We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

9

The Health Care Access Index as a Determinant of Delayed Cancer Detection Through Principal Component Analysis

Eric Belasco[1], Billy U. Philips, Jr.[2] and Gordon Gong[2]
[1]*Montana State University, Department of Agricultural Economics and Economics,*
[2]*Texas Tech University Health Sciences Center, F. Marie Hall Institute of Rural Community Health,*
USA

1. Introduction

In the past two decades, cancer mortality declined significantly in the United States (Byers, 2010). Although the reasons for the decline have not been well-established, many factors such as the reduction in the number of smokers, increased cancer screening, and better treatment may have played an important role (Byers, 2010, Richardson et al. 2010). However, disparities in cancer mortality persisted among different ethnic groups and social classes (Byers, 2010). Health status and health disparities among different social and ethnic groups are to a large degree determined by socioeconomic status and living conditions in general (Pamies and Nsiah-Kumi, 2008; World Health Organization [WHO], 2008). For example, life expectancy worldwide increased from 48 years in 1955 to 66 years in 2000 mainly as a result of improvement of overall living conditions in addition to advancement in medical science and large-scale preventive interventions (Centers for Disease Control and Prevention [CDC], 2011). Large health disparities exist between poor and rich countries or within any given rich or poor country (WHO 2008). In the case of cancer mortality due to delayed detection, socioeconomic status may determine health insurance coverage status, which in turn affects health behaviour including regular check-ups and participation in cancer surveillance among high risk groups. Regular cancer surveillance is critical for cancer control (Byers, 2010, Richardson et al. 2010). Lack of health insurance due to economic hardship may result in the delay in cancer detection.

A vexing question is how to determine socioeconomic status. Early studies associated cancer mortality with single socioeconomic indicators such as individual income, education level, below or above poverty level among others. For example, Ward et al. (2004) used the percentage of the population below the poverty line as a socioeconomic indicator and found that cancer mortality rate was 13% and 3% higher in men and women, respectively, among U.S. counties with ≥ 20% of the population below the poverty line as compared with those with < 10% below the poverty line from 1996 to 2000. On the other hand, Clegg et al. (2009) used education level as an indicator, and reported that lung cancer incidence was significantly higher among Americans with less than a high school education than those

with a college education. Clegg et al (2009) also used family annual income as an indicator and found that lung cancer incidence rate was 70% higher in those with less than $12,500 annual income compared with those with incomes $50,000 or higher.

An alternative approach to assessing socioeconomic status is to build a composite index based on many aspects of socioeconomic status using readily available census data, which is then employed to predict health status using information from disease (e.g., cancer) registries. For example, Singh (2003) used 17 socioeconomic indicators (such as household income, median home value) derived from US census data to build a composite index for socioeconomic deprivation by factor analysis and principal component analysis (PCA). Such a composite index is believed to reflect socioeconomic status more thoroughly with multiple indictors, while PCA address the issue of inter-correlation among factors. Such a composite index tends to have a high reliability coefficient (α equal to 0.95) (Singh, 2002). The composite index is then used to predict health status. For example, Singh (Singh, 2003) reported that US mortality of all-causes was significantly and positively correlated with a deprivation index derived from US census data. Crampton et al. (1997) developed a relatively simple socioeconomic status index termed the New Zealand Index of Relative Deprivation (NZDep91) which was constructed based on nine socioeconomic variables from New Zealand census data. The NZDep91 is subsequently used to predict hospital discharge rate and all-cause mortality (Salmond et al. 1998). Albrecht and Ramasubramanian (2004) (Henceforth, A&R) modified the NZDep91 and developed a Wellbeing Index (WI) by principal component analysis using ten socioeconomic variables from US Census 2000 data. The WI is recently shown to be highly correlated with delayed cancer detection (as assessed by the ratio of late- to early-stage cases) of female genital system (FGS), lung-bronchial and all-type cancers at diagnosis among Texas counties (Philips et al., 2011).

One of the main purposes of the current study is to determine whether the percentage of late-stage cancer incidence is correlated with a newly developed index of health accessibility, which is an extension of the previously mentioned WI. We term the new index the Health Care Accessibility Index (HCAI), which is derived from principal component analysis of ten socioeconomic indicators derived from US Census Bureau's 2005-2009 American Community Survey plus two additional factors that are more closely related to health, i.e., health insurance coverage and physician supply. By examining the relationship between HCAI and late-stage cancer detection, we are able to establish whether health inequities exist in certain communities that can be related to access to health care. A high percentage of late-stage cancer cases is problematic for communities due to the often relatively low survival rates of costly procedures. The derived HCAI is compared with WI in their association with delayed cancer detection in Texas counties to determine the optimal model by the Akaike's Information Criteria (AIC) (Akaike, 1974) and Schwartz Information Criteria (SIC) (Schwartz, 1978). Another difference between WI and HCAI is that the ten socioeconomic variables for computing the HCAI includes a new variable, the median income of a county and excludes the percentage of people with disability (because of its absence in the US Census Bureau's 2005-2009 American Community Survey database).

This study also addresses several practical statistical issues regarding the choice of socioeconomic variables in association with delayed cancer detection. Firstly, A&R arbitrarily classify the WI rankings as the deciles of the first component scores retrieved from PCA. It is quite frequent for groupings to be assigned based on terciles, quartiles, or

deciles, which may or may not produce a proper classification. If ad hoc metrics are used to denote an optimal number of groups to be used in any particular index, then the inference that derive from the index might not have meaning and might even add uncertainty to the representativeness of the index. For example, the grouping used for Texas might be substantially different from Delaware, given the different sizes and regional dynamics. We propose to use AIC and SIC to find the optimal number of groups and compare the goodness-of-fit between models. Secondly, it has not been statistically demonstrated that using the composite WI is superior to using each of the 10 individual socioeconomic variables in their correlation with health status such as delayed cancer detection. The optimal model of the two (a composite vs. multiple variables) will also be determined by AIC and SIC. Thirdly, we also propose the use of raw principal component scores in a regression in order to improve goodness-of-fit and compare these results to the previously mentioned models based on AIC and SIC. Using these proposed methods, we look to characterize the relationship between late-stage cancer detection with the HCAI in order to identify the existence or lack of existence of economically-rooted health inequities.

2. Review of PCA in regression analysis

2.1 Review of creating indices using PCA

Using PCA in economics is particularly appealing for applications where comparisons are warranted that comprise over a collection of variables. This method is particularly convenient and informative when the researcher is interested in the "ability" of a model to characterize a collection of variables rather than the marginal impacts between one variable and another (Greene, 2012, pg. 93). Further, the assumption of an exogenous shock is often used when evaluating marginal impacts, which are unrealistic with covariates that are highly correlated. For example, in evaluating the relative wealth or poverty in developing counties, asset indices are often built to reflect the relative wealth in order to make cross-country comparisons. For example, Booysen et al (2008) conducted a transregional survey in sub-Saharan Africa to evaluate the movement of poverty across regions over a particular time span. While "poverty" is a loosely defined term, this line of research commonly utilizes an asset index in order to evaluate the ability of citizens in each country to consume durable goods. Another example is Gwatkin et al. (2000) who use data from the Demographic and Health Survey (DHS) program to evaluate socio-economic differences between developing countries. They create a socio-economic status (SES) index, which was also used in studies such as Vyas and Kumaranayake (2006).

SES indices have also been developed in order to evaluate health outcomes in developing counties in studies such as Deaton (2003), where it is argued that health outcomes and the utilization of health services are largely different across different socioeconomic classes. Ruel and Menon (2002) create a child feeding index using responses in the DHS survey to assess the influences on child feeding practices. As in many related studies, the creation of an index is used to identify problem areas that need improved policy design or interventions.

Ewing et al. (2008) develop a sprawl index in order to evaluate the relationship between urban sprawl and health-related behaviours. As in the case of other index variables, urban sprawl can be defined only when many factors are combined. The main four factors that are

include are residential density, mixture between homes, jobs, and services, strength of business centers, and accessibility of the street network. Data were from the Behavioral Risk Factor Surveillance Systems (BRFSS) from 1998 to 2000 to evaluate health activity and lifestyle. On a similar topic is a study by Pomeroy et al (1997) who evaluated perceptions and the factors that led to the appearance of a successful resource management program.

