created the ICRI by regressing the 4 factors on our outcome of interest — an age-adjusted common cancers mortality rate (CCMR), incorporating the direction of the β-coefficients from regression models to sum factor scores. Finally, we mapped and assessed the geographic distributions of both ICRI and CCMR by county across the state.

Results

The ICRI was positively associated with the CCMR ($r = 0.59$, $P < .001$) and explained 32.2% of the variance in the CCMR across Illinois. The ICRI showed distinct geospatial patterns across the state, with the highest risk counties located in the east–central, far northern, and southern regions. The CCMR showed similar geospatial patterns.

Conclusion

Our study identifies counties in Illinois that may benefit most from interventions that target multiple cancer risk factors simultaneously. The ICRI may be adapted for use in other geographic locations where data are available.

Introduction

Cancer is the second leading cause of death in the US, with lung cancer accounting for almost one-quarter (23%) of cancer deaths. Other common sites of cancer death are colorectal (9%), female breast (7%), and prostate (5%) (1). Illinois residents experience higher rates of death from all 4 of these cancers compared with the US overall (1). In Illinois, approximately 14,140 deaths from these 4 common cancers are expected in 2022, with wide variation across the state (2).

Determinants of cancer incidence include those related to demographics, social and economic factors, health behaviors, the phys-

ical environment, and clinical care (3). For example, Singh et al (4) showed that cancer outcomes in the US correlate with socioeconomic status and race and ethnicity at both individual and population levels. Other researchers have demonstrated that health behaviors and cumulative environmental quality influence cancer risk (5–7). Krieger (8) proposed an ecosocial theory and discussed how social and biologic reasoning and dynamic and ecologic perspectives could affect distributions and result in social inequalities in cancer outcomes. Most cancer risk factors have been examined in isolation with respect to outcomes. Although some indexes have been constructed and used, their data were older (9,10) than ours, focused on just one cancer type (11), or did not include specific risk factors such as individual health behaviors, air pollution, English proficiency, and poor physical and mental health (9–11). Our study included these additional elements of risk and evaluated the index's relation to an outcome that reflects the 4 most common causes of cancer death in Illinois.

This article introduces the Illinois Cancer Risk Index (ICRI), which incorporates predictors of the most common cancer deaths into a single summary measure. The index can be used to identify counties in Illinois where public health intervention is needed most.

Methods

We identified 90 variables related to the following domains of risk: demographics, social and economic factors, health behaviors, the physical environment, and clinical care. We selected these on the basis of published literature (3), ecosocial theory (8), and available Illinois county data. For each variable, the most recent county-level data (2014–2018) were extracted from publicly available sources, including the Area Health Resources Files (12), County Health Rankings and Roadmaps (CHR) (13), the US Department of Agriculture Food Environmental Atlas (14), and the Illinois Environmental Protection Agency Bureau of Air (15).

We normalized each of the 90 variables and expressed them as either per-capita values or percentages by using the following formula:

$$\text{Variable}_{\text{normalized}} = (\text{Variable} - \text{Variable}_{\text{minimal}}) / (\text{Variable}_{\text{maximal}} - \text{Variable}_{\text{minimal}})$$

Data were available for more than 94% of our study variables for all 102 Illinois counties. We replaced missing data for the risk factors (1.1% of data points missing) and site-specific cancer mortality rates (30.4% of data points missing) by using the hot deck imputation method (16). Subsequently, we standardized data for each variable by applying a z score standardization to transform the different variables into comparable scales (16). We then con-

ducted redundancy and multicollinearity tests by using variance inflation factors (VIFs) to remove highly correlated variables (VIF >5) and improve the efficiency of factor analysis. Sixty-one variables were retained for subsequent exploratory factor analysis (16).

Exploratory factor analysis was used to extract 13 grouped variables called "factors" from these 61 variables. Analysis was initiated by estimating the variance component with principal component analysis. We used the Kaiser Measure of Sampling Adequacy (Kaiser MSA) in combination with variable communalities (proportion of each variable's variance that the factors explain) to extract these factors. We used the Bartlett χ^2 test to validate the estimated factors (all P values < .001). Factors containing 2 or fewer variables or those whose ascribed variables all had factor loadings less than 0.5 were excluded. Four factors representing 25 variables were retained for analysis (Table 1).