While many of these indices include continuous variables, Koenikov and Angeles (2009) and Filmer and Pritchett (2001) explore ways to incorporate discrete data into PCA. Savitry (2005) provides a comprehensive overview of all multivariate methods used in developing indices.

2.2 Review of using PC scores in regression analysis

While the developments of a reliable and robust index are quite extensive in the literature, little research has been done to evaluate the impact of incorporating an index into a regression. This is precisely because usually the index itself is the primary interest. For example, in the identification of poverty traps, one needs only an index for identification. However, if one were interested in the impact of a poverty trap on say personal liberties, then a regression would likely be needed.

There are many studies that conduct PCA to create an index that is used in a secondary regressions (Everitt and Hothorn, 2011; Everitt, 2011; Vyas and Kumaranayake, 2006). One method that is used is to rescale the principal component values. For example, Ewing et al. (2008) scale the raw principal component scores to have a mean of 100 and a standard deviation of 25. This is similar to normalizing the principal components. However, the distribution of scores might not be normal. Vyas and Kumaranayake (2006) report finding scores that are different distributions by county that include counties with normal and uniform distributions along with distributions that possess negative or positive skewness. Their results suggest that the characterization groups might be different across cases, which leads one to consider a data-driven approach. Pomeroy et al (1997) use the raw and unscaled principal component scores as the dependent variable. While marginal impacts are not easily interpreted, they can loosely be discussed directionally.

Another common approach is to use cut-off points. For example, Filmer and Pritchett (2001) split their sample into three populations based on cut-off points at the 60^{th} and 20^{th} percentile of the principal component scores. They define these groups to be 'low', 'medium', and 'high' socioeconomic groups. Cut-off points can also be defined a bit more arbitrarily. For example, A&R use deciles (ten groups) to define different ranges of socio-economic status. The implicit assumption made is that each group is distinctly different from the other. However, just as Vyas and Kumaranayake (2006) found differences by country it might also be possible to find that some populations need more grouping than others. Others, such as Booyens et al (2008) use quintiles in breaking up the population.

Other approaches that are more data-driven include the use of cluster analysis, which is described in some detail in Everitt et al. (2011). However, one limitation of cluster analysis as well as the previously discussed methods is that the existence of an observation in a particular group is mutually exclusive of other groups. This means that each observation must with certainty fit into a single category. However, an evolving area of research includes latent class analysis where the residency within each group is treated as an

unobserved factor that is estimated with some probability, rather than absolute certainty. However, these models are also with their own set of limitations. For example, the number of classes must be defined *a priori*. Additionally, the models have been shown to be highly non-linear and can often have difficulty finding optimized values when many classes are added.

3. Data and application

3.1 Data summary and description

This study was approved by Texas Tech University Health Sciences Center Institutional Review Board with exemption for review because of its use of published data.

3.2 Cancer data

Cancer stage data from 2004 to 2008 were provided by the Texas Cancer Registry, Cancer Epidemiology and Surveillance Branch, Texas Department of State Health Services (note that cancer data from 2008 are the latest data available). This database provides cancer data by year, age, county, Hispanic origin (Hispanic vs. Non-Hispanic) as well as population size for each county. We used cancer data between 2004 and 2008 cancer data (the latest available data is the 2008 data) to match the five years of American Community Survey data although there is a one-year lag. The five most common categories of cancer are studied including breast, colorectal, FGS, lung-bronchial, and prostate cancers. Female genital system includes cervix uteri, ovary, corpus and uterus, vagina, vulva and others. We pool the five-year (2004-2008) data to calculate the numbers of age-adjusted late- and early-stage cancer cases per unit (100,000) population using 2000 USA standard population (National Cancer Institute, NCI, n.d.). We use the percentage of late-stage cases among all staged cancer cases in our analysis. The number of unstaged cancer cases is not included in the denominator because the percentage of unstaged cases varies significantly by cancer type as well as by county, and inclusion of such cases in the denominator would result in uncertainty in estimating the percentage of late-stage cancer cases. Carcinoma in situ and localized cancers are considered as early-stage while cancers defined as "regional, direct extension only", "regional, regional lymph nodes only", "regional, direct extension and regional lymph nodes", "regional, NOS" and "distant" are considered as late-stage (Philips et al. 2011).

3.3 Socioeconomic status data

Socioeconomic status data are derived from the U.S. Census Bureau's (n.d.) 2005-2009 American Community Survey. Since this survey does not provide percentage of people with disability, the present study uses the remaining nine of the 10 socioeconomic variables originally used to build the Wellbeing Index (WI) developed by A&R and are listed in Table 1. We add median income (from the 2005-2009 American Community Survey) so that the total number of socioeconomic variables is still ten in the current study.

3.4 Data of factors more closely related to health

Data for the percentage of uninsured and percentage of obese individuals are obtained from Texas State Data Center (n.d.). The number of physicians and estimated population size in

each county from 2004 to 2008 are derived from Texas Department of State Human Services (DSHS, n.d.). Physician supply is the number of physicians per 1,000 residents in each county. Physicians considered are those with medical doctor (MD) and/or doctor of osteopathy (DO) degrees who worked directly with patients. Residents and fellows; teachers; administrators; researchers; and those who were working for the federal government, military, retired, or not in practice were excluded from the total of physicians by DSHS (n.d.).

	Description	Weighted Mean	Std Error	Min	Max
% Single Parent	% people in single parent households	18.89	0.54	0.00	27.55
% No High School	% people over 18 without high school	21.38	1.17	2.60	53.59
% Unemployed	% people unemployed	6.87	0.15	0.00	21.58
% Income Support	% people with income support	28.20	1.47	0.00	66.64
% Below Poverty	% people in households below poverty level	15.90	0.88	0.00	46.81
% No House	% people not living in own home	34.19	1.74	11.88	70.83
% Few Room	% people living in homes with too few bedrooms	4.97	0.47	0.00	15.79
% No Phone	% people in households without phone	5.33	0.20	0.00	18.23
%No Car	% people in households without car	6.55	0.40	0.00	17.55
Median Income	Median Income (in $1,000s)	49.63	1.73	20.38	82.55
%Uninsured	% people without health insurance	24.47	1.07	14.20	38.10
Physician supply	Number of physicians per 1,000 residents	0.16	0.01	0.00	0.33
% Obese	% people with BMI above 30	28.90	0.33	23.80	32.80
% Hispanic	% Hispanic	35.60	3.19	2.44	97.15
Population	Population (in 1,000s)	636.02	218.75	0.02	1,912

Table 1. Description and weighted summary statistics of relevant covariates

4. Statistical analysis and results

4.1 Computing the health care accessibility index

While access to health care may be determined by socioeconomic status in general, health insurance coverage and health care services (number of physicians and or hospitals relative to local population) may more directly impact on health as discussed above. In this study we add the latter two variables to the 10 socioeconomic indicators for a principal component analysis to build Health Care Accessibility Index (HCAI).

There are two well-known benefits to creating an index rather than including all of the components in the index into a regression. First, the index itself can serve as a variable of interest to identify (in our study) areas of extreme health care obstacles. Second, the index can significantly reduce the degrees of freedom in a regression while preserving information from the variables. In this study we also find a third benefit, which is that if we are careful about how we use the index in regression analysis, we can also improve goodness-of-fit.

While the A&R study presents an appealing start to our research, it lacks two important components. First, it only includes socioeconomic variables that are intended to provide an index for wellbeing or economic deprivation. While their index provides a good basis to evaluate the ability to pay for health services, it lacks health-specific variables. Second, the study presents a less than appealing method for incorporating the index into regression by ranking the first principal component scores into deciles. This essentially groups the observations into 10 different bins which, may or may not be the optimal number of bins to use in grouping regions. To fix the second issue of rank ordering, we use AIC and SIC to determine the optimal number of bins (more discussion on this follows in the next subsection).

In order to fix the first issue, we add three components which are essential to access to health care services which include median income, the percentage of the uninsured, and the number of physicians per 1,000 residents. All variables are computed at the county-level. These variables are intended to account for the access residents within a county have to health care services. Lack of health care insurance in clearly a hurdle in obtaining affordable health care services and is another factor included in this analysis. Finally, the number of physicians allows for an evaluation into geographic access to health care services. Does the county in question have an adequate medical infrastructure to prevent and detect illness when it arises? Some Counties in Texas (particularly in the western region) are geographically isolated. If adequate care is not geographically close to residents, the economic cost to receiving care increases in terms of time off work and travel costs.

In compiling the twelve variables, we are able to derive an index for access to health care services through the use of PCA. While the goal is to consolidate a group of many variables into a smaller set of linearly related variables (principal components), it is often the case that multiple principal components are needed to explain a substantial proportion of the variation in the independent variables. Results from the initial PCA is shown below in table 2.

The first principal component explains 41% of all variation in the variables included in the index. The influence is also spread across relatively evenly across all included variables, with the exception of direct patient physician supply (*DPC*). The principal component scores from this regression are used as the HCAI. Notice the negative score associated with median income, which is consistent with a negative relationship between income and obstacles to receiving health care services. The second principal component explains an additional 16% of the variation in the variables and is largely influenced by physician supply, percent uninsured, percent without their own house, and percent of single parents.

These results provide us with a couple of items. First, we have the first five principal components, which can be used rather than the original 13 variables in order to shrink the necessary variables while still preserving almost all of the variation in the variables. One notable and helpful point in this analysis is that each component is orthogonal to the others,

	Principal Component Score				
Variable	1	2	3	4	5
% Single Parent	0.31	0.22	0.22	0.01	0.21
% No High School	0.36	-0.13	-0.36	-0.03	0.15
% Unemployed	0.24	0.05	0.31	0.71	-0.43
% Income Support	0.18	-0.56	0.23	-0.05	0.28
% Below Poverty	0.39	-0.08	0.09	0.00	-0.02
% No House	0.18	0.45	0.21	-0.30	-0.24
% Few Room	0.27	0.17	-0.48	0.24	0.17
% No Phone	0.28	0.09	-0.04	-0.50	-0.48
%No Car	0.36	-0.03	0.22	0.12	0.05
Median Income	-0.35	0.31	-0.18	0.26	-0.08
%Uninsured	0.31	0.24	-0.41	0.07	0.11
DPC	-0.01	0.47	0.37	-0.04	0.57
Eigenvalue	4.89	1.94	1.36	0.76	0.74
Difference	2.95	0.58	0.60	0.02	0.11
Proportion	0.41	0.16	0.11	0.06	0.06
Cumulative	0.41	0.57	0.68	0.75	0.81

Table 2. PCA results from variables in health access index

meaning the independent variables will be uncorrelated and avoids the issue of multicollinearlity which arises from including all 13 variables into the regression. A notable problem when multicollinearity is particularly acute is that it is difficult to isolate marginal relationships between competing variables, which often leads to high standard errors. Second, the first principal component scores can be used to develop the index of interest in the following way. The index is created by recognizing that the first principal component can be written as a linear combination of the original variables such that

$$Prin1 = a_{1,1}x_1 + a_{1,2}x_2 + \cdots + a_{1,13}x_{13} \quad (1)$$

where $a_1 = (a_{1,1}, a_{1,2}, \ldots, a_{1,13})$ include the parameter estimates in table 2 and $x = (x_1, x_2, \ldots, x_{13})$ include the 13 parameters used in the index from table 1. $Prin1$ is then computed for each observation and can be ranked to present a ordering of health accessibility. The second component is then derived based on the remaining variability and results in a_2 which leads to $Prin2$. The second component is derived based on the restriction that $a_2'a_1 = 0$ so that $Corr(Prin1, Prin2) = 0$. Additional components can be derived in the same fashion so that they are uncorrelated with all prior components.

The rankings associated with the first principal component, $Prin1$, is used as the basis of our HCAI. To illustrate the regional dynamics of this new index, we present figure 1 below.

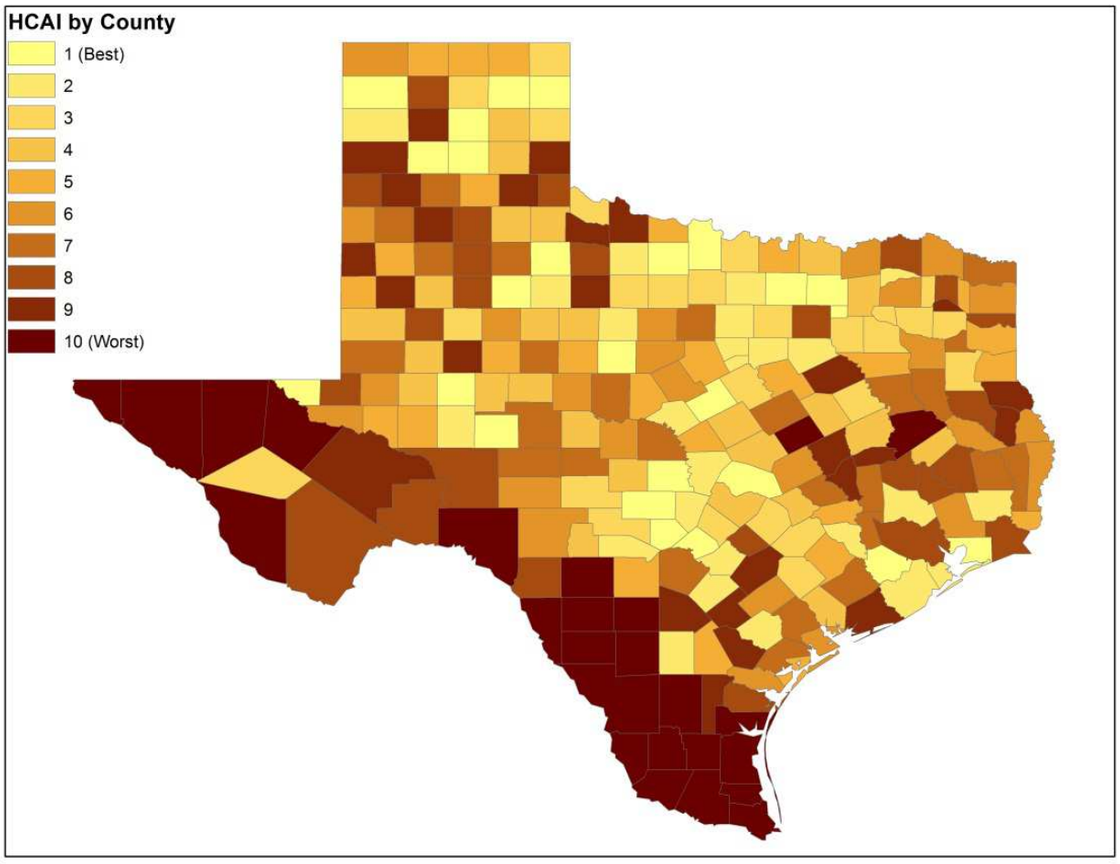


Fig. 1. Map of ordinal ranking of Texas counties by HCAI

Figure 1 shows that the counties that have poor health care access are focused along the Mexican border. These counties tend to have a large percentage of Hispanics in the population. Other counties further north are also found to have high HCAI rankings. Some of the better HCAI rankings are in the more urban areas of Texas that are near the largest metropolitan areas of Dallas, Fort Worth, Houston, San Antonio, and Austin. The large size of Texas, as well as its diversity in the variables used in this study, provides an opportunity to evaluate the how these factors influence late-stage cancer detection.

In the present study, the percentage of late stage cancer cases is hypothesized to correlate with the developed HCAI. Figure 2 below shows a scatter plot of the dependent variables along with the first component scores of HCAI for each cancer type that is evaluated in this study. Given the principal component scores shown in table 2, it is clear that a county that has a low degree of health access will possess a high component score. These same areas with a low degree of health access correspond to high rankings. Given this, it is not surprising that some of the slopes shown in figure 2 are significantly positive. Lung cancer is one example where the positive slope is particularly striking. On the other hand, the

relationship with the HCAI is shown to be positive but relatively flat for colon and prostate cancers. Given the different speeds of progression in different cancers, it should not be surprising that the relationships differ across cancer types.