Our outcome of interest was the average age-adjusted mortality rate from lung, colorectal, breast, and prostate cancers, which we refer to as the common cancer mortality rate (CCMR). Cancer mortality data were obtained from the Illinois State Cancer Registry and included the years 2014–2018 (15). We replaced missing data with hot-deck imputation for counties that had low counts because of suppressed data.

We used multiple linear regression to assess the bivariate and multivariate relationships between each of the 4 retained factors. We then used this regression analysis information to construct the ICRI. In brief, the direction of each association (the sign of the β-coefficient) was incorporated into the calculation of the ICRI. For each Illinois county, the ICRI was calculated by multiplying the standardized value of each variable by its respective factor loading and then summing all 4 factors together while accounting for the sign of the β-coefficient from regression models. We used the following equation:

ICRI = Factor Score 1 + Factor Score 3 + Factor Score 5 + Factor Score 6

For example, we calculated the ICRI for Cumberland County as

$$\text{ICRI}_{\text{Cumberland County}} = \text{Factor Score 1}_{\text{Cumberland County}} + \text{Factor Score 3}_{\text{Cumberland County}} + \text{Factor Score 5}_{\text{Cumberland County}} + \text{Factor Score 6}_{\text{Cumberland County}} = (-0.49) + (-1.91) + (-3.92) + (-3.05) = -9.37$$

Sensitivity analyses were performed in which the ICRI was regressed individually against breast, colorectal, lung, and prostate cancer mortality rates.

We generated and comparted maps of the ICRI and CCMR by county. We divided county-level CCMRs into quintiles for mapping. For the ICRI map, counties were classified for risk as very low, low, average, high, and very high. Low- and high-risk counties had index values that were more than 0.5 SD but less than 1.5 standard deviations from the mean. Very low-risk and very high-risk counties had an index more than 1.5 standard deviations from the mean. Average risk counties had an ICRI value within 0.5 standard deviations from the mean.

We used SPSS version 27.0 (IBM Corp) to conduct all analyses and constructed maps in ArcGIS Pro 2.8 (Esri Corp).

Results

The communalities of variables (the proportion of common variance found in a particular variable) were all higher than 0.5. All variables correlated with at least 1 other variable (correlation coefficient of at least 0.3), indicating that variables all shared some common variance. We retained 4 factors representing 25 variables because, together, they explained 70.1% of the total variance of the 61 originally identified, noncollinear variables across Illinois. These factors were calculated, along with the variables that each one comprises and their factor loadings (Table 1). Factor 1 included Black race and health behaviors (the variable with the highest factor loading was Black race at 0.931) and explained 27.3% of the total variance. Factor 3 included Hispanic ethnicity, rurality, measures of air pollution, and language barriers (the variable with the highest factor loading was the percentage of non-Hispanic White race at −0.787) and explained 17.8% of the total variance. Factor 5 included aspects of financial security (the variable with the highest factor loading was percentage of people living in poverty at 0.820) and explained 16.7% of the total variance. Factor 6 included density of primary care providers, education, and income ratio (the variable with the highest factor loading was the primary care provider [PCP] rate at 0.553). PCP rate is the ratio of population to primary care physicians and includes practicing physicians specializing in general practice medicine, family medicine, internal medicine, and pediatrics, and explained 8.3% of the total variance.

Regression analysis and index construction

In bivariate regression, Factor 5 was the only significant predictor of the CCMR; Factors 1, 3, and 6 and the ICRI were not. However, in multivariable regression, all 4 retained factors significantly predicted the CCMR. Together, they explained 33.5% of the variance in the county-level CCMR across the state of Illinois (Table 2).

The ICRI had a moderate positive association with the CCMR (r = 0.59, P < .001) and explained 32.2% of the variance in this outcome (Table 2). The ICRI had a mean of 0 and an SD of 4.6. Winnebago County in the northern portion of the state had the highest index value at 10.76, implying a high risk of mortality from lung, breast, prostate, and colorectal cancers in this area. Cumberland County in the east had the lowest value at −9.37, implying a low risk of mortality from these 4 cancers in that area.

We performed sensitivity analyses by regressing the ICRI against individual cancer mortality rates rather than the CCMR (Table 3). The ICRI explained the largest proportion of variation in prostate cancer mortality (64%), followed by deaths from breast (57%), colorectal (56%), and lung (23%) cancer.