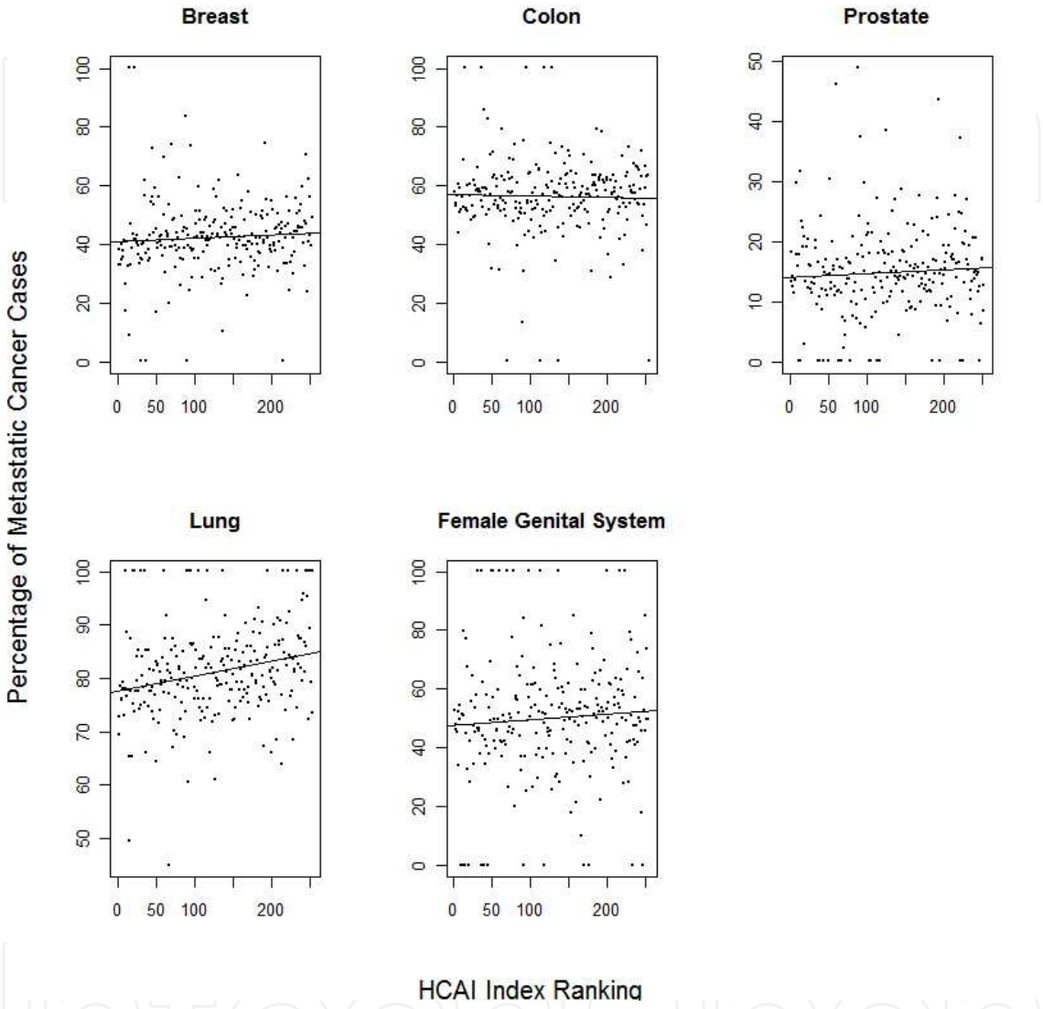


Fig. 2. Plot of the percentage of late-stage cancer cases with HCAI first principal component score, by cancer type

4.2 Weighted tobit regression model

The dependent variable of interest used in this analysis is the log of the sum between one and the percentage of late-stage cancer cases among all staged cancer cases. Some counties experienced no late-stage cancer cases and other experiences all late-stage cancer cases. These are all from relatively small counties. However, this does mean that censoring is an issue that will need to be dealt with in this study. As shown in table 3, there is some degree of censoring for all regressions in our analysis. For example, the percentage of late stage breast cancer cases was lower censored at 0 in four cases and upper censored in two cases. In the lower censoring case, the county experienced no late-stage cancer cases but

experienced only early-stage cases. Alternatively, some counties experienced only late-stage cases resulting in an upper bound of 100%. This double-bounded censoring occurred for all cancer types to varying degrees.

If censoring is not accounted for and usual least squares methods are used, we are assured of biased estimates (Amemiya, 1973). So, in order to obtain unbiased parameter estimates, we use the Tobit model (Tobin, 1958) to correct for censoring. The Tobit model assumes the dependent variable, Y, is a partially observed derivative of a latent variable, Y^*. When censoring occurs at zero, the two are related with the following relationship, $Y = \max(0, Y^*)$. In this particular application of a double-bounded Tobit model, we assume a latent variable, Y^*, with the following relationship to the observed variable, Y.

$$Y = \begin{cases} 0 & if\ Y^* < 0 \\ Y^* & if\ 0 \le Y^* \le 100 \\ 100 & if\ Y^* > 100 \end{cases} \quad (2)$$

A more comprehensive discussion on the Tobit model can be found in Amemiya (1974). Residuals are weighted by population size in order to account for the wide variety in the size of counties. Weighted statistics provide a more accurate representation of the population of interest, which is the state of Texas in this application.

Type	Weighted Mean	Std Error	Min	Max	Limit obs.	
					Lower	Upper
Breast	40.46	0.48	0.00	100.00	4	2
Colon	57.92	0.81	0.00	100.00	4	5
Prostate	14.45	0.41	0.00	48.87	18	0
Lung	79.28	0.59	44.98	100.00	0	20
FGS	50.69	0.55	0.00	100.00	13	13

Table 3. Percentage of late-stage cancer cases, by cancer type

4.3 Incorporating the index into regression analysis

While the selection of principal components is well established and can be found in most intermediate econometrics text books, the usage of that information in a regression is less understood. For this reason, we use five separate regressions for each cancer type to compare different methodologies that could be used to integrate an index or information in an index within a regression context. We use AIC and SIC to compare model fit and include two additional variables that are hypothesized to influence the percentage of late stage cancer cases. These variables include the log of the percentage of obese individuals and the log of percentage of Hispanic individuals. The five models of interest include the following specifications:

Model 1. The composite ordinal WI (from A&R) based on deciles according to the first principal component scores

Model 2. All 12 variables in the composite HCAI

Model 3. The raw scores of the first two principal components using variables in HCAI

Model 4. The composite ordinal HCAI based on grouping of the first principal component scores that minimizes AIC

Model 5. The composite ordinal HCAI based on grouping of the first principal component scores that minimizes SIC

The first method uses the method proposed by A&R, which assumes the use of deciles to group the ordered WI. This model provides a prior baseline under which to evaluate all of the proposed models. As A&R point out, once the index is used to define groups, those groups are compared in an ordinal manner, implying the se of ordinal and not continuous variables in a regression. Model 2 does not make use of PCA and places all the variables used in the HCAI directly into the regression. This model also provides another baseline model in that we can compare the methods used in PCA to that of this model which simply uses the variables without any PCA. In order for PCA to be effective in our analysis, it will need to demonstrate an improvement over model 2. In our original analysis we included two additional regressions that were omitted from our final analysis because they were not found to improve upon the utilized models. These models included (1) the raw WI variables and (2) decile grouping from the HCAI. Using raw WI variables was consistently outperformed by model 2, while the decile grouping of HCAI was consistently outperformed by models 4 and/or 5 unless they selected 10 groups (as occurred in a couple of instances).

The final three models make use of the PCA results based on the HCAI. These are the models of interest in the sense that we hypothesize that they will improve upon Models 1 and 2. Model 3 uses the raw scores of first two principal components in the regression. The scores of first two components are used since they minimize both AIC and SIC in all of the used regressions. Component scores are fitted values using the parameter values listed in table 2 and the associated variables for each observation. Therefore, each observation will contain a unique component score. The first two principal component scores explain 57% of the total variation in the used variables based on the results in table 2. While it might seem more appropriate to include say the first five principal components in order to explain 81% of the total variation, the final three components tend to be significant are the model does not fit the data as well (in terms of AIC and SIC) as when only the first two scores are included. This model provides a set of continuous variables that can be used to determine late-stage cancer detection. Models 4 and 5 use AIC and SIC to find the optimal number of groups to be used for HCAI. This search method is conducted by running all models up to 40 groups in order to find the single model that minimizes AIC or SIC. Observations are evenly divided across the selected number of groups which are used as categorical variables in our analysis.