Maps

Thirty-three Illinois counties were classified as at high or very high risk for the 4 common cancers (Figure) based on the ICRI, and 32 counties were classified as low or very low risk. Higher risk counties were primarily located in the east–central portion of the state, with several counties in the far northern and southern portions of the state also classified as such. Counties in the 2 highest CCMR quintiles were located predominantly in the northeast and southern parts of Illinois, with some counties in the central portion.
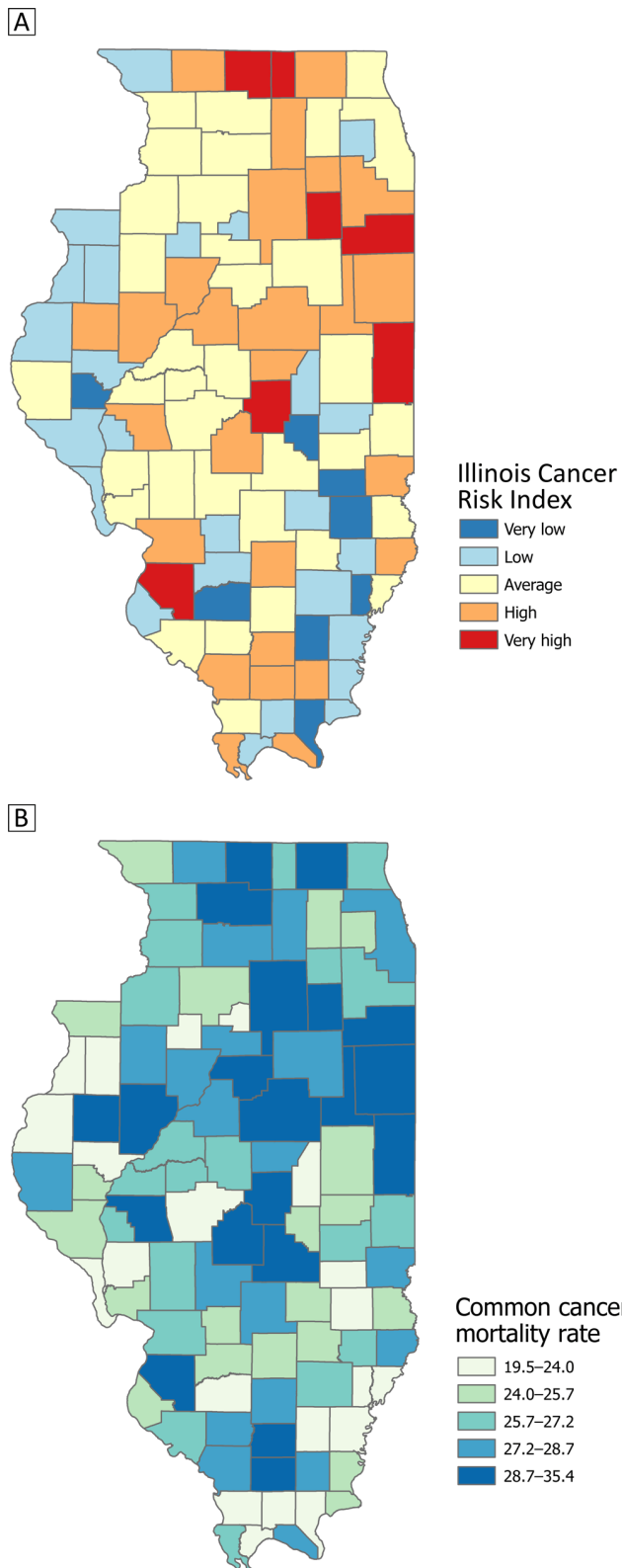
Figure 1. Distribution by county of risk of the 4 most common cancers — lung, colorectal, breast, and prostate — in Illinois. Map A displays risk by the Illinois Cancer Risk Index (ICRI) for each county. Higher risk counties were located in the east-central portion of the state, with some also located in the far northern and southern portions of the state. Map B plots the common cancer mortality rate (CCMR), 2014–2018, for each county in Illinois. Counties in the 2 highest CCMR quintiles were located in the northeast and southern parts of Illinois, with some also located in the central portion of the state.

Discussion

We constructed a novel cancer risk index — ICRI — by using population-based, county-level data from the state of Illinois. The ICRI represented a broad range of determinants of the 4 most common cancers in both Illinois and the US. To the best of our knowledge, our study incorporates one of the largest numbers of cancer risk factors (based on ecosocial theory [8]) to date and has important implications for screening, intervention resource allocation, and access to cancer care.