These five models are then compared by using AIC and SIC metrics to assess goodness-of-fit. In order to declare one of the proposed models as an improvement, it will need to have values of both AIC and SIC that are lower than models 1 and 2. Since maximum likelihood methods are commonly used to estimate Tobit estimates, an appealing goodness-of-fit measure is that of AIC and/or SIC since they are both easily derived from maximum likelihood outputs. The two can be expressed as follows

$$AIC = 2k - 2LL \tag{3}$$

$$SIC = klog(k) - 2LL \quad (4)$$

where k is the number of parameters and LL is the maximized log-likelihood value that corresponds with the optimal parameter estimates. Both information criteria measure are composed of two parts, which include two times LL and the penalty for adding new parameters. For models that contain more than 8 parameters ($k > 8$), SIC provides a heavier penalty for adding new variables. For this reason, AIC tends to support more overfit models while SIC tends to support more underfit models. Both AIC and SIC provide a basis upon which to select from nested models by minimizing the associated criteria. However, Koehler and Murphree (1988) point out that in their analysis the two criteria indicate different models 27% of the time. Their research suggests that SIC is often preferred due to the improved predication accuracy from SIC-preferred models. However, as pointed out by Kuha (2004), the preference for AIC or SIC can often rely on the data generation process and when AIC and SIC both select the same model, it is shown to be a more robust result. With these points in mind, we proceed by using AIC and SIC criteria separately to determine the appropriate model and determine whether our proposed models provide an improvement on past models.

A&R suggest that when using their index of socioeconomic deprivation, one should rank the first principal component then group this ranking into deciles, so that groupings will occur at 10% intervals. For example, if a county receives a grouping of 1, it is found to be in the lower 10% with regard to the index. We find this ranking method to be arbitrary and therefore suggest the use of AIC/SIC to determine the optimal number of deciles. This is achieved by re-running each model with the number of groups changing every time starting with 1 group (or rather no grouping at all) and up to 40 groups. Forty groups were selected because AIC and SIC did not show significant improvement after that point. With such a large and diverse state as Texas, it seems that the same rank grouping would be substantially different than for some smaller state. Of importance for this study is the generalizeability of our study to other areas. For example, does the method that we propose provide a general enough method on which to evaluate any region? While the selection of a specific grouping may work best in one situation, a more data-driven approach (such as the one we propose) will provide more flexibility in evaluating different situations.

Also included in each model will be two additional variables in order to evaluate the impact of different specifications on parameter estimates, since typically marginal inference is an important component of any analysis (particularly in economics). These variables include the log of percentage of Hispanic and the log of percentage classified as obese. Both variables are likely to have some degree of correlation with the index ranking but to varying degrees. This provides a look at two variables that have varying relationships with the index in question.

4.4 Empirical results

4.4.1 Model selection

The first tasks in the empirical section of this analysis is to determine grouping based on the candidate model that minimizes AIC and SIC (models 4 and 5) for all 5 cancer types models (as outlined in section 3.2). The selected numbers of groups are shown in table 4.

	AIC	SIC
Breast	11	5
Colon	10	1
Prostate	9	1
Lung	34	8
FGS	1	1

Table 4. Number of groups determined based on minimum AIC and SIC criteria for models 4 and 5

The variability in the selected number of groups, suggests that arbitrarily extracting a grouping number might not be optimal. For example, in our analysis, lung cancer makes use of many groups while FGS cancer suggests only 1 group (which means no group distinctions at all). To illustrate the meaning of a group, the use of 34 groups implies there are 34 groups (each with 7-8 observations). In the regression, this implies that 33 new classification variables are added which are evaluated relative to the omitted group. Each model is run using the weighted Tobit model to assess the goodness-of-fit as well as parameter estimates of variables outside of the proposed index for all 5 cancer-types as well as with all cancer types. The regression results are below in table 5.

Table 5 shows the five models used in this study along the column headers, while row headers provide space for AIC and SIC measures for each cancer type. The purpose of this analysis is to evaluate the newly developed models (3-5) and compare them to two baseline models (1-2). Recall that in order for models 3-5 to unambiguously improve the baseline models 1-2, they must improve both AIC and SIC measures. Each model that improves upon the baseline models is shaded in grey in table 5.

For each cancer type, at least one of the three proposed models improved over the baseline models. Additionally, the use of a data-driven method such as AIC or SIC provides a clear improvement over the baseline models in each cancer-type. It is interesting to note that the best fitting model is not unambiguously from using AIC or SIC. This is a common problem that has been discussed in other studies such as Koehler and Murphree (1988). The only instance when AIC does not improve upon the base models is in the case of lung cancer, which suggests the use of 34 groups, which is unusually high based on the other selections. Alternatively, SIC provides a heavier penalty for new variables and tends to support lower parameterized models. This is clearly the case when SIC suggests the use of 1 group 60% of the time, where no group distinctions are provided. In two out of those three instances, SIC provided a clear improvement over baseline models. The use of raw principal component scores are also considered in model 3 and are shown to clearly improve on the baseline models in breast and FGS cancer types. In the case of breast cancer, it is interesting to note that this model provided the lowest AIC and SIC, making it a robust selection for best model fit.

The results from regressions on colon cancer provide some interesting insights. First, AIC suggests the use of deciles, which is in line with A&R. While this research makes the argument that deciles will not always be satisfactory, it can be optimal in some situations. It

is also worth noting that model 4 improves in both AIC and SIC over model 1. These two models are identical with the exception of the new index that is used in model 4, which indicates its improved predictive power over the WI. The model suggested by SIC is to include only 1 group, which are really no group distinctions at all. This also presents the possibility that when an index might not always provide valuable information for a model, which can only be detected through the use of a data-driven method, such as information criteria. However, model 5 is not the favored model in this scenario as it does not improve upon the earlier models. For this reason, model 4 is the preferred model for colon cancer.

		WI	HCAI			
		Deciles	All Vars	Prin Scores	Optimal AIC	Optimal SIC
Type	Model	1	2	3	4	5
Breast	AIC	-384.601	-389.795	-404.036	-392.419	-383.926
	SIC	-338.822	-333.451	-382.907	-343.118	-355.755
Colon	AIC	-397.097	-405.139	-395.164	-408.757	-392.295
	SIC	-351.266	-348.732	-374.011	-362.926	-378.193
Prostate	AIC	-24.605	-23.846	-25.446	-34.810	-28.660
	SIC	21.174	32.498	-4.318	7.447	-14.574
Lung	AIC	-859.488	-855.730	-851.945	-878.167	-869.566
	SIC	-813.554	-799.195	-830.744	-747.431	-830.699
FGS	AIC	-132.891	-131.983	-144.948	-143.995	-143.995
	SIC	-87.269	-75.833	-123.891	-129.958	-129.958

Table 5. Weighted Tobit goodness-of-fit regression results from 5 alternative models, by cancer type (Dependent variable: logged sum of one plus percentage late-stage cancer cases)

Based on figure 3, lung cancer is one particular type of cancer that is expected to be highly correlated with the HCAI. Because of the high degree of positive correlation, it is not surprising that AIC and SIC both suggest relatively high grouping values. For example, AIC suggests using 34 groups, while SIC suggests the use of 8 groups. However, while the AIC selected model has a very low AIC value (-878.167), SIC is not as impressive given the relatively large penalty factor for each of the 34 groups. Thus, the SIC selected group is able to improve over the base models in terms of AIC and SIC and is therefore the preferred model for lung cancer.

Many variables under study are inter-correlated. For example, figure 3 shows that percentage of the uninsured in Hispanic is significantly higher than that in non-Hispanic populations in Texas. Figure 4 shows that the percentage of late-stage cancer cases in Hispanics is higher than non -Hispanics in Texas. Hispanics also tend to have higher percentage obese individuals and socioeconomically deprived individuals. These facts suggest the necessity to adjust for covariates in assessing the association between the HCAI and delayed cancer detection. This will also allow determination of the role of ethnicity in the delayed cancer diagnosis.