Our study differs from other reports that also describe cancer risk indexes. Although Scott et al (9) included factors from several domains, our study also examined air pollution, English proficiency, and poor physical and mental health. Wang et al (10) used data from 1998–2000 to focus exclusively on late-stage cancers and did not include rurality, air pollution, English proficiency, poor physical and mental health, or health behaviors such as alcohol consumption, smoking, and diet. In a separate publication, Wang et al (11) used breast cancer mortality throughout Illinois as the outcome of interest and did not include rurality, air pollution, English proficiency, poor physical and mental health, health behaviors, or ratio of population to primary care physicians. Overall, our study evaluated more domains than previous studies and created a unique outcome by averaging the 4 most common causes of cancer death in Illinois.

Demographic variables had high loadings in 2 factors included in our index. For the first factor, the percentage of the population that identified as non-Hispanic Black had the highest factor loading. For the second factor, the percentage of the population that identified as non-Hispanic White had the highest factor loading. Together, these 2 factors explained 45.1% of the total variance of the 61 variables across Illinois. These variables exhibited the connections between social disparities and cancer risks across Illinois counties, showing patterns similar to larger-scale studies. For instance, previously published research found substantially elevated cancer mortality rates at multiple sites among non-Hispanic Black populations compared with non-Hispanic White populations (8). On the other hand, Hispanic populations had lower cancer mortality rates than non-Hispanic White or non-Hispanic Black populations, although Hispanic populations tend to have later-stage dia-

gnoses and poorer quality of life at the national level and across Illinois (1).

In our index, Hispanic ethnicity is loaded passively with air pollution variables. This association may indicate the disparities between racial and ethnic minorities and environmental pollution (7). A recent study showed that Hispanic, Latino, and other minority populations were being exposed to higher levels of dangerous fine particulate matter 2.5 ($PM_{2.5}$) from air pollution than other groups; previous studies also found that the Hispanic population is at higher risk of premature death from exposure to $PM_{2.5}$ air pollution (17), which echo our results. Next, because about 14% of the Illinois population are non–US-born (18), it is important to consider the additional hurdles encountered by this group. Limited English proficiency can burden non–US-born people when they attempt to access health care in the US, although they often have lower cancer mortality rates than people born in the US. Over time, their cancer rates tend to equal or even exceed those of people born in the US as they acculturate, and this tendency is why English proficiency was loaded in our index.

Factor 1 included health behavior–related variables and poor physical and mental health days. Multiple studies have shown that health behaviors are strongly associated with cancer outcomes. For example, smoking and alcohol consumption are known risk factors for multiple cancers (19,20), with smoking the most predominant risk factor for lung cancer (3). Additionally, previous research demonstrated that people who develop mental disorders after a cancer diagnosis may be at higher risk of cancer death (21). Furthermore, mental health treatment offered to cancer patients after diagnosis can improve lung cancer survival, and reductions in the severity of mental illness may manifest in greater self-efficacy for managing chronic conditions and improvements in positive health behaviors, such as physical activity and stress management (22). However, past studies using Illinois data (10,11) and US data (9) did not consider these variables to create their single or multiple cancer sites index.

Living in a rural area — which had a high factor loading in our index — is also associated with higher cancer mortality rates because of limited health care access (23). Another variable, severe housing problems, was loaded in the same factor with rural area. Although housing is not frequently examined with cancer outcomes, studies have found that it is associated with increased cancer mortality disparities between Black and White populations (9,23). Rural residency and the health behavior variables loaded in our factors might represent latent variables related to protective behaviors associated with cancer outcomes.

Household income, median home value, and median gross rent were loaded negatively in Factor 5, echoing the findings from oth-er research that low- and middle-income counties have higher cancer mortality rates than high-income counties in the US (24). On the other hand, the variables "some college," "PCP rate," and "income ratio" were positively associated in Factor 6. These results are similar to some studies that found that education, primary care access, and income might have essential roles in cancer treatment and prevention (25,26). For example, use of breast and colorectal cancer screening is 20% to 30% lower among those with only a high school education than among college graduates (25). Also, a greater primary care physician supply was associated with lower cancer mortality (26), and low income is a barrier to health care access (25).

Although the ICRI explained only 32.2% of the variance in the age-adjusted CCMR, the combinations of the multiple cancer risk variables demonstrate the nature of cancer as a heterogeneous disease with many risk factors that may have a long-term impact on health (8). Although resource allocation can affect all the variables analyzed in this article, some variables, such as smoking and alcohol consumption, are also considered modifiable cancer risk factors. In addition, lifestyle modification could significantly reduce the burden of cancer (27); thus, public health professionals in Illinois may use the index to direct risk reduction and health promotion programs and policies at those counties most in need.