We acknowledge in this research that not all types of cancer will have a significant relationship with the HCAI index. While the types of cancers that do have a correlation show health inequities, not all cancer types are thought to have inequities. In particular, we expect for cancers where detection is at a high cost (such as breast, colon, and lung cancer) to be particularly susceptible to health inequities. For example, in order to detect lung colon cancer, a costly colonoscopy is necessary which will have a lower compliance rate in individuals with poor access to health care services.

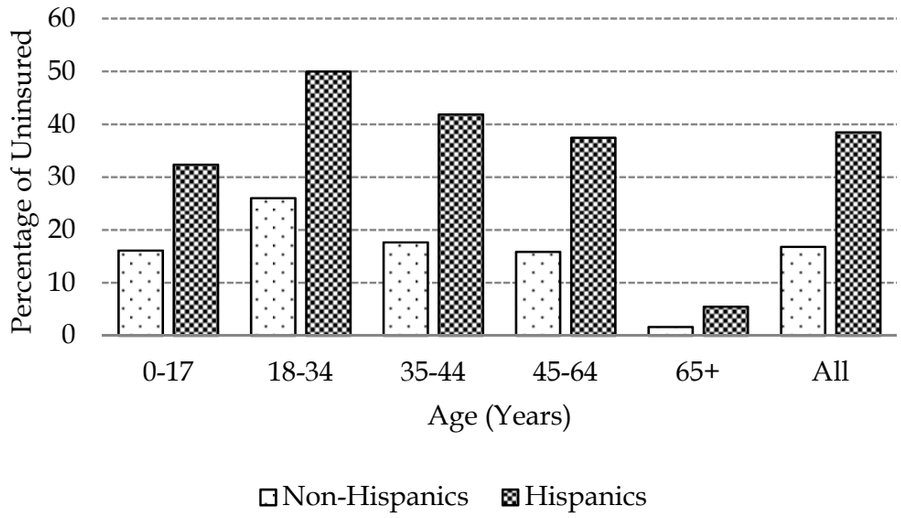


Fig. 3. Percentage of the Uninsured in Hispanic and Non-Hispanic populations in Texas.

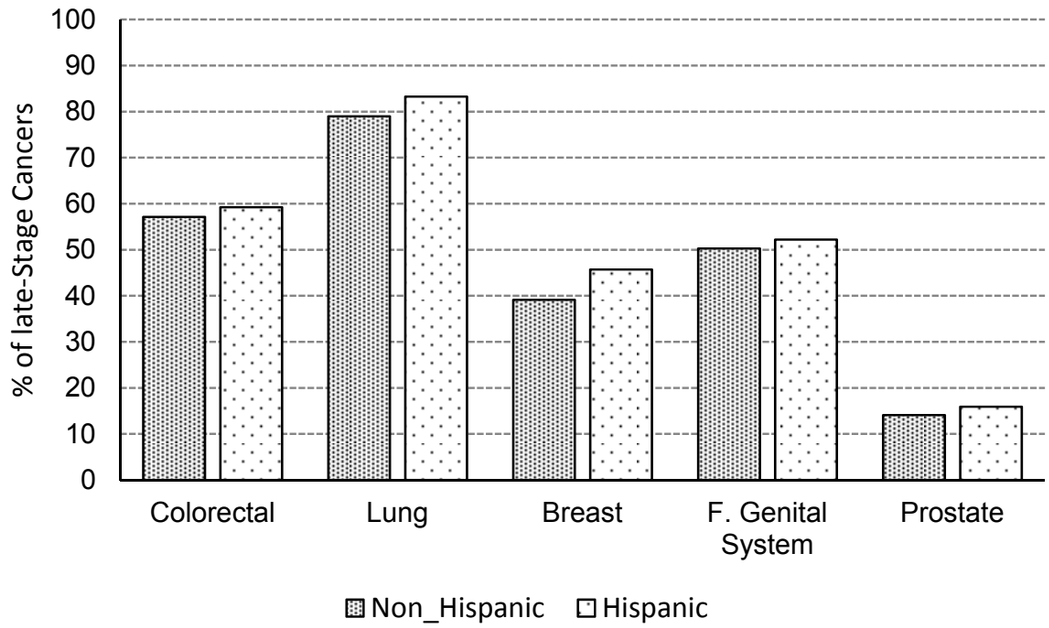


Fig. 4. Percentage of late-stage cancer cases in Hispanics and non -Hispanics in Texas.

4.4.2 Regression results

Below are included the parameter estimates resulting from the optimally selected model based on the criteria stated above. In all four cases, at least one of the newly proposed models outperformed the baseline models in terms of AIC and SIC. In the case of breast cancer, late-stage cancer detection was best determined using the raw principal component scores. Parameter estimate results are listed in table 6.

Variables	Estimate	SE	p-value
Intercept	2.800	0.691	<.0001
log(% Hisp)	-0.039	0.019	0.040
log(% Obese)	0.325	0.218	0.136
Prin1	0.024	0.005	<.0001
Prin2	-0.021	0.006	0.001
Sigma	0.106	0.005	<.0001

Table 6. Weighted Tobit regression results of parameter estimates of the % late-stage cases for breast cancer, using raw principal component scores from composite HCAI (model 3).

Based on the parameter estimates in table 2, it is clear that a high principal component score is associated with lower health care access. For this reason, it's not surprising to see a positive and significant parameter estimate of the %late-stage breast cancer cases on $Prin1$. However, $Prin2$ is negatively associated with % late-stage breast cancer cases. Percent Hispanic appears to have a negative impact on late-stage cancer detection. As shown in Figure 3, Hispanics, as a group, actually have a higher percentage of late-stage cancer cases relative to non-Hispanics. However, once we control for obesity and access to health care, we see that a 10% increase in the % Hispanics in a population is associated with a 0.39% decrease in late-stage breast cancer cases. As shown in figure 3, the percentage of uninsured (which is included in HCAI) is much higher in Hispanic populations than non-Hispanic populations. This speaks to the high degree of correlation between HCAI and % Hispanic variables. Obesity appears to have an insignificant impact, although it is worth noting that obesity is highly correlated with % Hispanic as well as the HCAI.

Table 7 shows the results for late-stage colon cancer incidents, where the use of deciles based on model 4 was selected based on its low AIC and SIC. Paradoxically, counties with the worst degree of access to health care (rank9) have a lower percentage of late-stage cancer cases. After detection of adenomatous polyps (not carcinoma yet) by screening, polypectomy is generally performed before the polyps evolve to early-stage cancer (Philips et al., 2011). Since counties with better health care accessibility tend to have higher CRC screening rate, many would-be early stage CRC cases are eliminated by polypectomy, resulting in a reduction in early-stage CRC cases. This may explain the paradoxical negative association between the percentage of late-stage CRC cases and HCAI. In order for the HCAI to be more useful in finding areas that are problematic in late-stage colon cancer determination, the data necessary needs to include the finding of colon cancer precursors, which is not present in our data.

Variables	Estimate	SE	p-value
Intercept	2.592	0.710	0.001
log(% Hisp)	0.026	0.020	0.182
log(% Obese)	0.390	0.220	0.076
rank1	0.106	0.041	0.009
rank2	0.119	0.043	0.006
rank3	0.104	0.038	0.006
rank4	0.021	0.039	0.582
rank5	0.079	0.051	0.120
rank6	0.076	0.047	0.106
rank7	0.053	0.032	0.092
rank8	0.121	0.027	<.0001
rank9	0.062	0.047	0.189
Sigma	0.102	0.005	<.0001

Table 7. Weighted Tobit regression results of % late-stage colon cancer cases, composite HCAI grouping based on 10 groups (model 4).

The parameter estimate results for table 8 are shown below for prostate cancer. Almost all of the included parameter estimates are insignificant, which is consistent with previous results likely for the same reason (Philips et al., 2011).