Our index explained the largest proportion of variance in prostate cancer mortality rates across Illinois and the smallest proportion for lung cancer. Many variables used in our study may relate to prostate cancer mortality in Illinois compared with other cancer types because early screening often can reduce prostate cancer risk for highly educated people with cancer screening resources (28). The low prediction of lung cancer mortality might be due to the noise introduced by smoking as the strongest risk factor for lung cancer (3). It also indicates that underlying latent sociodemographic variables other than smoking and air pollution could have an impact on lung cancer mortality in Illinois. Future studies could explore more variables specifically related to lung cancer mortality risk. Nevertheless, our index map showed a similar geospatial pattern that matches the county-level CCMR. Both maps demonstrate the compelling need for cancer-related public health resource allocation in east-central and northern Illinois.

Previous Illinois studies examined the associations between several risk factors and late-stage diagnosis for 4 common cancers and created a county-level index for breast cancer (10,11). Our study differs in that we comprehensively focused on ecosocial determinants of cancer risk factors according to ecosocial theory (8). Many variables used to construct our index are also associated with other cancer outcomes and other chronic noncommunicable diseases (27,29). This framework could potentially be used to create a similar scoring system for public health professionals.

Strength and limitations

Our study had several strengths. First, it incorporated numerous cancer risk factors at population levels to create a cancer risk index for Illinois. Our results identified physical and mental health variables and air toxin variables that previous studies did not include. Additionally, the use of a factor analysis framework can explore the underlying trends in the data, increase interpretability, and minimize information loss while reducing dimensionality. Furthermore, we identified and reduced factors from large feature sets associated with common cancer mortality in Illinois, and this information can assist in cancer intervention and prevention program planning at the county level.

Our study had some limitations. First, we only examined the state of Illinois. Further studies should explore larger data sets with counties from multiple US states and abroad. Second, we needed to impute data for some counties because of missing data; this was to ensure that an ICRI value was ascribed to each county. Third, we relied on data that were self-reported and aggregated at the county level, which may obscure nuances in individual behavior and be susceptible to social desirability bias. Fourth, caution must be applied in interpreting the index because we evaluated the ICRI by using its correlation with CCMR, and correlation does not signify causation. Despite these limitations, the calculated index provides informative data to advise public health professionals. To address these limitations in future studies, researchers can use larger and validated data sets and machine learning frameworks, which are becoming increasingly prevalent in cancer research, to model risk factors (30).

Conclusion

This study identified, reduced, and analyzed a substantial number of cancer predictors and incorporated them into a single novel county-level index for the state of Illinois. Our analysis found that the ICRI was moderately associated with the CCMR, which is the average mortality from the 4 most common cancers in the state. Public health professionals may use this framework to target resources and interventions to counties in Illinois that score highest on the risk index and are, therefore, most in need. Future research should apply this framework to construct indexes for other diseases and for multiple geographic locations.

Acknowledgments

Author Information

Corresponding Author: Lei Guo, MA, MS, School of Interdisciplinary Health Professions, Northern Illinois University, Wirtz Hall 323, DeKalb, IL 60115. Telephone: 734-358-3970. Email: rguo@niu.edu.

Author Affiliations: [1]School of Interdisciplinary Health Professions, Northern Illinois University, DeKalb, Illinois. [2]University of Illinois Cancer Center, Chicago, Illinois. [3]School of Health Studies, Northern Illinois University, DeKalb, Illinois. [4]Department of Engineering Technology, Northern Illinois University, DeKalb, Illinois.

References

1. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. CA Cancer J Clin 2022;72(1):7–33.
2. 2022–2027 Illinois comprehensive cancer control plan. Springfield (IL): Illinois Department of Public Health. Accessed August 30, 2022. https://dph.illinois.gov/content/dam/soi/en/web/idph/publications/idph/topics-and-services/diseases-and-conditions/cancer/2022-2027_IL-Comp-Cancer-Plan_03092022.pdf
3. Thun MJ, Linet MS, Cerhan JR, Haiman CA, Schottenfeld D, editors. Cancer epidemiology and prevention. Fourth edition. New York (NY): Oxford University Press; 2018.
4. Singh GK, Jemal A. Socioeconomic and racial/ethnic disparities in cancer mortality, incidence, and survival in the United States, 1950–2014: over six decades of changing patterns and widening inequalities. J Environ Public Health 2017;2017:2819372.
5. Hastert TA, Ruterbusch JJ, Beresford SA, Sheppard L, White E. Contribution of health behaviors to the association between area-level socioeconomic status and cancer mortality. Soc Sci Med 2016;148:52–8.
6. Jagai JS, Messer LC, Rappazzo KM, Gray CL, Grabich SC, Lobdell DT. County-level cumulative environmental quality associated with cancer incidence. Cancer 2017;123(15):2901–8. Erratum in Cancer 2019;125(10):1756.
7. James W, Jia C, Kedia S. Uneven magnitude of disparities in cancer risks from air toxics. Int J Environ Res Public Health 2012;9(12):4365–85.
8. Krieger N. Defining and investigating social disparities in cancer: critical issues. Cancer Causes Control 2005;16(1):5–14.
9. Scott LC, Barker LE, Richardson LC. Using population health measures to evaluate the environmental burden of cancer at the county level. Prev Chronic Dis 2019;16:E45.

10. Wang F, Luo L, McLafferty S. Healthcare access, socioeconomic factors and late-stage cancer diagnosis: an exploratory spatial analysis and public policy implication. Int J Public Policy 2010;5(2-3):237–58.

11. Wang F, Guo D, McLafferty S. Constructing geographic areas for cancer data analysis: a case study on late-stage breast cancer risk in Illinois. Appl Geogr 2012;35(1–2):1–11.

12. US Department of Health and Human Services. Area health resource files. Accessed November 14, 2021. https://data.hrsa.gov/topics/health-workforce/ahrf

13. County Health Rankings & Roadmaps. Accessed November 14, 2021. https://www.countyhealthrankings.org/

14. US Department of Agriculture. Food environment atlas. Accessed November 14, 2021. https://www.ers.usda.gov/data-products/food-environment-atlas/

15. Illinois Environmental Protection Agency Bureau of Air. Accessed November 14, 2021. https://www2.illinois.gov/epa/topics/air-quality/Pages/default.aspx

16. Finch WH. Exploratory factor analysis. In: Teo T, editor. Handbook of quantitative methods for educational research. Rotterdam (NL): SensePublishers; 2013.

17. Jbaily A, Zhou X, Liu J, Lee TH, Kamareddine L, Verguet S, et al. Air pollution exposure disparities across US population and income groups. Nature 2022;601(7892):228–33.

18. US Department of Commerce. American community survey; 2018. Accessed November 4, 2021. https://www.census.gov/programs-surveys/acs

19. Siegel RL, Jacobs EJ, Newton CC, Feskanich D, Freedman ND, Prentice RL, et al. Deaths due to cigarette smoking for 12 smoking-related cancers in the United States. JAMA Intern Med 2015;175(9):1574–6.

20. Nelson DE, Jarman DW, Rehm J, Greenfield TK, Rey G, Kerr WC, et al. Alcohol-attributable cancer deaths and years of potential life lost in the United States. Am J Public Health 2013;103(4):641–8.

21. Zhu J, Fang F, Sjölander A, Fall K, Adami HO, Valdimarsdóttir U. First-onset mental disorders after cancer diagnosis and cancer-specific mortality: a nationwide cohort study. Ann Oncol 2017;28(8):1964–9.

22. Berchuck JE, Meyer CS, Zhang N, Berchuck CM, Trivedi NN, Cohen B, et al. Association of mental health treatment with outcomes for US veterans diagnosed with non-small cell lung cancer. JAMA Oncol 2020;6(7):1055–62.

23. Blake KD, Moss JL, Gaysynsky A, Srinivasan S, Croyle RT. Making the case for investment in rural cancer control: an analysis of rural cancer incidence, mortality, and funding trends. Cancer Epidemiol Biomarkers Prev 2017;26(7):992–7.

24. O'Connor JM, Sedghi T, Dhodapkar M, Kane MJ, Gross CP. Factors associated with cancer disparities among low-, medium-, and high-income US counties. JAMA Netw Open 2018;1(6):e183146.

25. Goding Sauer A, Siegel RL, Jemal A, Fedewa SA. Current prevalence of major cancer risk factors and screening test use in the United States: disparities by education and race/ethnicity. Cancer Epidemiol Biomarkers Prev 2019;28(4):629–42.

26. Basu S, Berkowitz SA, Phillips RL, Bitton A, Landon BE, Phillips RS. Association of primary care physician supply with population mortality in the United States, 2005–2015. JAMA Intern Med 2019;179(4):506–14.