Variables	Estimate	SE	p-value
Intercept	2.395	1.420	0.092
log(% Hisp)	0.014	0.041	0.731
log(% Obese)	0.077	0.443	0.862
rank1	-0.009	0.080	0.916
rank2	-0.005	0.089	0.964
rank3	-0.046	0.074	0.538
rank4	0.054	0.084	0.522
rank5	0.095	0.107	0.374
rank6	0.044	0.064	0.495
rank7	0.139	0.062	0.024
rank8	-0.056	0.058	0.338
Sigma	0.214	0.010	<.0001

Table 8. Weighted Tobit regression results of % late-stage cases for prostate cancer, composite HCAI grouping based on 9 groups (model 4).

Table 9 shows the results for late-stage lung cancer detection. Based on figure 2, the relationship between lung cancer and HCAI has the strongest correlation. Based on this, it is not surprising to see so many significant variables in the results. For example, groups 1, 2, and 6 are statistically lower than the reference group 8 (omitted). Relative to the other groups, the omitted group (8) has the worst access to health care. It is worth mentioning that as the ranking group increases, the estimate decreases in general. Given that the HCAI also captures socioeconomic factors, it is also likely that counties with high HCAI have higher smoking rates than other counties given the relationship between income and the prevalence of smoking (Laaksonen et al., 2005).

Variables	Estimate	SE	p-value
Intercept	4.469	0.275	<.0001
log(% Hisp)	0.002	0.008	0.894
log(% Obese)	-0.018	0.085	0.836
rank1	-0.073	0.016	<.0001
rank2	-0.019	0.017	0.266
rank3	-0.048	0.015	0.001
rank4	-0.025	0.016	0.115
rank5	-0.020	0.017	0.226
rank6	-0.040	0.011	0.001
rank7	0.004	0.011	0.777
Sigma	0.042	0.002	<.0001

Table 9. Weighted Tobit regression results of estimate of % late-stage cases for lung cancer, composite HCAI grouping based on 8 groups (model 5).

Finally, table 10 shows the parameter estimates associated with late-stage detection for FGS cancer. While the optimal model does not contain any information for the HCAI, the results show significant relationships with the percentage of Hispanics (a negative association) and obese (a positive association) individuals. The latter finding is consistent with previous report that obesity is a risk factor for endometrial cancer (a type of FGS cancer), while the former again suggests that the higher percentage of late-stage FGS cancer in Hispanics (Fig. 3) is due to their lack of access to health care and or socioeconomic deprivation.

Variables	Estimate	SE	p-value
Intercept	0.996	0.910	0.274
log(% Hisp)	-0.080	0.031	0.010
log(% Obese)	0.955	0.296	0.001
Sigma	0.177	0.008	<.0001

Table 10. Weighted Tobit regression results for female genital system (FGS) cancer with no HCAI grouping (models 4 and 5).

5. Final discussion

These results demonstrated in the previous section suggest that data-driven methods such as AIC and SIC as well as the use of raw principal component scores show significant improvement in terms of model-fit when explaining the percentage of late stage cancer cases. Additionally, the newly developed HCAI improves upon the index derived by A&R when evaluating late stage cancer detection. These findings add insight into future studies that will hopefully more effectively utilize PCA in different applications.

The use of data-driven techniques clearly has the ability to provide more flexibility in using PCA in a variety of applications without using rigid rules for grouping ordered indices. While some indices may be appropriately developed, the usage of this information in a regression or grouping context still desires some advances. Using this methodology, we found that the HCAI is positively correlated with the percentage of late-stage cases for breast and lung cancers. This research identifies that in order to promote early cancer detection policy makers should increase the rate of health access as defined in the HCAI. Additionally, the relative lack of explanatory power for the other cancers in this study points to some necessary areas for future research. First, in the case of colon cancer, more detailed data are needed. In order to properly assess colon cancer, data are needed that take account for pre-cancer detection of polyps and adenomas. Without that information, late- and early-stage cancer cannot be properly identified because early detection in the case of colon cancer is detection before it is cancerous. For prostate and FGS cancers, the index was not very explanatory which indicates a lack of health inequality in these areas that can be traced back to health access. Without a doubt, each cancer type is different and in this research we have identified a few types of cancers (breast and lung primarily) where health inequalities exist. These cancers tend to have relatively high costs associated with detection.

The issue of creating an index of health accessibility to explain disparities in health outcomes is an area of much importance in developing and developed counties. The present study shows that health care accessibility as measured by the HCAI impacts delayed detection of several cancers consistent with the results based on data in 2000 (Philips et al., 2011). The positive correlations between the two variables are statistically significant or marginal for all these cancers studied after controlling for several other potential determinants. This finding suggests that socioeconomic deprivation, health insurance coverage and health care service significantly impacts delayed detection of these cancers independent of percentage of Hispanics or percentage of obese individuals in counties of Texas which is one the states with most severe physician shortage and lowest health insurance coverage. Thus, we not only have for the first time proposed the HCAI but also for the first time validated its utility in its correlation with delayed cancer detection.

Previous study showed that WI was significantly associated with delayed detection (assessed by the ratio of late- to early-stage cancer cases similar to the percentage of late-stage cases) of FGS and lung-bronchial cancers but not breast cancer (Philips et al., 2011) in contrast to the results of the current study. The difference in the results is due to difference in study design with which several covariates are entered in the regression in the current study. Particularly, the HCAI covers not only socioeconomic variables but health insurance and health service (physician supply) as well.

It is of interest to note that Hispanics have higher percentage of late-stage cancer cases and also have higher percentages of obese individuals, higher percentages of uninsured individuals. In our multiple regression analysis, we find that delayed cancer detection is actually negatively associated with Hispanic after adjusting for socioeconomic status, physician supply, and percentage of uninsured, which are all included in the HCAI, and percentage of obese individuals. This suggests that their delay in cancer diagnosis is likely due to these factors rather than Hispanic culture per se or genetic predisposition.

These findings provide the evidence-base critical for decision makers to establish policies to promote early detection for effective cancer control targeting specific berries such as physician shortage, lack of health insurance and to improve socioeconomic conditions in general.

One promising avenue that was not undertaken in this study is that of latent class models in grouping. Latent class models provide a couple of advantages over the methods used here: (1) class identification is treated as an unknown variable, which is different from the mutually exclusive technique use in most studies; and (2) classes need not be similar sizes. One major limitation of finite mixture models is that they have been shown to be highly non-linear when more than a few groups are used. In our particular application, this would be a major obstacle given the high number of classes suggested by some models.

6. Acknowledgment

The authors wish to acknowledge the assistance from Kris Hargrave. Cancer incidence data have been provided by the Texas Cancer Registry, Cancer Epidemiology and Surveillance Branch, Texas Department of State Health Services, 1100 W. 49th Street, Austin, Texas, 78756, http://www.dshs. state.tx.us/tcr/default.shtm, or (512) 458-7523.

7. References

Abseyasekera, S. (2005). Multivariate Models for Index Construction, In: *Household Surveys in Developing and Transition Counties: Design, Implementation, and Analysis*, pp. 367-388, United Nations, 92-1-161481-3, New York, NY, USA.

Albrecht, J. and Ramasbramanian, L. (2004). The Moving Target: A Geographic Index of Relative Wellbeing, *Journal of Medical Systems,* Vol.28, No.4, pp. 371-384, 1573-589X

Amemiya, T. (1973). Regression Analysis when the Dependent Variable is Truncated Normal, *Econometrica,* Vol.41, No.6, pp. 997-1016, 1468-0262

Amemiya, T. (1984). Tobit Models: A Survey, *Journal of Econometrics,* Vol.24, No.1-2, pp. 3-61, 0304-4076

Akaike, H. (1974). A New Look at the Statistical Identification Model, *IEEEE Transactions on Automatic Control,* Vol.19, No.6, pp. 716-723, 0018-9286

Berenger, V. and Verdier-Chouchane, A. (2007). Multidimensional Measures of Well-Being: Standard of Living and Quality of Life Across Counties, *World Development*, Vol, 35, No. 7, 1259-1276.

Booysen, F., Van Der Berg, S., Burger, R., Von Maltitz, M. (2008). Using an Asset Index to Assess Trends in Poverty in Seven Sub-Saharan African Countries, *World Development*, Vol. 36, No. 6, 1113-1130.