27. Stein CJ, Colditz GA. Modifiable risk factors for cancer. Br J Cancer 2004;90(2):299–303.

28. Kilpeläinen TP, Talala K, Taari K, Raitanen J, Kujala P, Pylväläinen J, et al. Patients' education level and treatment modality for prostate cancer in the Finnish Randomized Study of Screening for Prostate Cancer. Eur J Cancer 2020;130:204–10.

29. Rojas-Rueda D, Morales-Zamora E, Alsufyani WA, Herbst CH, AlBalawi SM, Alsukait R, et al. Environmental risk factors and health: an umbrella review of meta-analyses. Int J Environ Res Public Health 2021;18(2):704.

30. Abdullah Alfayez A, Kunz H, Grace Lai A. Predicting the risk of cancer in adults using supervised machine learning: a scoping review. BMJ Open 2021;11(9):e047755.

Tables

Table 1. Factor Loadings[a] and Sources of Data for Variables, the Illinois Cancer Risk Index, 2014–2018

Variable	Factor name[b]				Communalities[c]	Source
	1. Black race and health behaviors	3. Ethnicity, air quality, housing, and rurality	5. Financial needs and unemployment	6. Education, primary care provider, and income ratio		
Non-Hispanic Black population, 2014–2018, %	0.93[d]	—	—	—	0.96	County Health Rankings (13)
Non-Hispanic White population, 2014–2018, %[b]	—	−0.79	—	—	0.95	
Hispanic population, 2014–2018, %	—	0.79	—	—	0.90	
Rural population, 2010, %	—	0.75	—	—	0.81	
Current smoker[d], %	0.88	—	—	—	0.94	
Excessive drinking[e], %	0.85	—	—	—	0.90	
Poor physical health days[f], mean no.	0.79	—	—	—	0.85	
Poor mental health days[g], mean no.	0.62	—	—	—	0.89	
Limited access to healthy foods, 2014–2018, %	0.53	—	—	—	0.86	
Severe housing problems[h], %	—	0.73	—	—	0.86	
Carbon monoxide[i]	—	0.72	—	—	0.97	Illinois Environmental Protection Agency Bureau of Air (15)
Nitrogen oxides[j]	—	0.72	—	—	0.97	
Average daily PM$_{2.5}$[k]	—	0.72	—	—	0.65	County Health Rankings (13)
Sulfur dioxide[l]	—	0.72	—	—	0.97	Illinois Environmental Protection Agency Bureau of Air (15)
Volatile organic material[m]	—	0.72	—	—	0.97	
Population not proficient in English, %	—	0.69	—	—	0.90	County Health Rankings (13)
Population living in poverty, 2014–2018, %	—	—	0.82	—	0.93	Area Health Resources Files (12)

Abbreviations: — , variable not included in the model.

[a] Factor loadings: correlation coefficients between observed variables and common latent factors.

[b] Factor is a latent variable associated with a set of observed variables that have similar response patterns.

[c] The correlation coeffect r and P value of bivariate linear regression between factor and common cancers mortality rate.

[d] Percentage of adults who smoked 100 cigarettes in their lifetime and currently smoked, 2014–2018.

[e] Percentage of adults reporting binge or heavy drinking, 2014–2018. Binge drinking = consuming 4 or more drinks on one occasion for a woman or 5 or more drinks on one occasion for a man. Heavy drinking = 8 or more drinks per week for a woman or 15 or more drinks per week for a man.

[f] Number of physically unhealthy days in past 30 days (age-adjusted), 2014–2018.

[g] Number of mentally unhealthy days in past 30 days (age-adjusted), 2014–2018.

[h] Percentage of households with at least 1 of 4 housing problems: overcrowding, high housing costs, lack of kitchen, or lack of plumbing facilities, 2014–2018.

[i] Carbon monoxide stationary point source emission distribution at county level (tons/y), 2014–2018.

[j] Nitrogen oxides stationary point source emission distribution at county level (tons/y), 2014–2018.

[k] Average daily density of fine particulate matter in micrograms per cubic meter (PM$_{2.5}$), 2014–2018.

[l] Sulfur dioxide stationary point source emission distribution (tons/y), 2014–2018.

[m] Volatile organic material point source emission distribution (tons/y), 2014–2018.

[n] Adults aged 25–44 years with some post-secondary education, 2014–2018.

[o] Ratio of household income at the 80th percentile to income at the 20th percentile at county level, 2014–2018.

(continued)

Table 1. Factor Loadings[a] and Sources of Data for Variables, the Illinois Cancer Risk Index, 2014–2018

Variable	Factor name[b]				Communalities[c]	Source
	1. Black race and health behaviors	3. Ethnicity, air quality, housing, and rurality	5. Financial needs and unemployment	6. Education, primary care provider, and income ratio		
Median household income, 2014–2018	—	—	−0.81	—	0.95	County Health Rankings (13)
Median home value, 2014–2018	—	—	−0.66	—	0.94	Area Health Resources Files (12)
Median gross rent, 2014–2018	—	—	−0.60	—	0.89	
Unemployment rate, 2014–2018	—	—	0.56	—	0.72	USDA Food Environmental Atlas (14)
Some college[n]	—	—	—	−0.51	0.78	County Health Rankings (13)
Ratio of population to primary care physicians	—	—	—	0.55	0.7	
Income ratio[o]	—	—	—	0.54	0.743	
Correlation coeffect r and P value	$r = 0.37$, $P < .001$	$r = 0.36$, $P < .001$	$r = 0.04$, $P = 0.03$	$r = 0.26$, $P < .001$	—	—

Abbreviations: — , variable not included in the model.

[a] Factor loadings: correlation coefficients between observed variables and common latent factors.

[b] Factor is a latent variable associated with a set of observed variables that have similar response patterns.

[c] The correlation coeffect r and P value of bivariate linear regression between factor and common cancers mortality rate.

[d] Percentage of adults who smoked 100 cigarettes in their lifetime and currently smoked, 2014–2018.

[e] Percentage of adults reporting binge or heavy drinking, 2014–2018. Binge drinking = consuming 4 or more drinks on one occasion for a woman or 5 or more drinks on one occasion for a man. Heavy drinking = 8 or more drinks per week for a woman or 15 or more drinks per week for a man.

[f] Number of physically unhealthy days in past 30 days (age-adjusted), 2014–2018.

[g] Number of mentally unhealthy days in past 30 days (age-adjusted), 2014–2018.

[h] Percentage of households with at least 1 of 4 housing problems: overcrowding, high housing costs, lack of kitchen, or lack of plumbing facilities, 2014–2018.

[i] Carbon monoxide stationary point source emission distribution at county level (tons/y), 2014–2018.

[j] Nitrogen oxides stationary point source emission distribution at county level (tons/y), 2014–2018.

[k] Average daily density of fine particulate matter in micrograms per cubic meter ($PM_{2.5}$), 2014–2018.

[l] Sulfur dioxide stationary point source emission distribution (tons/y), 2014–2018.

[m] Volatile organic material point source emission distribution (tons/y), 2014–2018.

[n] Adults aged 25–44 years with some post-secondary education, 2014–2018.

[o] Ratio of household income at the 80th percentile to income at the 20th percentile at county level, 2014–2018.

Table 2. Associations Between Retained Factors and the Illinois Cancer Risk Index (ICRI) with the Common Cancer Mortality Rate, 2014–2018

Factor name	Multivariable linear regression[a]			Bivariate linear regressions				
	β	SE	P value	Intercept	β	SE	P value	r^2
Intercept	20.65	0.02	<.001	—	—	—	—	—
Factor 1	0.23	0.08	<.001	20.64	0.76	0.09	<.001	0.14
Factor 3	2.11	0.07	<.001	20.65	1.49	0.08	<.001	0.13
Factor 5	0.04	0.08	.01	20.64	0.19	0.09	.03	0.002
Factor 6	0.38	0.11	.004	20.66	0.65	0.09	<.001	0.07
ICRI	—	—	—	20.63	1.09	0.16	<.001	0.32

Abbreviations: — , variable not included in the model.
[a] r^2 = 0.335.

Table 3. Regression Analysis, Illinois Cancer Risk Index (ICRI) Relative to Breast, Colorectal, Lung, and Prostate Cancer Mortality Rates, 2014–2018

Age-adjusted cancer mortality rates (2014–2018)	Bivariate linear regressions				
	Intercept	β	SE	P value	r^2
Breast cancer	12.483	1.37	0.20	<.001	0.57
Colorectal cancer	12.324	0.10	0.13	<.001	0.56
Lung cancer	51.500	0.39	0.25	.02	0.23
Prostate cancer	8.116	1.33	0.16	<.001	0.64