Byers T. Two decades of declining cancer mortality: progress with disparity. Annu Rev Public Health 2010, 31:121-32.

Centers for Disease Control and Prevention (CDC). Ten great public health achievements--worldwide, 2001-2010. MMWR Morb Mortal Wkly Rep. 2011;60:814-818

Clegg LX, Reichman ME, Miller BA, Hankey BF, Singh GK, Lin YD, Goodman MT, Lynch CF, Schwartz SM, Chen VW, Bernstein L, Gomez SL, Graff JJ, Lin CC, Johnson NJ, Edwards BK. Impact of socioeconomic status on cancer incidence and stage at diagnosis: selected findings from the surveillance, epidemiology, and end results: National Longitudinal Mortality Study. Cancer Causes Control. 2009;20:417-435.

Cortinovis, I., Vella, V., and Nkiku, J. (1993). Construction of a Socio-Economic Index to Facilitate Analysis of Health Data in Developing Countries, *Social Science & Medicine*, Vol, 36, No. 8, 1087-1097, 0277-9536.

Crampton P, Salmond C, Sutton F. NZDep91: a new index of deprivation. Soc Policy J New Zealand. 1997;9:186-193

Davis, B. (2003). *Choosing a Method for Poverty Mapping, Food and Agriculture Organization of the United Nations*, 92-5-104920-3, Rome, Italy. Accessed at http://www.fao.org/DOCREP/005/y4597e/y4597e00.HT on Oct. 7, 2011.

Deaton, A. (2003). Health, Inequality, and Economic Develoment. *Journal of Economic Literature*, Vol 41, 113-158,

Everitt, B. and Hothorn, T. (2011). *An Introduction to Applied Multivariate Analysis with R*, Springer Science, 978-1-4419-9649-7, New York, NY, USA.

Everitt, B., Landau, S., Leese, M. and Stahl, D. (2011). *Cluster Analysis* (5th ed.), John Wiley & Sons, 978-0470220436, Chichester, UK.

Ewing, R., Schmid, T., Killingsworth, R., Zlot, A., Raudenbush, S. (2003). Relationship Between Urban Sprawl and Physical Activity, Obesity, and Morbidity. *American Journal of Health Promotion*, Vol. 18, No. 1, 47-57.

Filmer, D. & Pritchett, L.H. (2001). Estimating Wealth Effects without Expenditure Data – or Tears: An Application to Educational Enrollments in States of India, *Demography*, Vol, 38, No. 1, 115-132.

Greene, W. (2012). *Econometric Analysis* (7th edition), Pearson Education, Inc., 0-13-139538-6, Upper Saddle River, NJ, USA.

Gwatkin, D.R., Rutstein, S., Johnson, K., Pande, R., and Wagstaff, A. (2000). Socio-Economic Differences in Health, Nutrition, and Population. HNP/Poverty Thematic Group, World Bank, Washington, DC.

Koehler, A. & Murphree, E. (1988). A Comparison of the Akaike and Schwarz Criteria for Selecting Model Order, *Journal of the royal Statistical Society. Series C (Applied Statistics),* Vol.37, No.2, pp. 187-195, 1467-9876

Kolenikov, S. and Angeles, G. (2009). Socioeconomic Status Measurement with Discrete Proxy Variables: Is Principal Component Analysis a Reliable Answer? *Review of Income and Wealth*, Vol. 55, No. 1, 128-168.

Kula, J. (2004). AIC and BIC: Comparisons of Assumptions and Performance, *Sociological Methods & Research,* Vol.33, No.2, pp. 188-229, 1552-8294.

Laaksonen, M., Rahkonen, O., Karvonen, S. & Lahelma, E. (2005). Socioeconomic Status and Smoking: Analysing Inequalities with Multiple Indicators, *European Journal of Public Health,* Vol.15, No.3, pp. 262-269, 1464-360X.

National Cancer Institute . Surveillance Epidemiology and End Results. Standard populations - 19 age groups.
http://seer.cancer.gov/stdpopulations/stdpop.19ages.html (accessed 09 Mar 2011

Philips BU, Jr., Gong G, Hargrave KA, et al. Correlation of the Ratio of Late-stage to Non-Late-stage Cancer Cases with the Degree of Socioeconomic Deprivation among Texas Counties. International Journal of Health Geographics 2011;10: 12.

Pomeroy, R.S., Pollnac, R.B., Katon, B.M., & Predo, C.D. (1997). Evaluating Factors Contributing to the Success of Community-Based Coastal Resource Management: The Central Visayas Regional Project-1, Philippines. *Ocean and Coastal Management*, Vol. 36, No. 1-3, 97-120.

Richardson LC, Royalty J, Howe W, et al. Timeliness of breast cancer diagnosis and initiation of treatment in the National Breast and Cervical Cancer Early Detection Program, 1996-2005. Am J Public Health 2010, 100:1769-1776.

Ruel, M.T. and Menon, P. (2002). Child Feeding Practices Are Associated with Child Nutritional Status in Latin America: Innovative Uses of the Demographic and Health Surveys. *Journal of Nutrition*, Vol. 132, No. 6, 1180-1187, 1541-6100.

Salmond C, Crampton P, Sutton F. NZDep91: A New Zealand index of deprivation. Aust N Z J Public Health. 1998;22:835-837

Schwartz, G. (1978). Estimating the Dimension of a Model, *Annals of Statistics,* Vol.6, No.2, pp. 461-464, 0090-5364

Singh GK. Area deprivation and widening inequalities in US mortality, 1969-1998. Am J Public Health. 2003;93:1137-1143.

Singh GK, Siahpush M. Increasing inequalities in all-cause and cardiovascular mortality among US adults aged 25-64 years by area socioeconomic status, 1969-1998. Int J Epidemiol. 2002;31:600-613.

Texas State Data Center. Published Reports. http://txsdc.utsa.edu/Reports. accessed 13 Oct. 2011.

Texas Department of State Health Services . County supply and distribution Tables. http://www.dshs.state.tx.us/chs/hprc/PHYS-lnk.shtm (accessed 17 Mar 2011).

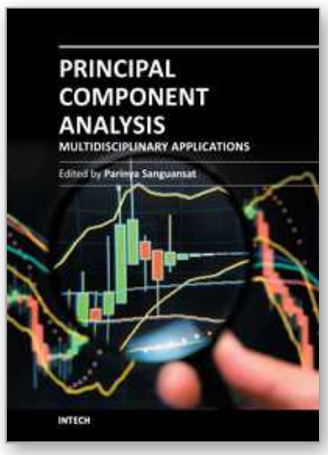
Tobin, J. (1958). Estimation of Relastionships for Limited Dependent Variables, *Econometrica,* Vol.26, No.1, pp. 24-36, 1468-0262

U.S. Census Bureau. 2005-2009 American Community Survey 5-Year Estimates.
http://factfinder.census.gov/servlet/DCGeoSelectServlet?ds_name=ACS_2009_5YR_G00_

Vyas, S. and Kumaranayake, L. (2006). Constructing Socio-Economic Status Indices: How to Use Principal Component Analysis. *Health Policy and Planning*, Vol. 21, No. 6, 459-468, 1460-2237.

Ward E, Jemal A, Cokkinides V, Singh GK, Cardinez C, Ghafoor A, Thun M. Cancer disparities by race/ethnicity and socioeconomic status. CA Cancer J Clin. 2004 Mar-Apr;54(2):78-93

World Health Organization. Commission on Social Determinants of Health. Closing the gap in a generation: Health equity through action on the social determinants of health. http://www.who.int/social_determinants/thecommission/finalreport/en/index.html (accessed 25 August 2011)



Principal Component Analysis - Multidisciplinary Applications
Edited by Dr. Parinya Sanguansat

ISBN 978-953-51-0129-1
Hard cover, 212 pages
Publisher InTech
Published online 29, February, 2012
Published in print edition February, 2012

This book is aimed at raising awareness of researchers, scientists and engineers on the benefits of Principal Component Analysis (PCA) in data analysis. In this book, the reader will find the applications of PCA in fields such as taxonomy, biology, pharmacy,finance, agriculture, ecology, health and architecture.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen