EPIDEMIOLOGY OF DIABETES AND RELATED MORTALITY:
EARLY SCREENING, SOCIOECOLOGICAL DETERMINANTS, AND THE VALUE OF PREVENTION

BY

KEDIR NESHA TURI

Dissertation

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Community Health in the Graduate College of the University of Illinois at Urbana-Champaign, 2015

Urbana, Illinois

Doctoral Committee

Associate Professor Diana Grigsby-Toussaint, Chair, Director of Research
Professor Reginald Alston
Associate Professor Mary Arends-Kuenning
Professor David Buchner
Professor Sara McLafferty
Abstract

The focus of my dissertation is prediction, prevention, and economic valuation of type 2 diabetes. I studied individual level type 2 diabetes risk factors, spatial spillover effect of diabetes-related mortality ecological risk factors, and individuals’ loss of well-being due to diabetes. In the first essay, titled “Diabetes Risk Prediction: Multivariate Nonlinear Interaction Approach,” I argue that the success in preventing or delaying the incidence of type 2 diabetes and subsequent complications depend on the early detection of undiagnosed cases and identifying people at high-risk. However, early detection of type 2 diabetes is seldom feasible because the symptoms show up late, and screening the entire population is very costly. Individuals who are prone (e.g., due to family history) to developing type 2 diabetes and those with undiagnosed diabetes need to be targeted for early screening. Thus, it is imperative to continue designing assessment mechanisms that help to identify individuals at high-risk based on simple, non-invasive, inexpensive, and routine clinical measurements. In this paper, I build a model that helps to predict type 2 diabetes with readily available, inexpensive, non-invasive, and easy-to-collect information. National Health and Nutrition Examination Survey (NHANES) data is analyzed to build this risk model. A non-parametric regression method, Multivariate Adaptive Regression Splines (MARS), is used to allow for interactions and non-linearity in the model. A risk prediction model using the MARS approach achieved a performance level of 87% accuracy with area under receiver operating character curve (AUROC) of 0.86, which is higher than similar models based on invasive and non-invasive measurements. Moreover, this model requires few measurements and limited information that may be obtained in settings such as community-based chronic disease prevention programs and workplace well-being programs. Therefore, this risk prediction model can be translated into a usable risk-scoring tool in community chronic prevention and employee wellness programs.

The second essay, titled “Spatial Spillover Effect from Socio-Ecological Determinants of Diabetes-Related Mortality in the US,” explores the spatial spillover effect from socio-ecological risk factors that are associated with type 2 diabetes-related mortality. I studied the spatial spillover effect of change in socioeconomic gradients (education, employment, and household income), retail food environments, and access to health-care on diabetes-related mortality rates (DRMR) across the United States. To examine mortality clusters and factors associated with the
clusters and spatial spillover effect, seven-year aggregates of multiple-cause mortality data from CDC WONDER compressed mortality database was merged with several sources of county-level data. The results show that high DRMR cluster counties are located throughout the Southern Plains, Southeastern, and Appalachian regions. High DRMR clusters are characterized by lower socioeconomic status, high density of fast food restaurants, lack of access to grocery stores, high proportion of African Americans, and low physical activity. Moreover, the impacts from change in socioeconomic gradients and the retail food environment in a particular county spill over to neighboring counties. The implication is that improvement in socioeconomic status and access to healthy food would significantly reduce DRMR in contiguous US counties.

The third essay, titled “What is the Value in Diabetes Prevention? A Subjective Well-Being Valuation Approach,” uses loss of well-being due to diabetes to quantify the monetary value of diabetes prevention in the US population. In this paper, I argue that the current preference-based health valuation approach is not appropriate for prevention-based programs valuation because they do not capture the social and economic value that an individual puts on a health condition. I utilize a recently developed subjective well-being valuation approach to quantify the monetary value of loss in well-being due to diabetes in the US population. This approach assumes that individuals derive overall life satisfaction from well-being, which is a function of health and income. Health, in turn, is produced by the combined input of an individual’s behaviors and medical technology. Thus, a marginal trade-off between health and income is used to derive the monetary value of health. The Panel Study of Income Dynamics (PSID) data was utilized for this study. The result shows that the monetary value for diabetes prevention is about $37,000, which is less than the current implicit threshold for program implementation. The resulting monetary value will help to quantify the societal value of diabetes prevention, which can be used to estimate the benefit side of the cost-benefit analysis.
Dedicated to my children
Acknowledgments

I would like to express my deepest gratitude to my advisor, Dr. Diana Grigsby-Toussaint, for her excellent guidance, caring, patience, and providing me with an excellent working atmosphere and financial support. Without her guidance and persistent help, this dissertation would not have been possible. I would also like to thank my dissertation committee members, Dr. Alston Reginald, Dr. Mary Arends-Kuenning, Dr. David Buchner, and Dr. Sara McLafferty for guiding my research. I am grateful to Dr. Hillary Klonoff-Cohen for her comments and suggestions on my dissertation proposal. I am indebted to my CompGen fellowship advisor Dr. Ruby Mendenhall for her guidance to expand the horizon of my research experience. I am also grateful to the CompGen Fellowship program for financial support through a pre-doctoral graduate fellowship.

I am thankful to my parents, who supported and encouraged me in all my endeavors. Thanks to my four sisters, brother, Uncle Desta, and other family members for their support and encouragement. I am grateful to my wife, Mulu Nuru, who was always there to cheer me up and stood by me during a sometimes-trying process of completing my Ph.D. degree. To our daughter Sitina, you are my inspiration. Finally, my appreciation goes to fellow graduate students and friends, specifically Mary Christoph and Dr. Beza Ayalew, for their support, feedback, and friendship.
Table of Contents
Chapter 1: Background and Significance ................................................................. 1
  1.1 Background ......................................................................................................... 1
  1.2 Contribution of the dissertation ....................................................................... 5
  1.3 References ......................................................................................................... 8
  2.1 Introduction ........................................................................................................ 11
  2.2 Methods ............................................................................................................. 28
  2.3 Results ............................................................................................................... 39
  2.4 Discussions ....................................................................................................... 42
  2.5 Conclusions ...................................................................................................... 45
  2.6 References ....................................................................................................... 47
Chapter 3: Spatial Spillover of Socio-Ecological Determinants of Diabetes-related Mortality Rate across US Counties ................................................................. 60
  3.1 Introduction ....................................................................................................... 61
  3.2 Methods ............................................................................................................. 69
  3.3 Results ............................................................................................................... 84
  3.4 Discussions ....................................................................................................... 96
  3.5 Conclusions ..................................................................................................... 100
  3.6 References ....................................................................................................... 101
Chapter 4: What is the Value in Diabetes Prevention? Well-being Valuation Approach ........ 108
  4.1 Introduction ....................................................................................................... 109
  4.2 Methods ............................................................................................................. 119
  4.3 Results ............................................................................................................... 127
  4.4 Discussions ....................................................................................................... 130
  4.5 Conclusions ..................................................................................................... 133
  4.6 References ....................................................................................................... 135
Tables and Figures .................................................................................................. 140
Appendix A: Summary of literature rated to chapter 2 .................................................. 166
Appendix B: Summary of data source and literature related to chapter 3 ....................... 168
Appendix C: Summary of data and literature related to chapter 4 .................................. 172
Chapter 1: Background and Significance

1.1 Background
Diabetes Prevalence

Diabetes is an increasingly challenging chronic disease in the world. Estimates from the International Diabetes Federation (IDF) indicate 382 million people are currently living with diabetes, with individuals between the ages of 40 and 59 being the most affected (IDF, 2014). The loss of well-being and economic burden of diabetes is enormous. Worldwide deaths related to diabetes were estimated to be 5.1 million, and financial costs of medical treatments for illnesses related to diabetes was estimated to be 548 billion US dollars in 2013 (IDF, 2014). With projected increases in diabetes prevalence by as much as 55% by 2035 (IDF, 2014), these costs are expected to continue to rise. In the United States, the prevalence rate of diabetes stands at 9.3% of the population, which is 21 million diagnosed and 8.1 million undiagnosed cases (CDC, 2014). Approximately 2 million new cases of diabetes are diagnosed every year in the United States (CDC, 2013). An additional 2 million individuals develop diabetes every year, but remain undiagnosed (CDC, 2013). The health consequences of diabetes are severe. Diabetes is associated with increased risk of cardiovascular disease and premature mortality and is the leading cause of blindness, kidney failure, and non-traumatic amputations resulting from microvascular complications (National Diabetes Statistics, 2011).

Risk factors for type 2 diabetes

The majority (approximately 97%) of individuals with diabetes have type 2 diabetes (Adeghate, Schattner, & Dunn, 2006). The risk factors for type 2 diabetes include family history of diabetes, overweight, unhealthy diet, physical inactivity, increasing age, high blood pressure, ethnicity, history of gestational diabetes, and poor nutrition during pregnancy. The underlying individual level risk factors can be narrowed to genetic, lifestyle, and environmental factors. Genetic susceptibility plays a crucial role in the etiology and manifestation of type 2 diabetes, with concordance in monozygotic twins approaching 100% (Kumar & Clark, 1999). Genetic susceptibility could be modified by environmental factors for an individual to have type 2
diabetes. The variability observed in diabetes prevalence among ethnic groups could be a combination of both genetic (thrifty gene theory) and environmental factors (Adeghate, Schattner, & Dunn, 2006). Several studies have reported that regular physical activity increases insulin sensitivity and glucose tolerance (van Dijk, 2012). Moreover, it has been recently shown that physical activity reduces the risk of type 2 diabetes (Colberg et al., 2010). In addition, excessive caloric intake and diet quality are major driving forces behind escalating obesity and type 2 diabetes epidemics (Hu, 2011).

Disparity in diabetes-related mortality

Adults diagnosed with diabetes have a greater than twofold increased risk of mortality compared to the general population (Egede, Nietert, & Zheng, 2005). In the United States, diabetes is the seventh leading cause of death (Murphy, Xu, & Kochanek, 2013). However, the trend in diabetes-related mortality is declining. Gregg et al. (2012) found that among diabetic adults, the CVD death rate declined by 40% and all-cause mortality declined by 23% between the earliest (1997) and latest (2006) samples of National Health Interview Survey (NHIS) linked to the National death index. Though the mortality rate from diabetes is declining due to medical advances (Saaddine et al., 2006; Murphy, Xu, & Kochanek, 2013), this benefit has not trickled down equally among all racial groups (Murphy, Xu, & Kochanek, 2013) and geographical areas (Murray, 2006). For example, racial ethnic disparities are still evident, with a black to white diabetes mortality ratio of 2:1 in 2010 (Murphy, Xu, & Kochanek, 2013).

Studies show that disparity in diabetes-related mortality rate is associated with genetic, socioeconomic, and environmental factors (Saydah & Lochner, 2010; Dray-Spira, 2013). In a retrospective cohort study, Lipscombe et al. (2010) compared changes in mortality from 1994 to 2005 by neighborhood income strata among people with diabetes aged 30 years. Lipscombe et al. found that the mortality rate was declining but the decline was significantly greater in the highest income group (by 36%) than in the lowest income group (by 31%). In a study where mortality was linked to the National Health Interview Survey (NHIS), Saydah, Imperatore, & Beckles (2013) found that level of education and financial wealth significantly contributed to diabetes-
related mortality, even after adjusting for demographics, healthcare access, and psychological distress.

**The economic costs of diabetes**

Diabetes is one of the most costly chronic diseases for both individuals and society. The cost of diabetes can be direct or indirect. Direct financial cost of diabetes to individuals and their families include medical care, drugs, insulin, and other supplies. Patients may also have to bear other personal costs, such as increased payments for health, life, and automobile insurance (WHO, 2015). The direct financial cost to the society is the diversion of scarce resources of the healthcare sector, including hospital services, physician services, lab tests to diabetes treatment, and daily management. The daily management of diabetes includes the supply of products such as insulin, syringes, oral hypoglycemic agents, and blood-testing equipment (WHO, 2015). Costs range from relatively low-cost items, such as primary care consultations and hospital outpatient episodes, to very high-cost items, such as long hospital inpatient stays for the treatment of complications (WHO, 2015). Diabetes also increases the cost of treating general conditions such as depression and cancer that are not directly related to diabetes (Egede, Zheng, & Simpson, 2002; Vigneri et al., 2009; Osborn & Holt, 2012). Therefore, a portion of health-care expenditures for these medical conditions is attributed to diabetes.

In addition, diabetes incurs indirect tangible costs both to patients and their families as well as to society. The indirect costs associated with diabetes include absenteeism from work due to illness, reduced work productivity while working due to health conditions, reduced workforce participation due to disability, and productivity lost due to premature mortality (Fu et al., 2009; Lee et al., 2008). Productivity loss occurs among those in the labor force as well as among the unemployed population. The indirect cost to patients and their families are in the form of lost income due to absenteeism and lost workplace promotion opportunities due to lack of productivity. The indirect cost does not include the intangible cost to the individual including loss of well-being and happiness because of diabetes-related illnesses.
Diabetes diagnosis and screening

The common diabetes diagnosis and screening methods are based on plasma glucose criteria, either the fasting plasma glucose (FPG) or the 2-h plasma glucose (2-h PG) value after a 75-g oral glucose tolerance test (OGTT) (ADA, 2015). Fasting is defined as no caloric intake for at least eight hours and the threshold for diagnosis is FPG ≥126 mg/dL (7.0 mmol/L) (ADA, 2015). In 2009, an International Expert Committee decided to add the A1C as a third option to diagnose diabetes (International Expert Committee, 2009). The diagnosis of diabetes with A1C is at a threshold of 6.5%. As with FPG and 2-h PG, the cutoff point for an A1C base diabetes diagnosis is associated with the risk of retinopathy. The advantage of using A1C as a diagnostic method as compared to the FPG and OGTT includes greater convenience (fasting not required), possibility of greater pre-analytical stability, and fewer day-to-day perturbations during stress and illness (International Expert Committee, 2009; ADA, 2015). However, the limited availability of A1C testing in certain regions of the developing world and the incomplete correlation between A1C and average glucose in certain individuals are some of its limitations (ADA, 2015).

The benefit of diabetes screening and early diagnosis to individuals include enhanced length and/or quality of life, which might result from a reduction in the severity and frequency of the immediate effects of diabetes or the prevention or delay of its long-term complications (WHO, 2003). The institutional benefits of diabetes screening are saving or redistribution of health-care resources, which might be possible as a result of reduced levels of care required for diabetes complications (reduced hospital admissions and lengths of stay, etc.) (WHO, 2003). The same methods discussed above are also used for screening asymptomatic individuals; the difference is that the test has to be repeated in case of diagnosis.

Diabetes prevention

The human and economic cost of diabetes can only be reduced by reversing the epidemic of diabetes and related complications. Studies have shown that lifestyle modification-based interventions such as diet and physical activity can prevent, delay, or reverse the incidence of diabetes and subsequent complications (Knowler et al., 2002; Mensink, 2005; Lindstrom et al.,
To a lesser extent, pharmaceutical intervention can also reduce the incidence of diabetes (Chiasson et al., 2002; DREAM, 2006); and a combination of both lifestyle modification and pharmaceutical intervention has the greatest effect (DPP, 2009). Cost effectiveness studies for such programs (DPP, 2003; Eddy et al., 2005; Herman et al., 2005; Hoeger et al., 2007) show that preventing diabetes with lifestyle modification or in combination with pharmaceuticals is cost effective.

Following the Affordable Care Act (ACA), efforts are underway to establish chronic disease prevention programs throughout the United States. ACA expansion of insurance coverage, new benefits for clinical preventive services, and funding for complementary public health and primary care programs is most likely to reduce the socioeconomic barriers to preventative and diagnostic services (Mayes & Oliver, 2012). In those states that have expanded Medicaid, the effect has been realized through the increased number of newly diagnosed diabetes cases (Kaufman et al., 2015). The new funding for complementary public health and primary care programs allows the programs to be implemented through primary care, community-based prevention, and workplace wellness programs. Lifestyle-based interventions such as healthy diet, physical activity, non-use of tobacco, and weight management are at the core of these programs. The sustained implementation of these programs coupled with the culturally appropriate public health campaign is predicted to reduce the epidemic of chronic diseases including diabetes.

1.2 Contribution of the dissertation
In general, my dissertation contributes to the primary and secondary diabetes prevention literature. The contribution of my dissertation is to identify individual and ecological risk factors and generate useful knowledge and tools for the primary and secondary prevention of diabetes in the US population. The knowledge generated is useful to design diabetes intervention programs at various levels (individual, ecological, and policy). It also contributes to the literature of health economics and resource allocation for chronic disease prevention by modeling the monetary value of diabetes prevention to the US population.

As discussed above, formal screening is expensive, inconvenient, and invasive. As a result, screening for diabetes has not been widely implemented despite mounting literature
showing early intervention could reverse or delay the progression of diabetes or related complications. Several questionnaire- or medical data-based screening methods have been developed to alleviate this problem. However, the underlying models either require information based on invasive procedures or their performance is low. In Chapter 2 of this dissertation, I identify individual level risk factors to build a type 2 diabetes prediction model that can be used for early screening before the disease causes damage to the micro- and macro-vascular and nerve systems. The prediction model was developed keeping in mind that primary prevention programs are implemented in limited resource settings such as community and workplace wellness programs. It also aims to reduce the inconvenience and cost of screening and eliminate the use of invasive procedures such as the need for fasting, fear of needles, wait time for laboratory test results, and the cost of specialized personnel, equipment, and reagents. The model uses simple anthropometric and blood pressure measures and questionnaire-based information to predict diabetes. It exploits the piecewise interaction between risk factors to improve the performance of the model in diabetes prediction.

Most previous studies have not examined the geographical distribution of diabetes-related mortality rate and correlated socio-ecological risk factors. The suspected underlying risk factors for diabetes-related mortality rate are socioeconomic and social capital conditions, access to health services, and access to food and food retail environments. The aim of Chapter 3 of this dissertation is to examine the geographical distribution of diabetes-related mortality rates and its relationship to these risk factors. The study identified US counties where diabetes-related mortality rates are high and explored the socio-ecological characteristics associated with these counties. The socio-ecological characteristics associated with management of diabetes included in this study are socioeconomic and social capital indicators, healthcare service indicators, demographic composition, and food retail environments and access. The study also identified factors that would have a multiplier effect on the neighboring counties if an intervention were implemented in one county. As the prevention and management of diabetes depends on the socio-ecological environments surrounding individuals, this contribution is relevant for designing diabetes-related mortality prevention interventions addressing socioeconomic and physical environments.
Currently, there is no method particularly developed to evaluate the benefit of chronic disease prevention. Comparative economic benefit analysis such as cost effectiveness and cost utility analysis are better fit to evaluate the progress of specific medical technology. Decision utility (preference) based methods, such as stated and revealed preference methods, do not account for the firsthand experience of individuals’ loss of well-being from the onset of diseases. In Chapter 4, I evaluate the monetary value of diabetes prevention from the perspective of individuals who have firsthand experience with diabetes. The study uses a well-being valuation approach developed in happiness and environmental studies to evaluate the value of diabetes prevention from the perspective of individuals’ who lost their well-being due to diabetes. The study is relevant to the discourse of resource allocation to chronic disease prevention, including diabetes.
1.3 References


Fu AZ, Qiu Y, Radican L & Wells BJ. Health care and productivity costs associated with diabetic patients with macrovascular comorbidity conditions. *Diabetes Care* 2009; 32:2187-2192.


Chapter 2: Diabetes Risk Prediction: Multivariate Nonlinear Interaction Approach

Abstract

The success in preventing or delaying the incidence of type 2 diabetes and subsequent complications depends on the early detection of undiagnosed cases and identifying individuals at high-risk. However, early detection of type 2 diabetes is seldom feasible because the symptoms show up late, and screening the entire population is very costly. Those individuals that are prone to developing type 2 diabetes and those with undiagnosed diabetes need to be targeted for screening. Thus, it is imperative to design assessment mechanisms that help to identify individuals at high-risk based on simple, non-invasive, inexpensive, and routine clinical measurements and questionnaires. In this paper, I build a model that helps to predict type 2 diabetes with readily available, inexpensive, and non-invasive information. National Health and Nutrition Examination Survey (NHANES) data is analyzed to build this risk model. A non-parametric regression method, Multivariate Adaptive Regression Splines (MARS), is used to allow for interactions and non-linearity in the model. A risk prediction model using the MARS approach achieved a performance level of 87% accuracy with area under receiver operating character (AUROC) of 0.86, which is higher than similar models based on invasive and non-invasive measurements. Moreover, the measurements and information needed for this model may be obtained in settings such as community-based chronic disease prevention programs and workplace well-being programs. Therefore, this risk prediction model can be translated into a usable risk-scoring tool in community chronic prevention and employee wellness programs.
2.1 Introduction

Type 2 diabetes (T2D) is the most common form of diabetes, accounting for 90-95% of all diabetes cases (CDC, 2011; WHO, 2015). People with diagnosed diabetes display evidence of progressive insulin resistance; beginning with cells’ inadequate use of insulin and as the need for insulin rises, the pancreas gradually loses its ability to produce insulin (Inzucchi, 2012). The age at which T2D is diagnosed is sliding from 30 years and older (with additional risks such as overweight or obesity and/or having a positive family history) to individuals as young as 15 years of age (Alberti, 2004). Younger populations, especially adolescents in racial and ethnic groups that are at high-risk for T2D, are frequently being diagnosed (Alberti, 2004; Constantino et al., 2013). The hazard of developing T2D at such a young age is more dangerous than type 1 diabetes (T1D) due to lifetime exposure to hyperglycemia and the higher probability of complications early in life (Constantino et al., 2013).

Even with well-known underreporting on death certificates, T2D is the seventh leading cause of death in the United States (CDC, 2011). T2D is associated with the increased risk of premature mortality; people with diabetes have a twofold risk of mortality compared to people without it (CDC, 2011). Moreover, complications from T2D have dreadful consequences for health-related quality of life. More than half of the people with diabetes eventually contract macro- and micro-vascular diseases, the majority of which are cardiovascular diseases (CVD) (Bowden et al., 2010). T2D is an independent risk factor for the development of CVD (Haffner et al., 1998; Greenland et al., 2003), coronary heart disease (CHD), and myocardial infarction (MI) (Greenland et al., 2003). When T2D is coupled with genetic susceptibility and other risk factors (hypertension, microalbuminuria, uncontrolled blood glucose, etc.), eventually it culminates in macrovascular disease (Bowden et al., 2010). Diabetes is also associated with microvascular diseases such as retinopathy, nephropathy, and neuropathy (Forbes & Cooper, 2013). In the United States, diabetes is the leading cause of new cases of blindness and kidney failure (CDC, 2011). In addition, diabetes patients suffer from nerve system diseases (neuropathy) and lower-extremity amputations (Forbes & Cooper, 2013). In the United States, about 60-70% of diabetic patients have mild to severe forms of nervous system damage, which causes impaired sensation or pain in the feet or hands, slowed digestion of food in the stomach, carpal tunnel syndrome, erectile dysfunction, or other nerve problems (CDC, 2011).
However, the incidence of diabetes and subsequent complications can be prevented, delayed, or reversed with lifestyle modification interventions such as diet and physical activity (Knowler et al., 2002; Mensink, 2005; Lindstrom et al., 2006; Li et al., 2008), because the risk factors for T2D are highly associated with the metabolic syndrome arising from obesity and inactive lifestyle. Pharmaceutical interventions (Chiasson et al., 2002; DREAM, 2006) or a combination of lifestyle modification and pharmaceutical interventions (DPP, 2009) could prevent the onset of diabetes complications. Studies show that the success in preventing or delaying the incidence of T2D and subsequent complications depends on the early detection of undiagnosed cases and identifying people at high-risk (Gillies et al., 2008; Simmons et al., 2010).

Despite the abundant research evidence regarding the importance of early screening and recommendations by the American Diabetes Association (ADA), laboratory based planned formal screening is rarely conducted (Ealovega et al., 2004; Inzucchi, 2012; Chatterjee, 2013). Moreover, early detection of the onset of T2D is seldom feasible because the symptoms show up late (Inzucchi, 2012), and screening the entire population is very costly (Waugh et al., 2007). To reduce the costs of screening and intervention programs, those individuals that are prone to developing T2D and those with undiagnosed cases need to be targeted (Engelgau, Narayan, & Herman, 2000). Screening a portion of the population that is prone to developing diabetes and confirming the high-risk individuals with common laboratory tests such as Fasting Plasma Glucose (FPG) and Oral Glucose Tolerance Test (OGTT), however, is economically inefficient (Norris, et al., 2008). Focusing on the early signs would help to target both risk of T2D and cardiovascular disease at the same time (Knowler et al., 2002), because lifestyle modification interventions have more impact at the early stage of metabolic disorders rather than later. Thus, it is necessary to continue to design screening mechanisms that help to identify individuals at high-risk based on simple, non-invasive, inexpensive, and routine clinical measurements and questionnaires.

**Methods for screening and diagnosis of T2D**

The current screening methods for T2D are based on the capillary or venous blood test for glucose. The screening tests are usually conducted after the individual becomes symptomatic. The tests are considered screening tests when an asymptomatic person is tested for identifying
undiagnosed diabetes. A test is considered as part of a diagnostic test if the test is repeated to confirm the result of the same test, or if the person is symptomatic and these tests are used as confirmatory tests. The screening and diagnostic tests currently in use include measurement of fasting plasma glucose (FPG), two-hour plasma glucose during an oral glucose tolerance test (2-h OGTT), and glycated hemoglobin (A1C). Fasting is defined as no caloric intake for at least eight hours before blood draw for the test. The cutoff point for all three tests are delineated based on their correlation with the presence of diabetic retinopathy symptoms (American Diabetes Association [ADA], 2013), which makes it difficult to use them for early screening before the disease causes damage. An individual can be classified as diabetic (T2D) if FPG ≥ 126 mg/dL (7.0 mmol/L), A1C ≥ 6.5%, a two-hour value in an OGTT (2-h PG) at or above 200 mg/dL (11.1 mmol/L), or a random plasma glucose concentration ≥200 mg/dL (11.1 mmol/L) in the presence of symptoms (ADA, 2013). The diagnosis of diabetes must be confirmed on a subsequent day by repeat measurement, repeating the same test for confirmation. In addition, individuals may be categorized as high-risk for diabetes (pre-diabetes) if the blood glucose is not high enough to be diagnosed as diabetes but is too high to be categorized as normal. Pre-diabetes occurs if FPG is 100 mg/dL (5.6 mmol/L) to 125 mg/dL (6.9 mmol/L), A1C is 5.7-6.4%, or 2-h OGTT is 140 mg/dL (7.8 mmol/L) to 200 mg/dL (11.0 mmol/L).

The OGTT and FPG, however, can be cumbersome, time-consuming, and expensive (ADA, 2013). Thus, the use of glycated hemoglobin (A1C) values for screening and identification of impaired glucose tolerance and diabetes was recommended by the International Expert Committee in a consensus report issued in June of 2009 (Gillett, 2009). The International Expert Committee recommendation was to use an A1C level ≥6.5% to diagnose diabetes, and individuals with A1C values between 6.0 and 6.4% were considered to be at particularly high-risk for developing diabetes (Gillett, 2009). Subsequently, the ADA affirmed this decision, but with A1C levels ≥6.5% to diagnose diabetes and A1C levels between 5.7 and 6.4% to represent increased risk for developing diabetes (ADA, 2013). The diagnosis should be confirmed with a repeat A1C measurement (ADA, 2013).
Risk factors for T2D

Non-modifiable risk factors

Age, gender, ethnicity, and family history are non-modifiable risk factors for T2D. T2D incidence and prevalence increase with age. In the US, 26.9% of all people aged 65 and above have diabetes (CDC, 2011). The peak ages for T2D may vary with country and dominant ethnic group in those countries. Studies in Asian countries show that the prevalence of T2D increased with age and peaked at 70–89 years of age in Chinese and Japanese subjects but peaked at 60–69 years of age followed by a decline at 70 years of age in Indian subjects (Mohan, 2003). The higher prevalence at higher age can partly be explained by the decrease in insulin secretion with age. This involves several biological factors including loss of Sirt1-mediated Glucose-Simulated Insulin Secretion (GSIS) due to age, decreased β-cell sensitivity to circulating incretins, decrease in mitochondrial function as age increases, and increased oxidative stress with age (Scheen, 2005; Gong & Muzumdar, 2011; Soriguer et al., 2014).

The risk factors for T2D are also different between men and women (Njolstad et al., 1998; Brunton, 2008). Thorand et al. (2007) found that elevated concentrations of C-reactive protein (CRP) have a stronger association with risk of T2D in women than men after adjustment for age and lifestyle factors including smoking, alcohol intake, and physical activity. Endogenous sex hormones may also be contributing to the difference in prevalence of diabetes between men and women by differentially modulating glycemic status and risk of T2D. A systematic review by Ding et al. (2006) found that high testosterone levels are associated with higher risk of T2D in women but with lower risk in men.

Prevalence and incidence of T2D is different among various ethnic groups (Brunton, 2008). Various ethnic groups have different predisposition in developing T2D under similar environmental exposures (Abate & Chandalia, 2003). Among the US population, Hispanics, African Americans, and Asians appear to have a much higher predisposition to T2D than individuals of Caucasian ancestry (Abate & Chandalia, 2003). The hypothesized mechanism that is thought to govern the relationship between ethnicity and T2D is a genetic predisposition of some ethnic groups to insulin resistance. According to the ‘thrifty gene’ hypothesis, a cluster of genetic deviation or polymorphisms that gave survival advantage to some ethnic groups by helping to cope with famine and hunting and gathering lifestyles, could now predispose them to insulin resistance and diabetes in environments of increased food supply and sedentary lifestyle.
(Carulli et al., 2005). In addition, family history, a proxy for shared genes, behavior, and environment among family members are significantly associated with incidence of T2D (Scheuner et al., 1998; Valdez et al., 2007; van’t Riet et al., 2010; Ning et al., 2013).

Modifiable risk factors

*Overweight and obesity*: WHO defines overweight and obesity as abnormal or excessive fat accumulation that may impair health. Thus, overweight and obesity are health risks because excess adiposity accumulates in the body. Adiposity is a well-established independent predictor of T2D. Between 60% and 90% of cases of T2D appear to be related to obesity (Thévenod, 2008). Several studies have shown that insulin resistance precedes the development of hyperglycemia in subjects who eventually develop T2D (Thévenod, 2008). Adipose tissue releases increased amounts of non-esterified fatty acids, glycerol, hormones, pro-inflammatory cytokines, and other factors that contribute to the development of insulin resistance in obese individuals (Kahn et al., 2006). However, not all insulin-resistant individuals develop T2D. Individuals develop T2D when insulin resistance is accompanied by dysfunction of the beta cells, which leads to insulin secretion and the inability to control blood glucose levels (Kahn et al., 2006).

Recent growing evidence shows that the location, or distribution, of excess body fat may incrementally influence the risk of developing T2D. Body Mass Index (BMI) has been used as the anthropometric measure of overall body fat (adiposity) for several years. Recently, other anthropometric measures of adiposity such as waist circumference (WC), waist-to-hip ratio (WHR), and waist-to-height ratio (WHtR) are being used to account for location and distribution of adiposity. A meta-analysis by Qiao & Nyamdorj (2009) compared BMI, WC, and WHtR in their relation to the incidence and prevalence of T2D. The best predictor among these measures did not find a statistically significant difference. A meta-analysis by Vazquez (2007), however, found that abdominal fat (WC) is a slightly better predictor of T2D than BMI. Another meta-analysis by Lee et al. (2008) found that upper body size (as measured by WHtR) is more predictive of diabetes than BMI. Thus, the best anthropometric body measure as a predictor of T2D is still debatable.

*Blood pressure*: Systolic blood pressure and the use of anti-hypertensive medication is significantly associated with the incidence of T2D after controlling for multiple risk factors in
the Atherosclerosis Risk in Communities (ARIC) cohort study (Gress et al., 2000). A cohort study of the Campania Salute Network in Italy also found that incidence of diabetes was significantly higher in patients with uncontrolled blood pressure than in those with controlled blood pressure (Izzo et al., 2009). Moreover, studies indicate that the association between blood pressure and T2D is not only strong but also consistent, even over a period of 10 years (Movahed, Sattur, & Hashemzadeh, 2010). Though the exact pathological pathway is not well understood, endothelial dysfunction has been suspected as a possible pathway for the strong association between T2D and hypertension. For example, studies have shown that markers of endothelial dysfunction are associated with new-onset of diabetes (Meigs et al., 2004; Meigs et al., 2006), and endothelial dysfunction is closely related to blood pressure and hypertension (Taddei et al., 2003).

The current recommended blood pressure treatment for individuals with diabetes focuses on tightly controlling systolic blood pressure at value below 130/80 mmHg. However, studies showed that the attempt to lower systolic to the recommended 130 mmHg has also pushed the diastolic blood pressure to lower than 70 mmHg (Osher & Stern, 2008). This is problematic, especially in older adults with diabetes, because diastolic blood pressure increases with age until age 60 and decreases afterward (Osher & Stern, 2008). The decrease in diastolic blood pressure after age 60 is presumed to result from arterial stiffness due to age and lifestyle-related structural and functional change to the vessel wall (Avolio et al., 1983). Studies showed that lower diastolic blood pressure, even when systolic is as low as 120 mmHg, increases the risk for cardiovascular diseases (D’Agostino et al., 1991; Messerli et al., 2006; Osher & Stern, 2008).

Physical activity: Physical inactivity is a well-established risk factor for T2D (The American Association of Diabetes Educators [AADE], 2011). Several studies have shown that a sedentary lifestyle and low cardiorespiratory fitness are involved at several points in the progression from normal glucose metabolism to T2D (Colberg et al., 2010). Physical activity protects against development of T2D by influencing glucose homeostasis (LaMonte, Blair, & Church, 2005). Improvements in glucose homeostasis occur through both acute responses and chronic adaptations (Albright et al., 2000; LaMonte, Blair, & Church, 2005). Acute responses are changes in glucose uptake, transport, and disposal that occur during and after each period of physical activity (Albright et al., 2000). Chronic adaptations are changes in the physiologic system structure or function that occur from prolonged exposure to a specified dose of physical
activity (Albright et al., 2000). The largest improvement and maintenance of glucose homeostasis occurs through a regular and adequate dose of physical activity (LaMonte, Blair, & Church, 2005).

**Cigarette smoking:** Active smoking is an independent risk factor for T2D (Foy et al., 2005; Willi et al., 2007; Cho et al., 2009). In a systematic review, Willi et al. (2007) found that active smoking could be independently associated with glucose intolerance, impaired fasting glucose, and T2D. In a four-year follow up study, Cho et al. (2009) found that smoking is an independent risk factor for T2D and showed a synergistic interaction with the status of low insulin secretion and high insulin resistance for developing diabetes. After multivariate adjustment, current and ex-smokers have a higher risk of developing T2D compared with never-smokers (Cho et al., 2009). In addition, Cho found that the risk for current smokers is higher than ex-smokers. Thus, smoking may be a modifiable risk factor for T2D.

The mechanism involves nicotine action on the brain and endocrine system that may increase the activity of the level of hormones such as catecholamines (Targher et al., 1997; Kapoor & Jones, 2005), which may impair the action of insulin and can induce insulin insensitivity (Targher et al., 1997; Kapoor & Jones, 2005). It has also been reported that catecholamines impair the pathways that are related to the production of insulin and the activity and synthesis of the proteins that transport glucose to cells (Bjorntorp, Holm, & Rosmond, 1999). Moreover, nicotine may initiate an increase in the breakdown of fats (lipolysis) and an increase in free fatty acids level in the blood, which increases insulin insensitivity in the muscle (Gastaldelli, Folli, & Maffei, 2010).

**Socioeconomic:** A prospective cohort study by Demakakos, Marmot, and Steptoe (2012) found that current and lifespan socioeconomic status of individuals is inversely associated with incidence of T2D. In a meta-analysis, Agardh et al. (2011) found consistent increased overall risk of T2D with low levels of education, occupation, and income compared to high levels of these determinants. In a longitudinal cohort study, Lee et al. (2011) reported that advanced education and increasing income were both inversely associated with incidents of T2D. The mechanism by which socioeconomic position predisposes to diabetes is through psychosocial factors. Psychosocial factors such as low job control, low social support, depression, and effort-reward imbalance are established as risk factors for CHD (Kumari, Head, & Marmot, 2004).
Household food insecurity: Food insecurity is significantly associated with a high incidence of T2D and low glycemic control even after controlling for age, sex, educational attainment, household income, insurance status and type, smoking status, BMI, duration of diabetes, diabetes medication use and type, and presence of a usual source of care (Selgman et al., 2007; Bawadi et al., 2012; Berkowitz et al., 2013). Though the exact mechanism on how food insecurity is related to T2D is not well studied, two hypotheses have been posited in the literature (Gowda et al., 2012). The first hypothesis is that a household might eliminate an expensive healthy diet and may opt for cheaper high calorie foods due to budget constraints. An extended exposure to high calorie foods may lead to overweight and obesity, which are well-established risk factors for T2D. On the other hand, several studies find that food insecure household members’ energy intake was not higher than food secure households members (Zizza et al, 2008). The contradictory results imply that food security is mediated or interacts with other risk factors; further research is needed.

The second hypothesis takes into account that a food insecure household member may go hungry or have irregular meals. The lack of access to food and shift in diet patterns may entail psychological stress that may trigger secretion of systemic inflammatory markers such as C-reactive protein (CRP) and alter immune function. Gowda et al. (2012) tested this hypothesis using NHANES data to find that food insecurity is associated with CRP secretion, which is a proxy for inflammation and a correlate for chronic diseases including T2D. The later hypothesis may explain why the empirical evidence is equivocal with respect to establishing overweight and obesity (as measured by BMI) as an absolute mediator for the link between food insecurity and T2D (Gowda et al., 2012).

Psychological factors: Several prospective cohort studies (Melamed et al., 2006; Heraclides et al., 2009) found that social and psychological stress increases the incidence of T2D. Novak et al. (2012) found that perceived chronic stress is an important predictor of diagnosed diabetes in a Swedish middle-age cohort after controlling for age, socioeconomic status, physical inactivity, cigarette smoking, alcohol abuse, obesity, systolic blood pressure, use of anti-hypertensive medication, and serum cholesterol level. The biological mechanism through which stress induces onset of diabetes is by influencing cellular functions and altering the levels of metabolites (Lloyd, Smith, & Weinger, 2005). The psychological reaction to stressors may lead to the activation of the hypothalamo-pituitary-adrenal (HPA) axis, which in turn leads to
various endocrine abnormalities including high cortisol and low sex steroid levels (Lloyd, Smith, & Weinger, 2005). Stress stimulates the release of hormones by the hypothalamus and endocrines that antagonize the actions of insulin and results in higher levels of blood glucose (Chida & Hamer, 2008).

The relationship between diabetes and depression is bidirectional, but the majority of the studies tilt toward the hypothesis that depression is a risk factor for T2D (Renn, Feliciano, & Segal, 2008). A meta-analysis of cohort studies (Knol et al., 2005; Mezuk et al., 2008) found a temporal and strong relationship between depression and T2D. The use of longitudinal studies in the meta-analysis allowed the authors to make a statement about the relative risk of developing T2D depending upon whether a person has been diagnosed with or has symptoms of depression. A meta-analysis by Knol et al. (2005) examined nine empirical studies that investigated the longitudinal relationship between depression and onset of T2D and suggested that adults with either a depression diagnosis or high depressive symptoms have a 37% increased risk of developing T2D compared to those with no or little depressive symptoms. A meta-analysis of thirteen studies by Mezuk et al. (2008) also demonstrated that depression and/or depressive symptoms consistently predicted T2D, corresponding with a 60% increased risk of T2D. Recent longitudinal studies (Richardson et al., 2008; Campayo et al., 2010) and a cross-sectional study (Kan et al., 2013) show that depressed individuals have a higher risk of developing T2D. The biological mechanism hypothesized in the literature is similar to one discussed for stress in the preceding paragraph.

**Interaction between risk factors**

The pathogenesis of T2D is still not well understood. However, it is well-known that genetic predisposition, aging, and environmental factors contribute to the risk of developing T2D. Environmental factors include lifestyle (physical activity, diet, smoking, and alcohol consumption) and socioeconomic environments. Additional environmental factors such as stress and toxins may also play a role (Ali, 2013). These risk factors work together synergistically with complex and dynamic interactions, which lead to the development of T2D. Thus, T2D should be studied in a framework of a complex and dynamic systems approach (Huang et al., 2009). Figure 2.1 depicts the synergistic relationship between environmental and genetic factors in developing risk for T2D. The following paragraphs highlight some of the mechanisms through which T2D
risk factors interact with each other to help justify the proposed interactions between risk factors in our prediction model. This study adopted the epidemiological meaning of interaction where the term interaction implies a mutual dependency of two or more risk factors contributing to disease risk (Ottman, 1996).

**Genetics and lifestyle**

In an effort to establish the pathways of diabetes risks, an increasing number of genome-wide association studies (GWAS) have identified and validated single nucleotide polymorphisms (SNPs) associated with T2D (Steinthorsdottir et al., 2007; Sladek et al., 2007; Saxena et al., 2007; Scott et al., 2007; Voight et al., 2010). Though the association of SNPs with T2D is significant, the effect size of each genetic allele is very small with odds ratios ranging from 1.05–1.32 (McCarthy, 2011). However, the impact of genes as risk factors is pronounced when it interacts with lifestyle-related risk factors such as poor diet, physical inactivity, and smoking. Even if a person is genetically predisposed, T2D and obesity are unlikely to develop without the individual being exposed to promoting environmental and lifestyle risk factors (Temelkova-Kurktschiev & Stefanov, 2012).

The available evidence to date on the interactions between genetic and lifestyle risk factors of T2D are two-way interactions. For instance, interaction between genes and physical inactivity was reported by Brito et al. (2009), interaction between genes and obesity was reported by Ning et al. (2013), and interactions between genes and cigarette smoking was reported by Liu et al. (2011) and Yang et al. (2012). Ning et al. (2013) reported a strong joint effect of family history of diabetes (a surrogate for gene and common environmental risk factors) and obesity on the risk of T2D in Finish populations but not in Chinese populations. Liu et al. (2011) reported that heavy smoking was significantly associated with T2D in Chinese populations, and this association was moderated by the CYP2A6 genotype. In the Strong Heart Family Study, Yang et al. (2012) also showed that genetic variants in the nicotinic acetylcholine receptor (nAChR) genes, which have been associated with smoking phenotypes, contribute to insulin resistance and T2D among American Indians.

The possible biological mechanism for interactions between cigarette smoking and genes involves the metabolizer genotypes. Heavy smokers with slow metabolizer genotypes may be at higher risk for T2D as compared with the light smokers with normal metabolizer genotypes. This
is due to the fact that in slow or poor metabolizer genotypes, the pancreas is exposed to greater circulating levels of nicotine, which may decrease insulin sensitivity and insulin secretion and may also contribute to apoptosis of islet β-cells (Xie et al., 2009). About 80% of inhaled nicotine is metabolically inactivated (converted to cotinine, and then to trans-3-hydroxycotinine by the CYP2A6 enzyme), which is eventually excreted through urine (Rossini et al., 2008). The rate of conversion is moderated by CYP2A6 genetic polymorphisms (Rossini et al., 2008).

At the center of interactions between physical activity and genes is a transcriptional co-activator known as Peroxisome proliferator-activated receptor gamma co-activator 1-alpha (PGC-1α), which regulates the genes involved in energy metabolism (Franks et al., 2003). PGC-1α plays a key role in mitochondrial biogenesis, glucose and lipid transportation and oxidation, and skeletal muscle fiber-type formation (Liang & Ward, 2006). Physical activity training plays a role in increasing these PGC-1α mRNA levels (Liang & Ward, 2006), where the transgenic overexpression of PGC-1α induces an increased resistance to muscle fatigue and increases the maximum volume of oxygen consumption (VO2max) (Franks et al., 2003). Some studies have also suggested that decreased VO2max is among the earliest indicators of insulin resistance and T2D; therefore, it is an important risk factor for disease progression (Leite et al., 2008).

Several genetic loci, each of which provides an opportunity to define the genetic architecture and pathophysiology of both T2D and obesity, have been discovered recently (McCarthy, 2010). Lower expression of a widely studied gene, TCF7L2 (transcription factor 7-like 2) locus, leads to decreased insulin secretion (Renstrom, 2012). TCF7L2 is responsible for the difference in T2D risk between the two homozygote classes (a genotype consisting of two identical alleles of a gene for a particular trait); it accounted for approximately twofold risks (Grant et al., 2006). On the other hand, fat mass and obesity-associated (FTO) locus, which is also associated with obesity, decreases insulin sensitivity especially in a high fat acid environment (Brunetti, Chiefari, & Foti, 2014). The strongest evidence on the effect of FTO on obesity is that it is responsible for a 2.5kg difference in weight between homozygote groups (Frayling et al., 2007). Based on this discussion, it is not hard to draw on the possibility that the risk for T2D is higher when genes associated with T2D and genes associated with obesity are present simultaneously.
Genes, age, and lifestyle

Several studies reported on the role of interactions between genes, age, and lifestyle factors in increasing the prevalence of T2D at higher ages (Soriguer et al., 2014). Insulin secretion decreases with age due to many biological factors, including loss of Sirt1-mediated GSIS due to age, decreased β-cell sensitivity to circulating incretins, decrease in mitochondrial function as age increases, and increased oxidative stress with age (Scheen, 2005; Gong & Muzumdar, 2011; Soriguer et al., 2014). Though the risk of T2D increases with aging due to biological deteriorations, which results in abnormal glucose regulation, it may also contribute to the risk of developing T2D through changes in lifestyle at an older age (Scheen, 2005; Soriguer et al., 2014). Lifestyle and physiological changes associated with age such as dietary patterns, physical inactivity, and increased adiposity are risk factors for T2D (Scheen, 2005). In addition, a number of studies have indicated that changes in insulin resistance observed during aging were also related to an increase in adiposity, especially abdominal adiposity (Imbeault et al., 2003; Ferrannini et al., 1996; Gayoso-Diz et al., 2011; DeNino et al., 2001). This indicates that age interacts with lifestyle factors such as physical inactivity and diet to promote the risk of T2D.

Demographic, socioeconomic, and psychological factors

Additional interactions between risk factors of T2D that are worth mentioning include the modification effect of gender and adiposity on psychological risk factors and the joint effect of physical inactivity and hypertension. Psychological risk factors of T2D are modified by gender and adiposity level. A multivariate-adjusted analysis of the Midlife in the United States (MIDUS II) survey data showed that weight discrimination exacerbated the effects of abdominal adiposity; individuals who had higher WHR and reported weight discrimination had the highest HbA1c levels (Tsenkova et al., 2011). The biological mechanism through which psychological factors influence the effect of obesity is that psychosocial vulnerability factors, such as perceived discrimination and work stress, moderate the effects of BMI and central adiposity on glucose metabolism by releasing catecholamine, which increases the production of peripheral and central proinflammatory cytokines (Tsenkova et al., 2011; Heraclide et al., 2012). In the White-Shell study, Heraclide et al. (2012) reported that gender and body weight status play a critical role in determining the direction of the association between psychosocial stresses and T2D, where stress has a harmful effect for women but protective effect for men. The protective effect of
psychosocial work stress on T2D risk among non-obese men may be related to weight loss due to stress among lean individuals, while the negative effect of work stress among obese women probably reflects gender-specific psycho-neuro-endocrine pathways (Heraclide et al., 2012).

**Physical activity, hypertension, and adiposity**

Hypertension and T2D often occur together. Hypertension is approximately twice as common in persons with diabetes as in those without (Lago, Singh, & Nesto, 2007). Hypertension is not only comorbid with T2D but also may contribute to the onset of T2D directly. Hypertension reduces the delivery of insulin and glucose (via decreased blood flow to skeletal muscle, the major up-taker of body glucose) due to increased vasoconstriction and vascular rarefaction, by diminishing insulin-sensitivity of skeletal muscle slow-twitch fibers, and by decreasing insulin post-receptor signaling through PI3K-Akt pathway (Sowers, 2004). Obesity is a known risk factor for both T2D and hypertension. The pathway for obesity leading to T2D is through the excessive free fatty acid released by adipose tissue, which leads to a decrease in insulin sensitivity of muscle, fat, and liver (Guilherme et al., 2008). On the other hand, obesity leads to hypertension through increased renal sodium reabsorption and impaired pressure natriuresis (Hall et al., 2010). Pressure natriuresis is thought to participate in regulating the volume of extracellular fluid levels when the normal neurohumoral mediators are impaired—the increase in water and sodium ion excretions that occur when blood pressure is elevated because of an increase in the circulating blood volume (Granger, Alexander, & Llinas, 2002). This mechanism involves activation of the sympathetic nervous system, which is mediated in part by increased levels of the adipocyte-derived hormone (leptin), stimulation of pro-opiomelanocortin neurons, and subsequent activation of central nervous system (melanocortin 4 receptors) (Hall et al., 2010).

Physical activity reduces the risk of obesity and hypertension, the two most important risk factors for T2D (Padilla, Wallace, & Park, 2005). Thus, the joint risk of physical inactivity, obesity, and hypertension on T2D is much higher than the individual or additive effect of these factors. Xu et al. (2014) reported that the odds of developing T2D decreased across two-way interactions of inactive-hypertension, active-hypertension, inactive-normotension, to active-normotension groups, with the active-normotension groups having the lowest risk.
**Previous diabetes prediction models**

Models for predicting the risk of T2D in the population have been extensively developed for standard use in primary healthcare (Buijsse et al., 2011). Most of these models are in western countries (Buijsse et al., 2011) and some Asian countries (Ackermann & Marrero, 2007; Kilikkinen et al., 2007; Saaristo et al., 2007; Schwarz et al., 2007; Schwarz et al., 2009), and one model based on survey from several continents including Africa and Latin America (Vistisen et al., 2012). Few models were also developed to predict pre-diabetes risk (Kenneth et al., 2008; Yu et al., 2010). The prediction performance of these models may vary over time (as prevalence of risk factors in a population changes), between populations and geographic areas (Echouffo-Tcheugui et al., 2011), and the models typically perform well in populations in which they were developed (Glumer et al., 2006). The aim of this study is to develop a model for T2D prediction in the United States population. Thus, the review of the previous prediction models in this essay is concentrated on the United States population. However, a complete list of literature on T2D risk modeling (including validation of existing models in various populations) was reviewed in Nobel et al. (2011) and Buijsse et al. (2011). The models are intended to predict the risk of T2D based on measures routinely performed at primary care clinics or on data from questionnaires or laboratory tests.

To date, about eight risk models have been developed based on the United States population (see Table A.1 for a summary). Five of these are based on the data extracted from cohort study sample populations established for purposes other than T2D. T2D status in these models was assessed based on one or a combination of the following criteria: Fasting Plasma Glucose (FPG) ≥ 126mg/dL, 2-hour Oral Glucose Tolerance Test (2h-OGTT) ≥ 200 mg/dL, or if started using insulin (or oral hypoglycemic medication) and self-reported physician diagnosis. Three of the models are based on the NHANES, which is nationally representative cross-sectional data. Stern et al. (2002) developed a model based on data extracted from the San Antonio Heart Study participant cohort study, which is composed of Mexican-Americans and non-Hispanic whites. The prediction model has a flawed estimation, where 2h-OGTT was used both as diagnosis criteria and as predictor; lacks important factors (diet, physical activity, and psychological factors); and the sample is not representative of the US population. Kanaya et al. (2005) used cross-sectional data extracted from the Rancho Bernardo Study (RBS), which is based on an affluent, white, and older adult cohort from Southern California. Kanaya et al. 
(2005) developed a model based on the demographic variables and invasive procedures (triglyceride and fasting glucose), which is not different from costly and invasive formal screening methods. Schmidt et al. (2005) and Kahn et al. (2009) separately developed a risk prediction model based on non-invasive clinical measures (waist circumference, height, weight, hypertension, and pulse rate) and a questionnaire using data from the Atherosclerosis Risk in Communities (ARIC) cohort participants (composed of non-Hispanic whites and African Americans), from 1987-1989 and 1996-1998. The models created by Schmidt et al. (2005) and Kahn et al. (2009) are not based on representative samples of the US population. In addition, arbitrary (quintile) segmentation of continuous risk factors (age, waist circumference, pulse height, and weight), which may lead to biased estimation and a missed opportunity for finding biological cutoff points for the risk factors, were used in these models. The T2D prediction model created by Wilson et al. (2007) is based on the Framingham Offspring cohort study, which is a 99% non-Hispanic white sample. The simplest clinical model included variables such as age, sex, FPG, BMI, HDL cholesterol, parental history of diabetes, triglyceride level, and blood pressure. This model is as invasive (to obtain triglyceride) and as costly as the formal laboratory-based screening of T2D itself. Further, the model is limited in validity, because the sample is not representative of the US population (by ethnicity), has no interaction between variables, and lacks important factors (diet, physical activity, and psychological factors) for T2D risk.

Borrell et al. (2007), Heikes et al. (2008), and Bang et al. (2009) developed a predictive score for undiagnosed T2D based on the NHANES III data. Borrell et al.’s (2007) model targets a dental clinic setting to predict T2D by including factors such as age, sex, ethnicity, family history of diabetes, self-reported hypertension and hypercholesterolemia, and periodontitis in the final model. Heikes et al. (2008) developed a simple, non-invasive model based on age, waist circumference, history of gestational diabetes, family history of diabetes, ethnicity, high blood pressure, weight, height, parental diabetes, and exercise. However, their model left out psychological and socioeconomic health determinants. Bang et al. (2009) developed a self-evaluation score based on non-invasive measures (age, sex, family history of diabetes, history of hypertension, BMI or waist circumference, and physical activity). Though the aim was to screen for T2D in high-risk populations for lifestyle interventions, the model precludes lifestyle-related measurements (diet and smoking). In addition, some continuous variables such as BMI and age
are categorized, which limited the use of the full information from these variables to establish diabetes-specific cutoff values.

**Research gap**

The risk prediction models developed to date, both in the United States and in other countries, mainly focus on non-modifiable risk factors such as age and family history, ethnicity, and on the consequences of adverse health behaviors such as high BMI, waist circumferences, and high blood pressure (Nobel et al., 2011). These models have not addressed several complex issues that may affect their applicability to identify high-risk individuals for screening and interventions. First, the T2D prediction models to date have completely ignored the importance of psychological, food insecurity, and other socioeconomic factors. Socioeconomic gradients such as income, education level, access to healthcare, access to health information, and attitude toward risk factors are important to understand and act upon for T2D risk factors. In the United States, adults with low socioeconomic position are disproportionately affected by diabetes and its complications (Brown et al., 2004). Among adults with diabetes, lower socioeconomic status (SES) is associated with many factors known to contribute to poor health outcomes, including reduced access to and underuse of recommended preventive care, poor metabolic control, and psychological distress (Brown et al., 2004). Further, food insecurity, which limits individual choice and induces reduced quantity and quality of food availability at all times as well as emotional distress, is also important in predicting pre-diabetic and diabetic glycemic level (Selgman et al., 2007). After all, prevention programs based on lifestyle modifications can only be implemented if the individuals are capable of choosing their diet and following up on the healthy eating recommendations for controlling glycemic level (Selgman et al., 2012). Similarly, despite the mounting evidence regarding the association between psychological factors and T2D, the prediction models seem to dodge this important risk factor. Prospective cohort studies (Melamed et al., 2006; Heraclides et al., 2009) have found that social and psychological stress increases the incidence of T2D. Novak et al. (2012) found that perceived chronic stress is an important predictor of diagnosed diabetes after controlling for age, socioeconomic status, physical inactivity, cigarette smoking, alcohol abuse, obesity, systolic blood pressure, use of anti-hypertensive medication, and serum cholesterol level in a Swedish middle-age cohort.
Second, the models are additive in nature (Vistisen et al., 2012) and do not take into account the fact that risk factors interact with each other not only globally, but also at a local level (Collins et al., 2011). That is, different parts of the variables interact with each other at different levels. If these local interactions between variables are not incorporated, the accuracy of the model in predicting T2D is compromised. In addition, the cutoff point used in the existing risk prediction models, for the risk factors such as anthropometric measures, are based on the existing classification threshold for normal, overweight, and obese individuals (Buijsse et al., 2011). However, the influence of such risk factors as anthropometric measures of obesity on glycemic level is continuous and non-linear (Wang et al., 2005) and may not accord with the existing cutoff thresholds (Buijsse et al., 2011). Thus, more investigation is needed as to what cutoff threshold should be used for relevant anthropometric measures of obesity in setting the criteria for screening high-risk individuals.

In summary, previous models ignored socioeconomic, food insecurity, and psychological factors; assumed linear relationship between predictor and T2D; assumed no interaction between variables predicting T2D; and used arbitrary (e.g., quintile-based segmentation) or non-diabetes specific cutoff values for continuous predictor variables. The current study aims to develop a prediction tool for identifying individuals at high-risk for T2D based on non-invasive clinical measures (anthropometric and blood pressure), psychological factors (stress and depression), economic factors (income level and education), food insecurity and diet quality, and non-linear interactions between these factors. The model we aim to develop is distinct in several ways: 1) this is the first model to include psychological factors, food insecurity, and diet quality in addition to the demographic and non-invasive clinical measures; 2) this is also the only model to consider non-linear interactions between predicting variables.

**Research Questions**

Do interactions between demographic factors (age, sex, and ethnicity), anthropometric measures (body mass index, waist circumference, and waist-to-height ratio), socioeconomic factors (income-to-poverty ratio, food insecurity, and education level), lifestyle (physical activity level, smoking status, and alcohol consumption), and depression improve the accuracy of predicting T2D?
Objective
To develop a model for predicting diabetes that relies on easy-to-obtain, non-invasive, routinely collected anthropometric body measures, demographic (age, sex, and ethnicity), anthropometric measures (body mass index, waist circumference, and waist-to-height ratio), socioeconomic (income-to-poverty ratio, food insecurity, and education level), lifestyle (physical activity level, smoking status, and alcohol consumption), and depression measures.

Hypothesis
The interaction between demographic (age, sex, and ethnicity), anthropometric measures (body mass index, waist circumference, and waist-to-height ratio), socioeconomic (income level, food insecurity, and education level), lifestyle (physical activity level, smoking status, and alcohol consumption), and depression measures accurately predicts T2D.

2.2 Methods
Data
This study utilizes a data set extracted from the National Health and Nutrition Examination Survey (NHANES), which is a cross-sectional and nationally representative survey of the non-institutionalized US civilian population. It collects demographic, health history, and behavioral information from participants in home interviews. Subsamples of participants are also invited for detailed physical, physiological, and laboratory examinations that are performed by trained personnel in specially equipped mobile centers (NHANES, 2014). Data from the 2005-2006, 2007-2008, 2009-2010, and 2011-2012 waves of NHANES were combined to obtain a larger sample size. The data used in this study is limited to non-pregnant participants and individuals aged 20 or older. The combined sample size, with non-missing responses to the question: “Other than during pregnancy, have you ever been told by a doctor or health professional that you have diabetes or sugar diabetes?” or those who have valid laboratory test results for diabetes, is 22,660. However, independent variables in the regression have lots of missing data. Therefore, the effective regression due to sample list wise deletion is 5,471.
Dependent variable

Type 2 diabetes

The classification of survey participants into diabetic and non-diabetic is based on the answer to survey questions and biochemical measures (plasma glucose and HbA1c). Survey participants were considered to have diagnosed diabetes: a) if they answered “yes” to the question, “Other than during pregnancy, have you ever been told by a doctor or health professional that you have diabetes or sugar diabetes?”; b) participants who answered “no” to the preceding question but who had a measured fasting plasma glucose ≥ 126 mg/dl or HbA1c ≥ 6.5% and were considered to have undiagnosed diabetes; and c) participants with self-reported diagnosed diabetes but with measured fasting plasma glucose < 126 mg/dl and HbA1c < 6.5% (which does not conform with the self-reported results) and were considered to be diabetic if taking insulin and/or oral hypoglycemic medication is reported. This approach of classifying survey participants into diabetes status has been used in several previous studies (Koopman et al., 2005; Selgman et al., 2007; Berkowitz et al., 2013). This classification approach is the most accurate method of ascertaining T2D, with reported sensitivity of about 95% in NHANES data (Midthjell et al., 1992). Participants who did not self-report diagnosis with diabetes and with measured fasting plasma glucose < 126 mg/dl and HbA1c < 6.5% and with no medication reported were classified as non-diabetic. Survey participants were classified as non-diabetic if they responded “no” to the self-reported diabetes diagnosis question, but self-reported to taking insulin and/or oral hypoglycemic medication. Demmer et al. (2013) also used this last classification approach.

Though type 1 and 2 diabetes are distinct diseases, NHANES data does not provide a clear way of differentiating between them. The survey neither asked if the diagnosis is for type 1 or 2 diabetes nor performed lab examination of islet cell autoantibodies to identify the type. This study follows a method outlined by Koopman et al. (2005), which used biological pathways of the diseases to differentiate between type 1 and 2 diabetes diagnoses in those who self-reported a diagnosis of diabetes. Accordingly, those who reported a diagnosis of diabetes at an age less than 30 and were only using insulin to manage their diabetes were classified as T1D (15 adults) and excluded from further statistical analysis. Those participants who reported a diagnosis of diabetes at an age less than 30 and were using both insulin and hypoglycemic medication or only hypoglycemic medication to manage their diabetes were classified as T2D. All participants who
self-reported a diagnosis of diabetes at an age greater than 30 were classified as T2D. According to this classification, 3,961 (about 17.5% of the sample) have diabetes. The sample weighted population estimate shows about a 12.9% diagnosed and undiagnosed diabetes prevalence in the United States, which is only marginally higher than the official CDC estimate of 12.3% in 2012. The slight difference in prevalence is caused by classification criteria differences, which is the addition of self-reported use of insulin and hypoglycemic medication or only hypoglycemic medication as classification criterion in this study.

**Independent variables**

Demographic variables considered as covariates are age, gender, marital status, and ethnicity/race (Tables 2.1 and 2.2). Age was measured in years and included in the analysis as a continuous variable. The age range of the sample for this study was adults from 20 years to 84 years old, with a mean age of 47. Gender was categorized as male (45.8%) and female (54.1%). Marital status was re-categorized to never married, married (including those living with a partner), and separated (including those separated, divorced, and widowed). The marital status category was re-coded to simplify the relationship status as it relates to the extent of social support. Self-reported race/ethnicity was categorized as Mexican-Americans (17.9%), other Hispanics (10.5%), African Americans (18.2%), non-Hispanic whites (48.8%), and others (including Asians, American Indians, and multiracial, 4.6%).

Several anthropometric measures of overweight and obesity were included in the NHANES data. These measures include weight, height, waist circumference (WC), triceps skinfold, and subscapular skinfold. The anthropometric measures of adiposity that are derived from these measures (or directly used) are: body mass index (BMI) to measure overall adiposity, WC to measure visceral adiposity, and subscapular skinfold (SubSK) to measure subcutaneous adiposity of back of the trunk (Table 2.2). BMI was calculated as weight in kilograms over height in meter squared (kg/m²), which ranged from 13 to 57 kg/m² with a mean of 27.5 kg/m². WC was measured at the abdomen right above the hip bone in centimeters (cm), which ranged from 59.7 cm to 166.4 cm with the mean of 95 cm. SubSK was measured with clipper (mm) at the inferior angle of the right scapula by pinching a double skinfold and underlying fat. SubSK measures ranged from 4.4 mm to 42 mm with a mean of 21.3 mm, respectively. The use of several anthropometric body measures account for the distribution and location of body fat in the
body. Since different anthropometric body measures are meant to measure adiposity at various locations of the body, this study utilizes all of the listed anthropometric body measures and selects the best predictor of diabetes in the statistical analysis.

NHANES utilizes WHO’s Global Physical Activity Questionnaire (GPAQ) to obtain self-reported physical activity through household interview. This questionnaire asks participants to recall their physical activity from the past 30 days. Participants report on frequency and duration of activities related to transportation (to and from work or school, shopping), activities related to work, and activities related to leisure (e.g., exercise, sports, and hobbies) (NHANES CDC, 2011).

Participants are asked to specify frequency and duration of work-related activities they perform for at least 10 minutes that are at a moderate intensity level (brisk walking, carrying a light load, etc.) or vigorous (carrying or lifting a heavy load, digging, or construction work, etc.). Participants are also asked to specify frequency and duration of transportation (walk or bicycle to work, school or shop) activities they perform for at least 10 minutes. For recreational activities, participants are asked to specify frequency and duration of activities they perform for at least 10 minutes that are at moderate (brisk walking, bicycling, swimming, or golfing) and vigorous intensity (basketball, running, etc.) levels. The questionnaire defines moderate intensity activities as activities causing light sweating or a slight-to-moderate increase in heart rate or breathing, and vigorous intensity activities as activities causing heavy sweating or large increases in breathing or heart rate (NHANES, 2011).

This study utilizes the combined duration of moderate and vigorous physical activity from work, transportation, and leisure. An estimate of average daily time spent in moderate to vigorous physical activity (MVPA) is computed by adding up the minutes of reported moderate and vigorous physical activity from transportation, work, and leisure activities from the seven days of recall and dividing the total number of minutes by seven.

\[
MVPA = \frac{\sum_{i=1}^{3} (\text{minutes in moderate activity} \times \#\text{days})_i + (\text{minutes in vigorous activity} \times \#\text{days})_i}{7 \text{ days}},
\]

where \(i\) is the type of activity (work, transportation, and leisure). The study sample calculated the MVPA range from zero to 1,371 min/day with a mean of 91 min/day (Table 2.2).
Smoking status, number of cigarettes, and frequency of smoking was self-reported and collected during the household interview. Participants who reported smoking at least 100 cigarettes in their entire life and smoking every day or some days at the time of the interview were classified as active (current) smokers. Participants who reported smoking a minimum of 100 cigarettes in their entire life and not smoking at all at the time of the interview were classified as ex-smokers. Participants who reported smoking fewer than 100 cigarettes during their lifetime (including new starters) were defined as never-smokers. This smoking classification is based on the definition that has been widely used in the literature (CDC, 2009; Clair et al., 2011). Accordingly, 69.5% of the study sample is considered “never-smokers”, and 30.5% of the study sample is considered current or ex-smokers (Table 2.1). Similarly, alcohol consumption was calculated using NHANES questionnaires, which included frequency and amount of consumed alcohol. The consumption was categorized into abstainers (69.5%), occasional drinkers (46.5%), moderate drinkers (16.5%), and heavy drinkers (21.6%).

Participants’ family history of diabetes was based on the following questions in NHANES: “Including living and deceased, were any of your biological, that is, blood relatives including grandparents, parents, brothers, sisters, ever told by a health professional that they had diabetes?” Prior to the 2005/6 survey round, “yes” responses were probed by asking participants to specify which of the biologic relatives were affected. Starting from the 2005/6 survey round, however, this probing for specifics was discontinued. Thus, the family history of diabetes was entered into the model as “yes” or “no” dummies. This question also did not distinguish if a relative’s diabetes status was type 1 or type 2. About 39.3% of the study sample indicated having relative(s) who were told by health professionals that they had diabetes (Table 2.1).

Depression in NHANES data is based on the Mobile Exam Center (MEC) administered patient health questionnaire (PHQ). The PHQ-9 is a nine-item, self-administered depression measure that has been well-validated in two large studies involving 3,000 patients in eight primary care clinics and 3,000 patients in seven obstetrics-gynecology clinics (Spitzer, Williams, & Kroenke, 2000). The self-administered PHQ-9 has a diagnostic validity comparable to the clinician-administered version. The PHQ-9 is the most commonly used depression questionnaire both in clinical and research settings (Kroenke, Spitzer, & Williams, 2001). This tool helps not only to diagnose depression but also to compute a severity score (Kroenke, Spitzer, & Williams, 2001). The total PHQ-9 score ranges from 0 to 27 because each item is rated on a 4-point scale.
(0 being “not at all” and 3 being “nearly every day”) (Kroenke, Spitzer, & Williams, 2001). The score of 5–9, 10–14, 15–19, and more than 20 represent mild, moderate, moderately severe, and severe depression, respectively (Kroenke, Spitzer, & Williams, 2001). This study computed the mean depression score and utilized the mean score as a continuous variable in the model. Using the continuous score helps to define what severity level of depression is relevant as a risk factor for T2D. The mean depression score for this sample ranged from 0 to 24 with a mean of 2.6 (Table 2.2).

In NHANES data, participants’ hypertension was measured based on the questionnaire and medical exam. Hypertension was ascertained based on a blood pressure (BP) measure. Blood pressure was measured three to four times by mercury sphygmomanometer using a standardized protocol (NHANES, 2014). Hypertension was defined as systolic blood pressure (SBP) ≥140 mmHg or diastolic blood pressure (DBP) ≥90 mmHg, or on antihypertensive medication; pre-hypertension was determined as SBP ≥120 mmHg or DBP ≥80 mmHg, but not meeting the criteria for hypertension (Chobanian et al., 2003). However, blood pressure was included in the statistical analysis as a continuous variable not as a category of hypertensive and non-hypertensive. Average SBP and DBP were calculated from three to four rounds of BP measurements. The distribution of SBP and DBP ranged from 78 to 228 mmHg and from 10 to 132 mmHg with a mean of 121.8 mmHg and 69.6 mmHg, respectively (Table 2.2).

The socioeconomic variables used are based on the household interview part of NHANES. Socioeconomic covariates that are included were education level and family income to poverty ratio (RFItPov). Educational attainment was categorized as less than high school diploma, high school diploma, and some college or higher, which was respectively about 28%, 24%, and 48% of the study sample (Table 2.1). The mean RFItPov was 2.5 with a minimum of 0 and a maximum of 5 (Table 2.2).

Food insecurity is defined as uncertainty about future food availability and access, insufficiency in the amount and kind of food required to lead a healthy lifestyle, or the need to use socially unacceptable ways (stealing or scavenging) to acquire food (Wunderlich & Norwood, 2006). The food security questionnaire in NHANES data contains 18 questions at household level; 10 items are related to adult food insecurity and eight items are related to child food insecurity. Our interest is to measure food insecurity of adults in the household. Thus, only 10 items are used to classify adults into four food security categories as set out in USDA
guidelines: full food security (0 points), marginal food security (1–2 points), low food security (3–5 points), and very low food security (6–10 points). Accordingly, the food security level was 69.8% full food security, 10.8% marginal food security, 12.3% low food security, and 7.1% very low food security.

Statistical Analysis

The current study utilizes Multivariate Adaptive Regression Spline (MARS) to build parsimonious and accurate diabetes prediction model. MARS is chosen as a statistical method because it accounts for both linear and non-linear relationships as well as piecewise interactions between predicting variables in determining diabetes. This approach enables us to model the interaction effect of predicting variables not only globally but also locally (between basis functions), since it segments the predicting variables into several pieces known as basis functions or splines. The piecewise approach of the method is also appropriate to pinpoint the accurate cutoff values of the anthropometric measures in relation to diabetes.

Multivariate Adaptive Regression Splines (MARS)

MARS is a non-parametric technique that makes no functional form assumptions about the underlying relationship between the dependent and independent variables. It is a procedure for fitting adaptive non-linear regression that uses piecewise linear basis functions and interaction between basis functions to define relationships between dependent and independent variables (Friedman, 1991; Trevor, Tibshirani, & Friedman, 2001). MARS is better at handling situations such as a high number of predictors, non-linearity, multicollinearity, and a high degree of interaction among predictors (Trevor, Tibshirani, & Friedman, 2001).

The basis functions (also known as splines) in MARS are a series of linear functions formed by segmenting independent variables. The basis functions are expressed as pairs with an inflection point (knot or value of independent variable), which are unknown a priori and need to be estimated from the data. Let Y be dependent variable (dummy for diabetes in our case) and X is set of predictor variables \( x_1, ..., x_p \). The basis function for each predictor variables \( x_j \) (\( j = 1, ..., p \)) is expressed as:

\[
(x_j - t)_+ \quad \text{and} \quad (t - x_j)_+ \tag{1}
\]
where \( t \) is a knot (a data value in \( x_{ij} \)) that marks a reflection point in predictor variable \( x_j \) \((i = 1, ..., N \) are observations in the predictor). Since “+” means positive parts of the function, the reflected pairs are:

\[
(x_j - t)_+ = \max(x_j - t, 0)
\]

and

\[
(t - x_j)_+ = \max(0, x_j - t)
\]

Initially, MARS forms reflected pairs for each predictor with knots at each observation in \( x_{ij} \), which create a \( 2 \times p \times N \) basis function in \( R^p \) space:

\[
C = \{(x_j - t)_+, (t - x_j)_+\}, \quad t \in \{x_{1j}, x_{2j}, ..., x_{nj}\}
\]

The basis function over the entire of \( R^p \) space can be represented by:

\[
h(X) = (x_j - t)_+
\]

The model building process starts with a constant function, \( h_0(X) = 1 \), and adds all possible new pairs of basis functions in \( C \) and their products in a forward stepwise process. The MARS algorithm is an iterative forward stepwise process that adds new pairs of functions to the model if it produces the largest decrease in the residual sum of squares by considering all possible pairs of new functions:

\[
h_m(x)(x_j - t)_+ \quad \text{and} \quad h_m(x)(t - x_j)_+
\]

where \( h_m(x) \) is a basis function that is already included in the model and does not depend on the newly added predictor \( (x_j) \). At the end of the iterative forward stepwise process, the model has the form:
\[ f(X) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(X) \]  

(6),

where \( h_m \ (m = 1, \ldots, M) \) is a basis function in \( \mathcal{C} \) or a product of two or more basis functions. \( \beta_0 \) is a parameter for the constant basis function \( h_0(x) \), and \( \beta_m \) is a parameter for \( m^{th} \) basis function or product of basis functions. These parameters are estimated by the least square method (Trevor, Tibshirani, & Friedman, 2001). The advantage of MARS is that the basis function and product of one or more basis functions in \( h_m \) allows for modeling both additive and interaction effects of predictor variables in determining the dependent variable.

MARS algorithm searches for all possible knots in each predictor variable and adds the resulting basis functions (or the products) until the change in residual error is too small to continue or the user defined maximum number of functions is attained. At the end of the process, the model in equation (6) over-fits the data (Trevor, Tibshirani, & Friedman, 2001). Thus, the backward pruning process starts to remove terms that are not sufficiently contributing to the model. The backward pruning process takes out one term at a time, where terms that upon their deletion cause minimal increase in residual squared error are removed at each stage of the iterative process. A predictor variable can be dropped from the model completely if none of its basis functions sufficiently contribute to predictive performance of the model. The optimal model minimizes the generalized cross-validation:

\[ GCV(\lambda) = \frac{\sum_{i=1}^{N} (y_i - f_\lambda(x_i))^2}{(1 - M(\lambda)/N)^2} \]  

(7)

\( M(\lambda) \) is an effective number of parameters in the model, which account for both the number of terms in the models and number of parameters used to select the optimal position of the knots. \( \lambda \) is the optimal number of terms (basis function and their products) in the model. If \( r \) is linearly independent of the basis functions in the model, \( c \) is a constant used as penalty for complexity (for each basis function included), and \( k \) is the number of knots selected in the forward stepwise process, then,

\[ M(\lambda) = r + ck \]  

(8)
The empirical recommendation is to use $c = 2$ and limit the interaction terms to 3 (Trevor, Tibshirani, & Friedman, 2001).

**Model validation**

Multivariate adaptive regression splines models were built through K-fold cross validation. K-fold cross-validation separates the data into K equal sized parts. The model is fit with the K–1 parts, and the prediction error of the fitted model is calculated when predicting the kth part, or the out-of-fold data (Trevor, Tibshirani, & Friedman, 2001). The process is conducted for $k = 1, 2, ..., K$, and the prediction error is averaged across all out-of-fold predictions. The out-of-fold R-squared is averaged (cross-validated $R^2$) from the left-out subset, which is an estimate of the model performance on independent data (Trevor, Tibshirani, & Friedman, 2001). The generalized R-squared (GRSq) is based on the raw GCV and is a generalization of model performance.

**Handling missing values**

MARS creates a basis function for any of the variables with missing data. It creates a basis function with a dummy of 1 when the data is missing on a given variable and 0 when the data is not missing. MARS searches for a systematic relationship between missing data and dependent variables and interactions between missing data basis functions and other variables’ basis functions in predicting the dependent variable. Thus, MARS determines if other variables can act as surrogates for the missing variable. For our purpose, the data was analyzed both with and without missing values. This helps to determine the sensitivity of the model to the missing values in the predictor variables.

**Model evaluation**

The MARS prediction performance is evaluated with area under the recover operating characteristics (ROC) curve. Statistics used to construct ROC graphs are sensitivity and specificity. Other statistics such as positive predictive value (PPV), negative predictive value (NPV), and accuracy are also important indicators of model performance. Let true positive (TP), true negative (TN), false positive (FP), and false negative (FN); the indices are calculated as follows.
Sensitivity is defined as a percentage of correctly identified individuals with a diabetes diagnosis:

\[ \text{sensitivity} = \frac{TP}{TP + FN} \]

Specificity is defined as a percentage of correctly identified individuals with non-diabetes:

\[ \text{specificity} = \frac{TN}{TN + FP} \]

Positive predictive value (PPV) measures the percent of the times that the positive value is the true positive value (i.e., the percent of all positive diabetes prediction that are true positives diabetes diagnosis).

\[ \text{PPV (precision)} = \frac{TP}{TP + FP} \]

Negative predictive value (NPV) measures the percent of the times that the negative value is the true negative value (i.e., the percent of all positive non-diabetes prediction that are true non-diabetes).

\[ \text{NPV} = \frac{TN}{TN + FN} \]

Overall accuracy of the prediction can also be calculated as:

\[ \text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \]

On an ROC graph, sensitivity is plotted on the y-axis and specificity is plotted on the x-axis. All possible combinations of sensitivity and specificity statistics create ROC space (area). The graph starts at coordinate (0,0) and ends at coordinate (1,1). The coordinate (0,0) represents a situation where a predictor is never issuing a positive prediction; such a predictor commits no false positive errors but also gains no true positives. The other extreme coordinate (1,1) provides positive predictions unconditionally (Fawcett, 2006). The position of a point in the ROC space shows the tradeoff between sensitivity and specificity (i.e., the increase in sensitivity is
accompanied by a decrease in specificity) (Fawcett, 2006). The two-dimensional measure of ROC is reduced into a scalar measure with the use of the Area under ROC Curve (AUC):

\[
AUC = \int_{0}^{1} ROC(t) dt,
\]

where \( t = (1 - \text{specificity}) \) and \( ROC(t) \) is sensitivity. AUC provides an interpretable measure of the accuracy of a prediction model. A larger AUC shows that the diagnostic test is more accurate.

### 2.3 Results

#### Additive MARS model

In this study, the analysis of two MARS models was conducted to test the hypothesis that including non-linearity and non-linear interactions in the model would improve the prediction of diabetes status. The model building process was started with 19 variables that, according to the literature, are associated with an increased probability of diabetes but do not involve invasive measurement procedures. The initial model included age, gender, ethnicity, marital status, education level, family history of diabetes, cigarette smoking, alcohol consumption, household food security, depression scores, RFItPov, WC, BMI, SubSF, SBP, sleep duration, and MVPA.

The reference model for the performance of this prediction model is a logistic regression model without interaction between variables but all candidate variables included. The first model was a non-linear additive MARS model where no interaction between basis functions was allowed in the model. This approach helps to identify a non-linear relationship between the predicted probability of diabetes status and predicting variables. In addition, the result from the additive MARS model helps to establish cutoff points for continuous variables.

The result from the additive MARS model shows that the forward selection process transformed the variables into 34 basis functions (piecewise linear relationship between predictors and probability of diabetes) including the intercept. Fourteen variables were used out of 19 variables to form 33 basis functions. The variables involved in the forward selection process of basis were age, WC, family diabetes history, ethnicity, ratio of family income to poverty, BMI, subscapular skinfold, systolic blood pressure, diastolic blood pressure, MVPA, ethnicity, education level, sleep duration, and depression scores. As discussed earlier, the
forward MARS selection process over fits the prediction model; thus, one basis at a time is removed in the backward selection process. Basis functions that least increase the sum of error squares when removed from the model were dropped during the backward selection process. The final additive MARS model through the backward selection process has 25 basis functions but is still composed of the same 14 variables selected during the forward process. The basis, knots (or levels for categorical variables), and regression coefficients are shown in Table 2.3. The model component plots show individual variables’ relationship to the fitted function (Figure 2.2 and Figure 2.4).

The results from the additive MARS model show that the probability of the onset of diabetes is non-linearly associated with predicting variables such as age, WC, SBP, DBP, MVPA, and sleep duration. Figure 2.2 shows the non-linear contribution of predictors to the risk of diabetes. Note that the contribution of each predictor is interpreted in the context multivariate model. We can see that the contribution of WC to the risk of diabetes is zero up to 88 cm and continues to grow linearly afterwards. The contribution of age to the risk of diabetes is none until age 42, where it increases steeply until age 69. A significantly high risk of diabetes starts at age of 47. Age is not the main contributing factor to risk of diabetes after age of 69; other factors become more important. The contribution of SBP to the risk of diabetes is less than 125 mmHg. The risk slightly increases from 125 mmHg up to 157 mmHg. The real risk of diabetes starts at SBP above 157 mmHg. On the other hand, an increase in DBP to 60 mmHg decreases the risk of diabetes and the risk of diabetes increases after 78 mmHg, which can be regarded as a U-shaped curve. Any amount of MVPA is better than sedentary life, but the high diabetes risk reduction was achieved starting 41 min per day, which is close to the current recommendation (30–60 min MVPA per day) by CDC and WHO. There was no diabetes risk reduction by extending MVPA beyond 132 min per day. Sleep duration has a U-shaped association, where sleep duration less than seven hours and greater than eight hours increases the risk of diabetes. The risk of diabetes increases with an increasing level of depression scale score. However, the depression scale score of 11 and above increases the risk of diabetes significantly.

The predicting power of the additive MARS model was assessed with AUC, which is 0.8504 (95% Wald CI: 0.837, 0.864) for this additive MARS model (Figure 2.3). The DeLong et al. (1988) test for statistical difference between AUC of the additive MARS model and logistic model (with all 19 variables included) is not significant. In addition, the model has 86.75%
accuracy to predict true positive and true negative rates. The sensitivity and specificity of the model can be set based on the purpose of the screening for diabetes. A cutoff at probability of 0.19 would yield balanced sensitivity and specificity of 78%.

**Two-way interaction MARS model**

The second model incorporated two-way interaction between basis functions. The coefficients for the standalone basis and two-way interactions are presented in Table 2.4. The final selection process produced 20 basis functions including the constant. Twelve two-way interactions between basis were selected into the final model. The interactions between the basis of WC and SubSF, age and SBP, depression score and DBP, ethnicity and DBP, ethnicity and sleep hours, education and alcohol consumption, ethnicity and age, and BMI and SBP were selected into the final model because they significantly contribute to the risk of diabetes.

As shown in Table 2.4 and Figure 2.4, interaction between WC > 69 cm and SubFS < 27.8 mm decreases the risk of diabetes. This indicates even if the risk of diabetes increases linearly with WC > 69 cm, the lower range of subscapular fat is protective. The interaction between age lower than 69 and SBP > 192 mmHg increases the risk of diabetes. On the other hand, interaction between age lower than 69 and SBP < 192 mmHg decreases the risk of diabetes. The interaction between age lower than 69 and SBP > 162 mmHg increases the risk of diabetes. The interaction between depression score < 11 and DBP < 77 increases the predicted diabetes, which indicates even if the symptoms of depression are at lower range, individuals with inflexible blood vessels are at increased the risk of diabetes. The interaction between lower range DBP (< 69 mmHg) and older age (above 50) increases the risk of diabetes significantly. From Figure 2.4, the greatest risk of diabetes is observed at intersection of DBP 50 mmHg and age above 60. Similarly, the interaction between age above 50 and increased SBP (above 125) increases the risk of diabetes. The highest risk is observed at age above 69 and SBP above 200 mmHg. Figure 2.4 also shows that as the gap between DBP and SBP widens, the risk of diabetes also increases.

Having a relative with a history of diabetes, as surrogate for genetic predisposition, is one of the major predictor of diabetes both independently and in interaction with other variables. The interaction between having a relative with a history of diabetes and MVPA shows the difficult effort it takes to overcome the risk of diabetes from genetic predisposition. Even at MVPA as
high as 233 min/day, the risk of diabetes remained non-zero. Similarly, the interaction between ethnicity and other risk factors points to the importance of genetic predisposition and socioeconomic status even when other risk factors are favorable for developing diabetes. Being of white ethnic origin and having DBP < 57 did not increase the risk of diabetes. Being ethnic white also interacts with sleep duration; sleeping less than seven did not increase the risk of diabetes in whites.

The two-way interaction MARS model has an AUC of 0.858 (95% Wald CI: 0.845, 0.871). The two-way interaction MARS model improved the prediction of diabetes over the non-linear additive model slightly, with an AUC of 0.0082 improvement (Figure 2.5). The statistical test for the difference between the AUC of the two-way interaction model and the additive model with Delong et al. (1988) was significant ($\chi^2 = 12.63, p < 0.001$). There was also very small improvement in the accuracy of the prediction for true positive and negative rates over the additive MARS model (87.37% vs. 86.75%).

2.4 Discussions

The objective of this study is to test the hypothesis that non-linear interactions between non-invasive, simple, and easily available diabetes risk factors improve the prediction of diabetes in the representative US population. The study implemented a non-parametric regression approach known as Multivariate Adaptive Regression Spline (MARS) to test this hypothesis. This approach allows the breaking of the non-linearly associated variables into pieces of linear segments (known as basis) and the interactions between these linear segments of various variables. In addition, the logistic regression model was used as a reference model. The fitness of the models were evaluated with 10-fold cross-validation on the same data. This approach is useful to evaluate model fitness in the same data when a new data set with similar representative population and measurements is not available.

The improvement in AUC of the new models over the reference model was statistically evaluated with Delong et al.’s (1988) $\chi^2$ based test. An additive non-linear MARS model can predict diabetes with 86.75% accuracy and AUC of 0.85. This level of accuracy and AUC is higher than using linear regression models such as the logistic regression model. The AUC achieved with this model is higher than similar other studies based on non-invasive measurements-based models such as Kahn et al. (2009) with AUC of 0.71, Schmidt et al. (2005)
with AUC of 0.71, and Bang et al. (2009) with AUC of 0.79. The additive MARS model achieved the same AUC of 0.85 with Heikes et al. (2008), which used the classification and regression tree from the NHANES III data. The predictive capacity of this model is even better than or similar to the models that incorporated invasive measures including lipid profiles. Such models have AUC ranging from 0.73 to 0.85 (Kanaya et al., 2005; Stern et al., 2002; Wilson et al., 2007; Borrell et al., 2007).

More importantly, the additive non-linear MARS model has discovered the non-linear relationships between some of the predicting variables and risk of diabetes. Among the most important non-linear associations are the relationships of SBP and DBP to the risk of diabetes, which are J- and U-shapes respectively. An increase in SBP up to 125 mmHg has no significant risk of diabetes and SBP above 157 mmHg increases the predicted risk of diabetes. Similarly, an increase in DBP up to 60 mmHg decreases the risk of diabetes and DBP over 78 mmHg increases the predicted risk of diabetes. This finding indicates that the risk of diabetes decreases within the BP range of 125–157/60–78 mmHg. Note that the cutoff point for the blood pressure control both in individuals with and without diabetes is still up for debate. The UKPDS study results showed that patients with tight BP control (achieving a mean BP of 144/82 mmHg) had reduced macrovascular and microvascular events (Adler et al., 2000). Recent meta-analysis and the Cochran systematic review of randomized control studies and cohort-based observational studies did not find evidence for tight BP control of less than 130/80 mmHg as recommended in some guidelines (Arguedas, Perez, & Wright, 2009; Catena et al., 2015).

Another interesting finding is that the protection effect of MVPA starts at low MVPA minutes per day, but the significant reduction of risk of diabetes is achieved around 41 minutes per day and no gain in reducing the risk of diabetes occurs after MVPA of 132 minutes per day. Note that the WHO MVPA recommendation for diabetes and metabolic syndrome risk reduction starts at 21 minutes per day (150 minutes per week) (WHO, 2011). In addition, this study finds that sleep duration has a U-shaped association with risk of diabetes, where sleep duration less than seven hours and greater than eight hours increases the risk of diabetes. This result is similar to the cohort study of Yaggi, Araujo, and McKinlay (2006) who finds short and long sleep durations are risk factors for diabetes. However, the current study finds that less than seven hours as a natural cutoff point for short sleep opposed to six hours in Yaggi, Araujo, and McKinlay’s study.
In a subsequent analysis, a model that included two-way interaction between basis was used to test the hypothesis that interaction between the basis improves the prediction performance of the model. The two-way interaction model improved the prediction performance of the model over the non-linear additive model at AUC of 0.86 and accuracy of 87%. The difference in AUC between non-linear additive and two-way non-linear interaction models is statistically significant. This level of performance, as measured by the AUC, is a new achievement in both invasive and non-invasive diabetes risk prediction models in the US population. Moreover, the model identified interesting interactions between various variables. For example, having a relative with a diabetes history makes it harder for an individual to reduce the risk of diabetes through physical activity. The study also finds that the effect of sleep duration is mediated by ethnic group and age. In addition, the risk from increased WC and SBP and decreased DBP is mediated by age.

The strengths of this study are that the model uses easy-to-obtain information and non-invasive measurements with superior performance and accuracy. The information used for this model can be obtained through survey and non-clinical measurements at minimal resource settings such as community chronic disease prevention and workers’ well-being programs. The measurement for screening does not involve fasting or blood drawing. As a result, the participation rate for this type of screening method is expected to be higher than the screening methods involving invasive and inconvenient screening measurements such as a 2-hour glucose test. The second strength is that the NHANES data used to train and test the model is representative of the US population. The performance of the model was internally tested on a portion of the same sample, using an n-fold cross-validation method. The performance of this model can easily be replicated in any sample composed of a representative US population.

However, this study has its own limitation and should be interpreted carefully. First, the prediction model is not validated and calibrated as an independent and representative sample. The model should be validated in an independent, representative sample before implementation. Second, developing the risk prediction score is only part of the public health problem; the cost effectiveness and the ability of the tool to garner high participation rates are issues that need to be considered during the screening tool development process. In this regard, although the information and measurements in this model are easily accessible, the cost of collecting these measurements and information has not been evaluated. In addition, NHANES data is research
data; trained personnel accurately and repeatedly measured the anthropometric and blood pressure measurements. Replicating research standard accuracy in measuring anthropometric and blood pressures in a community setting may not be realistic. Similarly, constraints as related to mobilizing the community to participate in the screening process are unknown at this stage. Thus, the cost effectiveness as well as participation rate of using this model as a screening tool compared to the clinical method need to be evaluated in the future.

2.5 Conclusions

Recent studies showed that diabetes can be delayed or prevented with lifestyle changes (Knowler et al., 2002; Mensink, 2005; Lindstrom et al., 2006; Li et al., 2008); thus, there is consensus that early detection is beneficial (Gillies et al., 2008; Simmons et al., 2010). US Preventative Service Task Force (USPSTF) also recommended early screening based on systematic review evidence that treatment of Impaired Fasting Glucose (IFG) or impaired Glucose Tolerance (IGT) was associated with delayed progression to diabetes (Selph et al., 2015). However, using the blood glucose screening method is costly, invasive, and inconvenient especially if fasting is required. Therefore, simple, non-invasive, and low-cost methods of screening need to be developed for use in limited resource settings such as community prevention and work-place well-being programs. “Limited resource” in this study is referring to the absence of on-site laboratory tests and clinicians as well as the participant’s lack of resources such as insurance, transportation, and childcare to participate in clinical-based screening services. The lack of insurance indicates lack of regular physician office visits and lack of an electronic health record, which can be used to predict diabetes based on the readily available biomarkers. The goal of this study is to improve the performance of the diabetes prediction model with non-parametric regression by incorporating non-linearity and interactions in the model developing process. The model developed with a non-linear interactions approach was able to improve the performance over the existing US population based diabetes prediction model in a statistically significant way. The model also identified non-linear relationships and two-way interactions. The non-linearity of the model helped to identify cutoff points for some of the variables such as systolic and diastolic blood pressure, depression score, sleep hours, and waist circumferences. In addition, the model adds new, never-used-before information as diabetes prediction markers such as depression score and sleep duration. In a nutshell, the model
was able to squeeze maximum prediction information out of easy measurements and was non-invasive in formation, which can be obtained in limited resource and non-clinical setting. Therefore, the developed risk prediction model can serve as a screening tool in the larger effort to curb the risk of diabetes by encouraging changes in lifestyle. However, the practical application of this model as a screening tool needs further scrutiny such as validation in an independent, and representative sample, and determining cost-effectiveness and acceptability for the screening process.


2.6 References


Arguedas JA, Perez MI, & Wright JM. Treatment blood pressure targets for hypertension. The Cochrane Library (2009).


Constantino MI, Molyneaux L, Limacher-Gisler F, Al-Saeed A et al. Long-Term Complications and Mortality in Young-Onset Diabetes Type 2 diabetes is more hazardous and lethal than T1D. *Diabetes care* 2013; 36(12): 3863-3869.


Kumari M, Head J, Marmot M. Prospective study of social and other risk factors for incidence of type 2 diabetes in the Whitehall II study. Archives of internal medicine 2004; 164(17): 1873.


Moore PA, Zgibor JC, DASANAYAKE AP. Diabetes a growing epidemic of all ages. The Journal of the American Dental Association 2003; 134(suppl 1):11S-15S.


Qiao Q, Nyamdorj R. Is the association of type II diabetes with waist circumference or waist-to-hip ratio stronger than that with body mass index? Eur J Clin Nutr. 2009; 64(1):30-34.


Scheen AJ. Diabetes mellitus in the elderly: insulin resistance and/or impaired insulin secretion? Diabetes Metab 2005; 31:5S27–5S34


Chapter 3: Spatial Spillover of Socio-Ecological Determinants of Diabetes-related Mortality Rate across US Counties

Abstract

The spatial structure in the diabetes-related mortality rate in US counties is well established in previous studies. However, the association between the spatial variation in diabetes-related mortality rate and spatial variation in socio-ecological factors has not been fully addressed. I studied the spatial spillover effect of change in socioeconomic gradients (education, employment, and household income) and retail food environments, and access to healthcare on diabetes-related mortality rates across the United States. Seven-year aggregates of multiple cause mortality data from CDC WONDER’s compressed mortality database was merged with several sources of county-level data to examine mortality clusters, the factors associated with the clusters, and the spatial spillover effects. The results show that high diabetes-related mortality rate clusters are in counties located throughout the Southern Plains, Southeastern, and Appalachian regions of the US. Lower socioeconomic status, high density of fast food restaurants, lack of access to grocery stores, high proportion of African Americans, and low physical activity characterize high diabetes-related mortality rate clusters. Moreover, the impacts from improvement in socioeconomic gradients and the retail food environment in neighboring counties would spill over, which would reduce the diabetes-related mortality rate in a particular county. The implication is that improvement in socioeconomic status and access to healthy food would significantly reduce diabetes-related mortality rates in the contiguous US counties.
3.1 Introduction

Adults with diagnosed diabetes have a greater than twofold increased risk of mortality compared to the general population (Egede, Nietert, & Zheng, 2005). According to the International Diabetes Federation (2014), diabetes was responsible for more than 5 million premature deaths globally in 2014—an 11% increase since 2011—and accounted for 8.4% of all-cause mortality in the age group of 20–70 years. The prevalence of diabetes in the United States increased from 4.5 to 9.3% between 1995 and 2012, and diabetes is the seventh leading cause of death (Murphy, Xu, & Kochanek, 2013). As of 2012, approximately 21 million people were living with diagnosed diabetes and another 8.1 million with undiagnosed diabetes (The Centers for Disease Control and Prevention [CDC], 2014). Despite the improvement in medical technologies and practices, diabetes-related mortality continues to pose challenges due to increasing prevalence of diabetes, especially among minority populations with low socioeconomic status and constrained access to healthcare bearing the worst effects.

Though diabetes-related mortality rate (DRMR) in the United States showed a sign of decline mainly due to medical advances (Saaddine et al., 2006; Murphy, Xu, & Kochanek, 2013), this benefit has not trickled down equally among all racial groups (Murphy, Xu, & Kochanek, 2013) and geographical areas (Murray, 2006). The reason behind the regional disparity in DRMR is not well understood. Most studies on determinants of DRMR disparity pointed to the individual behavior and socioeconomic characteristics (Saydah & Lochner, 2010; Dray-Spira, 2013; Saydah, Imperatore, & Beckles, 2013); however, DRMR is also driven by the socio-ecological factors that are beyond individual control (Weng, Coppini, & Sönksen, 2000; Brown et al., 2004). Socio-ecological factors previously linked to DRMR includes availability of and access to healthy food as measured by density of supermarkets, grocery stores, and convenience stores or fast food restaurants (Alter & Eny, 2005); attributes of physical environments. Attributes of physical environment such as conditions of residential area, conditions of workplaces, highways, accessibility of natural amenities and leisure, and pollution may promote or hinder physical activity (Wilcox et al., 2000; Deshpande et al., 2005; Gordon-Larsen et al., 2006). DRMR is also associated with socioeconomic factors such as median income, income inequality, minimum wage, and availability of jobs (Saydah & Lochner, 2010; Dray-Spira, 2013); proportion of minority populations in the county (McLaughlin & Stokes, 2002), and access to health services (Sommers et al., 2012). For example, Massing et al. (2004)
found that counties’ cardiovascular disease specific mortality rates are associated with income inequality among counties. Further, Shi et al. (2005) showed that counties with high income inequality experienced 11 to 13% higher heart disease and cancer mortality compared to their counterparts with high primary care resources and low inequality, respectively. Thus, in addition to individual health behavior and socioeconomic status, socio-ecological factors surrounding individuals’ lives are important determinants of DRMR as well.

Further, previous studies showed that US DRMR displays palpable spatial structure (i.e., spatial variation, pattern, and dependence) (Voeks et al., 2008). Spatial structure in DRMR may manifest in the form of spatial dependence due to resources’ and health services’ link to geographical areas and spatial interactions between ecological determinants because of social, cultural, economic, and political interactions of populations in neighboring geographic areas. However, this spatial structure has not received much scholarly attention. As a major contributor to mortality in the US (Zimmet, 2003), the spatial structure of DRMR and its relationship with socio-ecological determinants needs special scholarly attention. Understanding the dynamic interactions between a county’s DRMR and socio-ecological factors within the same county and neighboring counties is important in designing policy interventions and resource allocation for DRMR reduction effort. One needs to understand how DRMR in one county affects the DRMR in neighboring counties and how socio-ecological factors in one county affect the DRMR in neighboring counties.

The current study utilizes county-level age and sex standardized DRMR and county-level ecological determinants (socioeconomic, demographic, health services, and built environment) along with spatial statistics and spatial econometric models to disentangle: 1) the presence of spatial dependence in DRMR (if mortality in one county is affected by or affects the mortality rate in neighboring counties); 2) the ecological factors that contribute to the spatial variation in a diabetes-specific mortality rate; and 3) the spillover effect of change in ecological determinants of DRMR. This study hypothesizes that DRMR at the county-level are spatially interdependent, and there is a spillover effect from change in ecological determinants. This is because people move across counties for jobs (one may live in one county while working in another county); hence, they may be exposed to different socioeconomic and physical environment (natural and built) factors in another county. Further, social and institutional culture may diffuse across
counties’ borders influencing peoples’ food choice and culture of physical activity, and the socioeconomics of one county may influence or be influenced by the neighboring counties’ policy (for example, competition between counties for business based on tax reductions and land use policy may determine the type of industry in each county).

County is the spatial unit of analysis for this study. It is the smallest spatial unit for which reliable data on cause-specific mortality is publically available in the US. Data on causes of DRMR at smaller spatial units (for example, census tracts) is suppressed due to concerns about the identification of individuals (confidentiality concern) (O’Carroll et al., 2001). Moreover, counties are sociopolitical and administrative geographic units that provide context within which many social, economic, and public health policies are formulated and implemented (Singh, 2003). Most human service and public health administration is conducted at the county-level.

**Conceptual framework: socio-ecological determinants of diabetes-related mortality**

In this section, the relationship between diabetes specific mortality and various socio-ecological factors is discussed. The purpose of this discussion is to provide a framework for a selection of variables in the analysis as described in the methodology section. The conceptual framework is motivated by the socio-ecological theory. Socio-ecological models are used to explain the complex relationship between socioeconomic (e.g., poverty) and structural factors (e.g., access to care and facilities), individual practices, the physical environment, and health (McLeroy, Bibeau, & Steckler, 1988).

After the onset of diabetes, self-management is essential to prevent complications and mortality. However, self-management requires access to a variety of resources, including healthcare services, nutrition and physical activity related resources, as well as support for the initiation and maintenance of healthy behaviors (Evert et al., 2014). In the context of socio-ecological theory, a range of factors, which may include immediate environments, as well as social, organizational, governmental policy, and economic factors (McLeroy, Bibeau, & Steckler, 1988; Glasgow, 1995; Sallis & Owen, 2002), influence individuals’ health behavior. The degree of influence of these factors on the individuals’ health behavior and access to
resources is mediated by individuals’ geographical locations (e.g., residence and workplaces), ethnicity/race, and socioeconomic positions (Cutler, Lleras-Muney, & Vogl, 2008).

Figure 3.1 describes the conceptual framework of pathways in which policies and environmental factors interact with demographics, socioeconomic gradients, and the geographic location of individuals to determine the health behavior of the individuals and the associated outcomes. The focus of this paper is on the measurable socio-ecological factors that are related to diabetes-related mortality. Empirical evidence linking the elements of the theoretical framework and pathways are discussed below.

**Access to physical activity**

Adequate physical activity is a cornerstone to the management of diabetes and prevention of premature death. A meta-analysis by Sluik et al. (2012) and Kodama et al. (2013) indicates that diabetes patients who were regularly engaged in physical activity had lower mortality risk; even patients who performed moderate amounts of physical activity were at lower risk of premature death compared to those who were totally inactive. Ecological factors such as the built environment are crucial in either facilitating or hindering the efficacy and ability to be physically active. Handy et al. (2002) defined the built environment as integration of land use patterns; the distribution across space of activities and the buildings that house them; the transportation system, the physical infrastructure (such as roads, sidewalk, bike paths, etc.), as well as the service this system provides; and urban design—the arrangement and appearance of the physical elements in a community. The mechanism through which the built environment may affect physical activity level could be classified as those that limit or promote leisure activity (e.g., leisure sport facilities such as swimming pools) and those that limit or promote integration of walking or biking in daily routines (e.g., transportation-related factors such as sidewalk and bike routes). The ability to perform leisure physical activity may be bounded by access to and availability of public and private recreation facilities (Kahn et al., 2002) and green spaces (such as parks) in the vicinity. Integration of transportation-related physical activity into daily routine may be associated with the proximity and characteristics of the pedestrians and cycling infrastructure, including sidewalks, bicycle lanes, and trails (Heath et al., 2006).
Availability and access to healthy food

Diet is an integral part of the continuous treatment and self-management of diabetes. The American Diabetes Association recommends that diabetes patients consume a diet low in fat and high in fiber-containing fruits, vegetables, and whole grains in order to control glycemic level (Funnell et al., 2009). However, adherence to this recommendation is not only influenced by individual choice but also by access and availability of healthy food. Moore et al. (2008) found that individuals who live in areas with a low density of supermarkets are less likely than those who live in areas with high supermarket density to consume healthier diets. On the other hand, Alter and Eny (2005) showed that density of fast food restaurants is associated with higher cardiac-related mortality rates. Recently, Daniel et al. (2010) found that density of fast food restaurants is associated with higher cardiac and general mortality rate.

Access to health services

Access to healthcare as measured by the number and availability of facilities of high-quality primary care providers in a geographic area may reduce negative health outcomes due to lower socioeconomic status (Shi et al., 2003; Brown et al., 2004). In line with this, rural US counties with low density of physicians and hospital beds have limited access to healthcare and have a higher overall mortality rate (Shi et al., 2003). In the case of diabetes, access to healthcare has far more consequences because outcomes need to be followed up with a healthcare provider. In addition to availability, the health outcome of diabetes patients is determined by access to healthcare as measured by health insurance coverage. In a follow-up survey after a new diagnosis of diabetes, Burge et al. (2000) found that 60% of the uninsured newly diagnosed patients were able to obtain medical care, while only 6% of insured patients did not follow up with medical care. Beckles et al. (1998) found that insured diabetic patients tend to have three times more chance of undergoing a dilated eye examination as compared to the uninsured patients.
Income inequality

According to Lynch, et al. (2000), income inequality affects DRMR through psychological processes such as perceptions of social hierarchy based on relative socioeconomic position. Such perceptions may produce negative emotional reactions that may in turn trigger a psycho-neuro-endocrine chemical imbalance and stress-induced behaviors such as overeating and smoking (Lynch et al., 2000). At a macro level such as county, the evidence that income inequality is associated with mortality rate is strong. Meta-analysis by Kondo et al. (2009) showed significant associations between income inequality and health outcomes. Krieger et al. (2008) showed that disparity in mortality among racial and geographical areas in the US fall and rise with temporal income inequality in the population. McLaughlin and Stokes (2001) found that higher income inequality at the county-level was significantly associated with higher overall mortality. Murray et al. (2006) also divided the US population into eight unequal geographic categories based on association between all-cause mortality and socioeconomic inequality (based on race, location of the county of residence, population density, race-specific county-level per capita income, and cumulative homicide rate). According to Murray et al., the “Eight Americas” are Asians (America 1), below-median-income whites living in the Northland (America 2), middle America (America 3), poor whites living in Appalachia and the Mississippi Valley (America 4), Native Americans living on reservations in the West (America 5), African Americans middle-America (America 6), poor African Americans living in the rural South (America 7), and African Americans living in high-risk urban environments (America 8). The health outcome for America 5, 6, 7, and 8 is the worst, where chronic diseases are the major cause of mortality.

Income

As income increases, overall mortality decreases with the same proportion throughout the income distribution (Deaton, 2002). Those deprived of income are the most affected by illness and mortality (Deaton, 2002). The mechanism through which income deprivation affects health is that it limits the access to quality healthcare and drives people to risky health behavior. Deaton and Paxson (2001) used the National Longitudinal Mortality Study (NLMS) cohort (ages 25–59)
to find that doubling people’s income reduced the probability of death by 27% during the first year of follow up, 25% during the second year, 23% in five years, and 17% in nine years, after controlling for years of schooling. However, Deaton and Paxson (2001) emphasized income inequality over absolute income itself as a source of psychological distress that induces ill health and premature mortality. In reference to DRMR, the rate could increase due to economic deprivation that may limit people’s ability to adhere to lifestyle (diet and physical activity) and medication recommendations.

Demographic composition

Disparity in overall mortality among ethnic groups remains an issue in the US (Hoyert, 2012). However, the underlying cause for this disparity is unknown, and so is whether socioeconomic positions of the ethnic groups or ethnic group composition by itself is an underlying cause of disparity in mortality rate. McLaughlin and Stokes (2001) found that higher minority racial concentration interacted with income inequality to produce higher mortality rate in the counties. Considering that diabetes prevalence is higher among minorities such as African Americans and Hispanics (Carter, Pugh, & Monterrosa, 1996), the mortality rate from diabetes may also be higher in counties with a high concentration of these ethnic groups.

Racial segregation

Separation of ethnic/racial groups into residential areas is one distinct characteristic of the social organization of urban areas in the United States (Acevedo-Garcia et al., 2003). According to Williams and Collins (2001), racial residential segregation is a fundamental source of racial disparities in health outcomes. Enforced physical separation of races through norms that limit interactions between races and institutional racism (such institutions as mortgage markets) may contribute to the socioeconomic disparity between ethnic groups, which is a root cause for health outcome disparity (Williams & Collins, 2001). Segregation plays a major role in shaping access, utilization, and quality of healthcare services across the whole range of clinical care, which has severe consequences for the long-term health of individuals and communities (White, Haas, &
Ample empirical literature shows that all-cause mortality in the African American community is positively associated with residential segregation and with residential areas that are predominantly African American (Acevedo-Garcia et al., 2003). Thus, counties with higher racial segregation might display higher overall mortality and DRMR.

Research questions

1. Is there spatial dependence (autocorrelation) among counties in the diabetes-related mortality rate?

2. Is the variation in county diabetes-related mortality rate explained by the variation in socio-ecological factors (demographic, socioeconomic, and built environment)?

3. Are there spatial spillover (multiplier) effects from change in socio-ecological factors (demographic, socioeconomic, and built environment) in a county to diabetes-related mortality rates in other counties?

Objectives

1. To identify spatial dependence and variation in diabetes-related mortality rates.

2. To determine whether geographic variation in diabetes mortality is explained by socio-ecological factors (demographic, socioeconomic, and built environment).

3. To investigate whether there are spatial spillover effects from socio-ecological factors (demographic, socioeconomic, and built environment) that are associated with diabetes-related mortality rates.

Hypotheses

1. There is spatial dependence between the counties’ diabetes-related mortality rates.

2. Counties with high diabetes-related mortality rates have a high concentration of minorities (African Americans and Hispanics), higher level of racial segregation, socioeconomic gradients (high unemployment rate, low median household income, and
the proportion of college graduates), high income inequality, unfavorable built environment (high density of fast food restaurants, low density of department stores, and low density physical activity facilities), and health services.

3. There are spatial spillover effects from changes in a high concentration of minorities (African Americans and Hispanics), higher level of racial segregation, socioeconomic gradients (high unemployment rate, low median household income, and the proportion of college graduates), high income inequality, unfavorable built environment (high density fast food, low density of department stores, and low density physical activity facilities), and health services.

3.2 Methods

Theoretical framework for spatial dependence and spillover

Inherent to the health data, this paper assumes that there is a spatial structure involving both dependent and independent variables. This section provides the theoretical motivation for the choice of the statistical model that will be presented in the following section. We try to embed our assumption of interrelatedness of observations into the theory of spatial dependence.

The idea of spatial dependence sprung from Tobler’s (1970) first law of geography, which states: “Everything is related to everything else, but near things are more related than distant things.” This statement implies not only geographical proximity and bordering of spatial entities but also movement and interactions between them (Miller, 2004), which include diffusion of disease, innovation, physical resources, knowledge, and policy; and interaction among people and institutions at various geographical scales. Considering the first law of geography, spatial dependence exists not only when an observation in one location is correlated with an observation in another location but also when nearby locations have more similar patterns of observation than locations that are farther apart for a given variable (Haining, 2003). Spatial regression models explicitly account for this interrelatedness of data (observations) based on spatial proximity (Haining, 2003).
The source of spatial dependence may vary depending on the type of data at hand. This paper is concerned with spatial analysis of mortality and various socio-ecological data spanning social and economic policies, social and physical environment, and knowledge and technology related to diabetes. We believe that social, health, and economic policies, as well as knowledge and innovations with regard to health problems, are the underlying cause of mortalities; hence, diffusion and spillover of the socio-ecological factors, policies, and innovations across spatial units are responsible for spatial dependence of mortality.

Three common mechanisms through which socio-ecological characteristics and socioeconomic policies, cultures, and innovations may diffuse across geographical areas can be noted. First, socio-ecological characteristics and socioeconomic policies, cultures, and innovations may spread to neighboring administrative geographical units because of the preferences, norms, and values of the residents in neighboring administrative geographical units (Elkins & Simmons, 2005). This is known as social contagion theory in sociology, where it refers to the spread of thoughts, ideas, or behaviors from one individual to an entire group of people (Coleman, Katz, & Menzel, 1966). Social contagion theory emphasizes learning processes where residents of a neighboring county (state) learn the outcomes of policy, socio-ecological characteristics, culture, and innovations in another neighboring county (state), and eventually adapt, adopt, or reject similar policy, socio-ecological characteristics, culture, and innovations in their state (or county). The social contagion occurs because people move across counties for jobs (one may live in one county while working in another); hence, they may be exposed to socioeconomic, culture, and physical environment (natural and built) factors in another county. This learning process is not limited to individuals but also involves institutions and businesses. Second, economic characteristics, innovations, and policies may diffuse due to competition between neighboring geographical units. Geographical units may adopt or adapt similar or counter economic characteristics, culture, and policy in response to the economic characteristics, culture, and policy of the neighboring geographical units (Elkins & Simmons, 2005). Third, counties may adopt social and economic policies as dictated by states. This mechanism is referred to as coercion in international policy analysis (Elkins & Simmons, 2005). The environmental, socioeconomic, and cultural policies adopted by the state are also the policy
of the counties in most cases. The diffusion of policy has been empirically tested with cigarette smoking regulation at the state level (Pacheco, 2012).

Measuring spatial dependence

Exploratory spatial dependence analysis

At the exploratory stage of the analysis, global Moran’s $I$ and Local Indicators of Spatial Association (LISA) were used to detect global spatial dependence between county-level diabetes-related mortality rates and to detect significant spatial clusters or outliers for each county.

Global Moran’s $I$

Global Moran’s $I$ is a correlation coefficient weighted by spatial structure calculated for $N$ spatial unit (county) on a variable at locations $i,j$ as:

$$I = \frac{N}{\sum_{i} \sum_{j} w_{ij}} \frac{\sum_{i} \sum_{j} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i} (x_i - \bar{x})}$$

(1),

where $\bar{x}$ is the mean of the variable $x$, and $w_{ij}$ represents the elements of the queen contiguity weight matrix.

Moran’s $I$ statistical expectation is:

$$E(I) = -\frac{1}{N-1}$$

(2)

and variance is:

$$V(I) = \frac{NS_3 - S_3 S_4 (1-2N)}{(N-1)(N-2)(N-3)(\sum_{i} \sum_{j} w_{ij})^2}$$

(3)

where:

$$S_1 = \frac{1}{2} \sum_{i} \sum_{j} (w_{ij} + w_{ji})^2$$
\[ S_2 = \sum_i (\sum_j w_{ij} + \sum_j w_{ji})^2 \]

\[ S_3 = \frac{N^{-1} \sum_i (x_i - \bar{x})^4}{(N^{-1} \sum_i (x_i - \bar{x})^2)^2} \]

\[ S_4 = (N^2 - 3N + 3)S_1 - NS_2 + 3(\sum_i \sum_j w_{ij})^2 \]

Under the null hypothesis of no spatial autocorrelation, Moran’s I statistics are assumed to behave asymptotically normal. Then, the output statistics is:

\[ I^* = \frac{I - E(I)}{\sqrt{V(I)}} \]  

(4)

However, typically Moran’s I does not follow a normal distribution. Alternatively, Monte Carlo (random) permutation of test statistics under a null hypothesis is generated and used for significance testing. The Monte Carlo permutation method randomly rearranges the values among the geographic units and calculates I each time (e.g., 999 times) and compares the observed I with the 999 randomly generated Is. If the observed I falls into an area of either more than 95% or less than 5%, it is said that I is pseudo-significant at a 5% level (positive/negative).

The pseudo significance (p-value) from the Monte Carlo permutation is computed as \((M + 1) / (R + 1)\), where R is the number of replications and M is the number of instances where computed Is from the permutations is equal to or greater than the observed I value (for positive local Moran) or less or equal to the observed value (for negative local Moran) (Anselin, 2005). For instance, if an observed Moran’s I value is higher than any of the randomly generated Moran’s I values, the pseudo p-value would be 1/100=0.01 for 99 permutations or 1/1,000=0.001 for 999 permutations.

**Local Indicators of Spatial Association (LISA)**

Local Indicators of Spatial Association (LISA), developed by Anselin (1995), measure local, location-specific spatial correlation. This measure enables us to assess the significance of
local spatial clustering around an individual county. LISA identifies the degree and location of spatial clusters in detail. The local version of Moran’s $I$ at county $i$ is given as:

$$I_i = \frac{\sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{s_i^2 \sum_j (x_j - \bar{x})^2}$$

where:

$$s_i^2 = \frac{\sum_j w_{ij}}{n-1} - \bar{x}^2$$

The procedure for hypothesis testing for LISA is the same as global Moran’s $I$ above. The interpretation of LISA results, however, is specific to an individual’s location. In the LISA analysis, if the test statistic is not significant the conclusion is that the data shows no local spatial association at location $i$. When LISA statistics show statistical significance, however, the results fall into four quadrants with two possible patterns of local spatial association. When county $i$ and its neighbors display higher than average values, a High–High (HH) association (also known as hot spot) is indicated; and when county $i$ and its neighbors display lower than average values, the spatial tendency is ruled as a Low–Low (LL) association (also known as cold spot). In addition, locations with High–Low (HL) and Low–High (LH) spatial association are also identified. To make spatial autocorrelation visually interpretable, the local Moran’s $I$ is represented by cluster maps, in which the locations of statistically significant spatial clusters are highlighted to identify the patterns of associations.

**Spatial econometric models**

Following Elhorst (2010), we start with the non-spatial Ordinary Least Square (OLS) model and proceed with specifying various spatial extensions. In this model, we assume that the diabetes-related mortality rates (DRMR) at county $i$ (where $i = 1, ..., N$) is a function of county-level ecological characteristics:

$$Y = \alpha l_N + X\beta + u$$

where $Y$ is a $N \times 1$ vector of the dependent variable (which is diabetes-related mortality), $X$ is $N \times K$ matrix of explanatory variables (county-level ecological variables), and $u$ is a $N \times$
1 vector of independently and identically distributed error term for all \( i \) with zero mean and variance \( \sigma^2 \).

As we discussed earlier, due to the spatial nature of our data we anticipate spatial dependence. The spatial phenomena could arise from three situations:

1) *Endogenous spatial interaction effect*, where the diabetes-related mortality rate (dependent variable) in one county is correlated with diabetes-related mortality rate in another county. The endogenous spatial dependence for diabetes-related mortality rate may arise from, for instance, shared primary care services across counties or social network and cultural diffusion, which are partly formed based on spatial proximity, or by shared local environmental characteristics among individuals living in proximate counties. In the presence of spatial dependence, the Best Linear Unbiased Estimator (BLUE) properties of OLS estimation are violated and, thus, produce biased and inconsistent estimates (LeSage & Pace, 2004; Anselin, 1988).

2) *Exogenous spatial interaction*, where socio-ecological factors (independent variables) in one county depend on the socio-ecological factors (independent variables) in neighboring counties. Manski (1993) stated that economic agents’ (i.e., businesses such as healthcare providers, supermarkets, etc.) behavior is affected by the characteristics and behaviors of other economic agents in the vicinity. Such interaction between economic agents produces spillover effects (Anselin, 2002). Instances of such interactions includes when the decision made about food supply chain in one county affects the food supply chain in neighboring counties, which has implications on the availability and access to healthy food. Similarly, spatial agglomerations and competition of health providers (such as hospitals) may affect healthcare accessibility and health outcomes across counties, and unemployment rate and the availability of high-wage jobs in one county affects the economic situation in the neighboring counties.

3) *Interaction among error terms*, where error terms are spatially auto-correlated (if omitted variables or unobserved shock to the dependent variables followed a spatial pattern).

Extending equation (6) to a general model to allow for spatial lag in both dependent and independent variables (ecological variables) and for spatial error, we obtain the following notation:
\[ Y = \rho WY + \alpha l_N + X\beta + WX\theta + u \quad (7) \]

\[ u = \lambda Wu + \varepsilon \]

where \( W \) is an \((n \times n)\) standardized matrix defining who is a neighbor of whom by means of 0 to 1 values, and \( \varepsilon \) an error assumed to be independently and identically distributed. \( \alpha \) is county-specific fixed effects. The spatial weight matrix is usually standardized so that the sum of each row equals 1, which implies that the spatially lagged version of \( y \) contains the average \( y \) value of the neighbors.

Based on equation (7), we can specify and test various spatial models that may explain the spatial dependence in diabetes-related mortality rates and ecological determinates. Setting \( \theta = 0 \) and \( \lambda = 0 \), the model becomes a spatial auto regressive (SAR) model:

\[ Y = \rho WY + \alpha l_N + X\beta + \varepsilon \quad (8) \]

and setting \( \rho = 0 \) and \( \theta = 0 \), the model becomes a spatial error model (SEM):

\[ Y = \alpha l_N + X\beta + u \quad (9) \]

\[ u = \lambda Wu + \varepsilon \]

Whether the diabetes-related mortality data exhibits SAR or SEM can be tested based on the Anselin (1996) robust Lagrange Multiplier tests.

SAR and SEM do not capture the global and local interactions (spillover effect) (Elhorst, 2010), weakening their expounding power of the spatial process at hand. Further, we are also interested in the interaction between explanatory variables; apparently SAR and SEM do not capture this phenomenon. The spatial Durbin model overcomes the aforementioned limitations. By setting \( \lambda = 0 \), we can obtain:

\[ Y = \rho WY + \alpha l_N + X\beta + WX\theta + \varepsilon \quad (10) \]
Choice of spatial weights matrix

The spatial weights matrix $w$ is an $n \times n$ non-negative matrix that specifies which county is a neighbor with which county. A list of counties appears as both rows and columns in the matrix. The non-zero matrix elements of $w_{ij}$ indicate that county $i$ (row) and county $j$ (column) are neighbors. The main diagonal of the matrix represents self-neighbors, which are excluded by convention. Therefore, the diagonal elements of $w$ are set to zero ($w_{ii} = 0$). Spatial weight matrixes are often row standardized ($w_{ij}^s = \frac{w_{ij}}{\sum_j w_{ij}}$, thus, $\sum_j w_{ij}^s = 1$) to facilitate the interpretation of weighted averaging of the spatial effect of neighbors later in the spatial correlation and regression models (Anselin, 1988).

The choice of spatial weights matrix is crucial in specifying spatial regression models. However, there is no specific guideline as to what spatial weights matrix is appropriate for a given empirical analysis (Anselin, 2002, p. 289). Nevertheless, rules of thumb have been suggested for empirical specification of the weights matrix (Griffith, 1996, p. 80; Getis & Aldstadt, 2004); these focus on heterogeneity in spatial unit size and the sample size of spatial units, definition of neighbors, and use of lower order and higher order neighbors definition. Based on the law of large numbers, it is suggested that a relatively large number ($N > 60$) of spatial units be used to overcome the influence of the due to unequal size spatial units. This study, with 3,109 spatial units, exceeds the suggested threshold of spatial units.

The type of spatial weights matrix, which defines what is considered neighbors, is largely determined by the empirical research problem at hand. This study used a queen contiguity spatial weights matrix, where spatial units are neighbors if they share boundaries and vertexes. Because the socio-ecological interactions between spatial units are two-way interactions, this study chose to use a queen contiguity based spatial weights matrix. Further, one needs to decide whether to define neighbors based on the traditional sense of geographical structure sharing borders (first order contiguity) or also need to include neighbors of neighbors (higher order contiguity). In this regard, it was suggested that specifying first order and second order contiguity neighbors is more robust than specifying higher order neighbors, which may lead to reduced power during
estimation. This study considers this suggestion along with a comparison among models with a various orders weights matrix using Akaki Information Criteria (AIC).

**Interpretations of spatial lags parameters: the direct and indirect impacts**

The difference between equation (6) and (7) is that equation (7) involves lags of dependent variables, independent variables, and error terms. The presence of these lags in the model alters the interpretation of the model parameters into two of the sub models. The interpretation of parameters in the Spatial Autoregressive (SAR) and Spatial Durbin models are different from the least square regression parameters interpretation if $\rho \neq 0$ and $\theta \neq 0$ (Pace & LeSage, 2006). This is, however, not the case for the Spatial Error Model (SEM). The resultant change to the interpretation of the parameter specifically substantiates the Spatial Durbin model by allowing differentiation of direct and indirect impact of the independent variables on the dependent variable by including information from the neighboring spatial units (LeSage & Pace, 2009; Fischer, 2010).

For explicit illustration of the impact of this deviation, let us consider rewriting the Spatial Durbin model in equation (10) as follows:

\[
(I_{N} - \rho W)Y = \alpha l_{N} + X\beta + WX\theta + \epsilon \tag{11}
\]

\[
Y = (I_{N} - \rho W)^{-1}\alpha l_{N} + (I_{N} - \rho W)^{-1}X\beta + (I_{N} - \rho W)^{-1}WX\theta + (I_{N} - \rho W)^{-1}\epsilon \tag{12}
\]

The matrix of partial derivatives of $Y$ with respect to the $k^{th}$ explanatory variable in $X$ for $N$ counties (leaving out the spatial unit subscript) can be derived as follows:

\[
\frac{\partial Y}{\partial x_k} = (I_{N} - \rho W)^{-1}(I_{N} \beta_k + W \theta_k) \tag{13}
\]
where $\frac{\partial Y}{\partial x_k}$ is an $N \times N$ matrix, $(I_N \beta_k + W \theta_k)$ is an $N \times N$ matrix of independent and lagged explanatory variable parameters ($\beta_k$ consists of the parameters for the $k$ independent variables, and $W \theta_k$ comprises weighted parameters for lagged independent variable, which are off diagonal elements of the matrix), and $(I_N - \rho W)^{-1}$ is an $N \times N$ spatial multiplier matrix, which can be expanded as an infinite series: $(I_N - \rho W)^{-1} = I_N + \rho W + \rho^2 W^2 + \cdots$, with Debreu and Herstein (1953) (LaSage, 2008). The power on the infinite series represents the order of neighborhood relationship (i.e., $W$ represents a first order neighbor, $W^2$ represents a second order neighbor, etc.).

The result in equation (13) reveals that the partial derivatives of the Spatial Durbin model possess three interpretations (Elhorst, 2009). First, if a particular explanatory variable in a particular spatial unit (county) changes, it will not only affect the dependent variable in that county but also the dependent variables in other neighboring counties (neighbor as defined by weights matrix) (Elhorst, 2009). The effect of change in an explanatory variable in a county on the dependent variable of the county itself is called a direct effect. The effect of change in explanatory variable in a county on the dependent variables of other neighboring counties is known as an indirect effect. The direct effect is accounted for by the main diagonal elements, and the indirect effect is accounted for by off-diagonal elements of the partial derivative matrix $(I_N - \rho W)^{-1} (I_N \beta_k + W \theta_k)$.

Second, the direct effects and the indirect effects vary from county to county (Elhorst, 2009). The direct effects vary because the diagonal elements of the matrix, $(I_N - \rho W)^{-1}$, are different for different counties and conditional on $\rho \neq 0$. The indirect effects are different because both the non-diagonal elements of the matrix, $(I_N - \rho W)^{-1}$, and the spatial weights matrix, $W$, are different for different counties, conditional on $\rho \neq 0$ and/or $\theta_k \neq 0$. For $k$ independent variables and $N$ counties, the numbers of estimated parameters are $k$ different $N \times N$ matrixes. It is easy to see that the amount of estimated parameters is unmanageable even for small number of variables and spatial units. Pace and LeSage (2006) showed that an average of these spatial effects across all spatial units can be summarized with a scalar measure. Thus, the spatial effect is reported as: average direct effect, the average (over $N$ counties) impact of change in $k^{th}$ explanatory variable in county $i$ on the county $i$ dependent variable while
accounting for the feedback system of the impacts passing through neighboring counties and back to county $i$; *average indirect effect*, the average (over $N$ counties) impact of change in $k^{th}$ explanatory variable in county $j$ on the dependent variable in county $i$; and *average total effect*, the average (over $N$ counties) impact of change in $k^{th}$ explanatory variable in all county $j(N−1$ counties) on the dependent variable in county $i$. Finally, the hypothesis test (t-statistics and p-value for the average effect parameters is obtained through the Monte Carlo simulation (Bivand et al., 2013).

Third, the indirect effects can be differentiated into local effects and global effects (Elhorst, 2009). The indirect effects that occur given $θ_k ≠ 0$ are known as local effects; they arise only from a county neighborhood set (as defined by $W_{ij}$) matrix. The indirect effects that occur given $ρ ≠ 0$ is referred to global effects; they arise from units that do not belong to a county neighborhood set because matrix $(I_N − ρW)^{-1}$ does not contain zero elements. If both $θ_k ≠ 0$ and $ρ ≠ 0$, then the indirect effect is not separable.

**Data sources**

Data for this analysis is obtained from various sources as described in Table B.1. The county-level mortality rate in the contiguous US is the dependent variable of this study. Multiple cause mortality data was extracted from the CDC WONDER compressed mortality database (CDC WONDER, 2013).

The independent variables address density of health services and facilities, socioeconomic status, demographic composition, ethnic segregation, income inequality, density food markets, social capital, and density of physical activity facilities. The data were compiled and derived from several sources: health service variables data were derived Health Resources and Services Administration (HRSA); food service and supermarket variables were derived from US Census County Business Patterns; and ethnic segregation, income inequality, and social capital variables were extracted from the Research Triangle Institute (RTI) international database.
Dependent variable

Diabetes-related mortality rates were standardized based on 2000 US census age–sex population structure. Diabetes-related mortality rate is defined as multiple cause of death where diabetes is either directly the cause of death or contributed to the death. The mortality rate was measured as the total number of deaths per 100,000 populations in a county. Diabetes-related mortality rates were aggregate rates for year 2003 to 2010. It is common practice to age standardize mortality rates, as it is suitable for ecological mortality research (Kawachi & Blakely, 2001), and to compare data across space with two populations that have different demographic structures (Preston et al., 2001; Yang, Noah, & Shoff, 2013). It removes the effect of differences in age (or other confounding variables that affect mortality rate) between the populations. Aggregate mortality rates from eight years were used to reduce the number of counties with missing mortality data. Even using seven years of aggregated data, the mortality rates are missing for 417 counties (187 are suppressed because they have less than 10 deaths, and 230 are flagged as unreliable), which is about 13% of the total sample. Data are suppressed due to confidentiality concerns or if the rate is not reliable because of small numbers of deaths (if mortality is less than 20, it is flagged as unreliable). The suppressed and flagged data fundamentally created right-censored data, which would result in biased inferential statistics if the appropriate method were used in statistical analysis. In spatial econometrics analysis, tools that are appropriate for censored data have not developed yet. The next best option is to impute the missing values. Therefore, for 230 counties where the mortality rate is flagged as unreliable, the number of deaths were used as reported but the crude rate was imputed based on the reported number of death and other county characteristics (proportion with college education, proportion of unemployed, proportion of minorities, and urbanization index of the county). The purpose of the imputation was to reduce the error related to the unreliability of the data. The suppressed 187 deaths were imputed using an interval regression technique, which limits the outcome to a censored range (0 to 10). Then, crude mortality rates were calculated. Finally, the crude mortality rates were age-standardized using the direct age standardization method and the 2000 US Census population. In both cases, the imputation was conducted 20 times, and an average of the multiple imputations was calculated. The multiple imputations corrects for the biases since single-imputation inference tends to overstate precision because it omits the between imputation
component of variability (Little & Rubin, 1987). The death rate was log transformed to limit the influence of outliers.

**Independent variables**

*Healthcare service:* Health service variables were extracted from the Health Resources and Services Administration (HRSA) Area Health Resource Files (AHRF, 2012-13), a database maintained by the US Department of Health and Human Services. From the AHRF data, three variables were created to describe the healthcare infrastructure and were included in the analysis: proportion of the population aged 18 to 65 years without health insurance, total number of active medical doctors, total number of non-federal primary care physicians, total number of hospital beds, and total number of hospitals. The average of eight years (2003–2010) of data was calculated for all three variables to smooth out variation over years. The values of the total number of active medical doctors and total number of hospitals is expressed as per 10,000 of population, based on the inter-census population estimates.

*Ethnic segregation:* Ethnic segregation was represented by the Index of Spatial Proximity (ISP), which measures the extent to which subgroup populations (whites, African Americans, and Hispanics) are clustered together in adjacent tracts, suggesting the presence of a cultural enclave (Massey & Denton, 1988). The data is extracted from the Research Triangle Institute (RTI) international database. This variable is only available for the years 1990 and 2000. We used year 2000, which is closer to other data years. The value of the variable ranged from 0 to 2. An SPI index of 1 indicates that there is no differential clustering between minority and majority populations, a value greater than 1 indicates that the majority and minority members live spatially closer to own group, and a value less than 1 indicates members of each group live closer to each other’s group than to members of their own group (Massey & Denton, 1988).

*Ethnic composition:* Two variables representing ethnic composition are included in the analysis: the proportion of African Americans and the proportion of Hispanics were included in the analysis as independent variables to represent compositions of the two minority groups in the
counties. The proportion of the non-Hispanic white population was excluded to avoid a multicollinearity problem that may arise in the analysis. These variables were extracted from the American Community Survey (ACS) five-year (2005 to 2010) estimates (US Census Bureau, 2010).

**Socioeconomic factors**: Socioeconomic status is a multidimensional gradient; thus, it was measured as a composite score of several variables (Carpiano, Lloyd, & Hertzman, 2009). Factor analysis was used to derive a socioeconomic factor score from four variables. The variables included in the factor analysis are a percentage of the population aged 25 years and older with at least a bachelor’s degree (factor loading = 0.71), unemployment rate (factor loading = –0.52), percentage of population under poverty (factor loading = –0.83), and household median income (factor loading = 0.89). The socioeconomic factor explains 97% of the variation (with Eigenvalue = 2.25). Regression method was used to predict scores for the socioeconomic factors. The data was extracted from Health Resources and Services Administration (HRSA) Area Health Resource Files (AHRF, 2012-13), and average values (for year 2003 to 2010 estimates) were pooled for unemployment rate, percentage of the population under poverty, and household median income. For the percentage of the population aged 25 years and older with at least a bachelor’s degree, data was extracted from American Community Survey (ACS) five-year (2005 to 2010) estimates (US Census Bureau, 2010).

**Income inequality**: Income inequality was measured with the Gini index. The Gini index varies between zero and one. A value of one indicates perfect inequality, where only one household has any income. A value of zero indicates perfect equality, where all households have equal income. Data was extracted from American Community Survey (ACS) five-year (2005 to 2010) estimates (US Census Bureau, 2010).

**Retail food environment**: The retail food environment is depicted by density of grocery stores and supermarkets and density of full-service restaurants and limited-service restaurants (fast food restaurants). The number of grocery stores and supermarkets, number of full-service restaurants, number of limited-service restaurants for years 2003–2010 were extracted from county business pattern data. The numbers of business establishments (grocery stores and supermarkets, number of full-service restaurants, and number of limited-service restaurants) in
the counties were divided by inter-census population estimates for respective years to obtain density. An eight-year average of the number of grocery stores and supermarkets, full-service restaurants, and limited-service restaurants per thousand residents were calculated. Finally, a modified version of the Retail Food Environment Index (RFEI) from the UCLA Center for Health Policy Research (2008) was constructed as follows:

\[
RFEI = \frac{\text{#fastfood restaurants per 1000}}{\text{#supermarkets per 1000} + \text{# full service restaurants per 1000}}
\]

**Food accessibility:** This is represented by the percentage of households that have no vehicle and reside more than 1 mile (10 miles for rural residents) from the nearest grocery store or supermarket. The accessibility variable is extracted from Economic Research Service (ERS) food environment atlas data (2012). This variable is available only for 2010. Though the availability of infrastructure such as grocery store density and real state may not change by much over eight years, care is needed when one-year data is regressed against eight-year aggregate mortality data.

**Social capital index (SCI):** The social capital index was obtained from Rupasingha and Goetz (2008), which constructed a social capital index based on the population weighted number of religious organizations, civic and social associations, business associations, political organizations, professional organizations, labor organizations, bowling centers, physical fitness facilities, public golf courses, sport clubs, managers, and promoters in the counties. Rupasingha and Goetz (2008) used a principal component analysis method to construct the index for years 2005 and 2009. This study utilizes the average of the year 2005 and 2009 index to overlap with the aggregated mortality rate of the years 2003–2010.

**Percentage of adult meeting physical activity guidelines:** The proportion of the sample population who meet the physical activity guideline is constructed by ERS based on the 2010 Behavioral Risk Factor Surveillance Survey (BRFSS) response data. This variable is extracted from Economic Research Service (ERS) food environment atlas data (2012).
Rural–Urban continuum: The extent of urbanization was measured with the percentage of the population living in metropolitan and micro-metropolitan areas. This variable was calculated based on the US census data from 2000 and 2010. In addition, the rural–urban continuum code developed in 2003 by US Department of Agriculture was extracted from an area resource file. The code ranges from one to nine, with decreasing urbanization as the code increases (one being most metropolitan and nine being most rural).

3.3 Results
Descriptive statistics

The result of descriptive statistical analysis is summarized in Table 3.1. The diabetes-related mortality rate ranges from 5.84 to 229 per a population of 100,000, with a mean of 60.62 based on 3,109 counties. The indicator for county segregation, Spatial Proximity Index (SPI), ranges from zero to 1.68 with a mean of 1.06. This is an indicator that the segregation levels in the contiguous US counties vary widely, but on average the segregation level is close to neutral. As discussed previously, an SPI greater than 1 is an indication for residential segregation of ethnic groups. That is, neighborhoods with homogenous ethnic compositions live in proximate distance to a neighborhood with similar ethnic composition as compared with neighborhoods with different ethnic composition. The composition of minority ethnic groups of US counties also varies widely. Specifically, the Hispanic population ranges from zero to 99% of the total county population with an average of about 8%, and the Black population ranges from zero to 87% with the average being about 9%. Counties with a high proportion of Hispanic residents are in the Southern and Southwestern states bordering Mexico, while counties with a high proportion of Black residents tend to be located in Southern and Southeastern states.

The socioeconomic index varied widely between counties (ranging from –2.90 to 4.62 with a mean of zero), indicating socioeconomic status disparity among contiguous US counties. The average income disparity also varied widely, with counties with a lower socioeconomic index having a higher income disparity. The Gini index ranged from 0.21 to 0.65, with higher income disparity and a lower socioeconomic index clustered in Southern regions (Southwest and
Southeast regions), where the proportions of Hispanic and African American populations are higher.

Medical facilities and personnel were also measured with per capita hospital number and per capita active medical doctors in the county. The average number of hospitals per 1,000 county residents is 0.58. Density of hospitals ranges from none to 8.86 per 1,000 residents; where the highest hospital density is in the Northern mountain area, Northern central states, and some Southwest states, and the lowest density is in the some Midwest, South-Atlantic, and Mid-Atlantic states. Similarly, the average number of active medical doctors per 10,000 county residents is 1.25 with a very wide variation (ranging from zero to 26.94). The highest concentrations are in the New England region, and the lowest in the Midwest and Southwest parts of US. The average proportion of uninsured residents in a county is about 64% (ranging from zero to 73%). Counties with the highest uninsured proportion of residents are clustered in the Southwest and South-Atlantic parts of the US, while the lowest is clustered around the Northeastern Midwest and New England states of the US.

The Social Capital Index (SCI) is a composite measure for the social infrastructure in counties. The SCI was constructed based on the work of Putnam (2000) and Rupasingha and Goetz (2008). It comprises density of religious, labor, professional, and political organizations, sport, and recreational facilities in the county. The SCI ranged from −3.86 to 15.97 (with a zero mean). The highest SCI range is mostly concentrated in the Northern Central states, the Northern part of the mountain states, and New England states. The proportion of adult residents who meet physical activity guidelines ranges from about 7% to about 51%; those counties with a high proportion of meeting the guidelines are clustered around mountainous, North-Pacific, and New England states.

Physical food access and availability was represented by two variables. Retail Food Environment Index (RFEI) is a measure of physical healthy food availability. RFEI ranged from none to 2.29 (with a mean of 0.63)—the higher value representing an unhealthy food environment. The highest values are mostly prevalent in the South and Midwest contiguous US states. The physical access dimension is measured by the percent of residents who reside 1 mile (10 for rural residents) from supermarkets and have no autonomous access to a vehicle. It ranged
from zero to 29.51%; the highest values are clustered in the Southern contiguous US states (similar to RFEI).

**Exploratory spatial data analysis results**

**Global Moran’s I**

The overall spatial autocorrelation of diabetes-related mortality rate (DRMR) among contiguous US counties was analyzed with Global Moran’s I statistics. Moran’s I statistics were computed using a logarithmic transformed mortality rate along with a raw standardized six-order queen spatial weights matrix. Even after logarithmic transformation, our data violates normality assumption. Thus, the permutation and Monte Carlo method was implemented to obtain the pseudo p-value. The null hypothesis is there is no spatial overall pattern in the county’s DRMR. The alternative hypothesis is there is a positive spatial correlation between the county’s DRMR. The result from Global Moran’s I ($I = 0.129$, $SD = 0.004$, $P$-value = 0.001) is that there is a small but positive and statistically significant spatial dependence among the county’s DRMR. Thus, it is concluded that county diabetes-related mortality rates are spatially associated.

**Univariate Local Indicator for Spatial Association (LISA)**

Once we know that counties’ diabetes-related mortality rates (DRMR) are spatially correlated, we can identify the local clustering and heterogeneity in spatial dependence. Clusters of counties’ mortality rates were discovered using LISA (also referred to as Local Moran’s $I$). The Monte Carlo simulation (with 999 permutations) was used to generate the statistical significance test value (pseudo p-value). The statistical significance level for this analysis was 0.05 (any correlation with p-value less than 0.05 is considered statistically significant). The cluster analysis was performed in GeoDa, and the map was plotted in Esri’s ArcMap software programs. The result is presented as a cluster map in Figure 3.2.

Accordingly, the map is classified into five categories (four quadrant significant correlation categories and one insignificant category). Four hundred and one counties are hot
spot (High–High) clusters, with high mortality rate counties associated with other high mortality rate counties. Three hundred and fourteen counties were identified as cold spot (Low–Low) clusters, with low mortality rate counties associated with other low mortality rate counties. Sixty-seven and 88 counties are identified as Low–High and High–Low clusters, respectively. The remaining 2,235 counties’ mortality rates did not show any statistically significant cluster. DRMR exhibited High–High clusters throughout the Southern Plains, Southeastern, and Appalachian regions of the US. The Low–Low clusters were observed in the Western mountain regions of the US, in some parts of the Midwest, Florida, and New England.

Note that the spatial clusters labeled in the LISA cluster map are limited to the core of the clusters; it does not extend to the neighboring counties with similar value. That is, the cluster is classified as such when the value at a location (either high or low) is more similar to its neighbors (as summarized by averaging the neighboring values based on the spatial weights matrix) than would be the case under spatial randomness. Counties with such characteristics are color-labeled on the cluster map, but the neighboring counties with similar characteristics are not color-labeled as statistically significant clusters.

**Global bivariate Moran’s I**

Global bivariate Moran’s I is a Pearson correlation type of measure that accounts for the spatial nature of the data. Global bivariate Moran’s I is employed to identify socioeconomic, demographic, and environmental variables that are associated with a county’s diabetes-related mortality. The Global bivariate Moran’s I statistics (and its significance), for association between DRMR and neighboring counties’ explanatory variables, are provided in Table 3.2. The statistics show that there is spatial correlation between DRMR in one county and values of explanatory variables in neighboring counties (neighbor as defined by spatial matrix). Level of residential segregation, income disparity, high proportions of African American and Hispanic residents, fast food dominated food environment, low access to grocery stores, high proportions of uninsured, and high income disparity seem to positively correlate with diabetes-related mortality rate in the neighboring county. County DRMR is negatively associated with neighboring counties’ socioeconomic index, proportion of uninsured residents, proportion of adults who meet physical
activity guidelines, density of hospital beds, density of active medical doctors, increased social capital index, and metropolitan status.

**Bivariate Local Indicator of Spatial Association (LISA)**

The Global bivariate Moran’s *I* provided the mean association between a county’s diabetes-related mortality rate (DRMR) and neighboring counties’ socioeconomic, demographic, and health service factors. However, the heterogeneity in this statistic is not evident, and it does not indicate for which counties the associations are significant. Bivariate LISA was used to identify areas where these associations are significant. The results from bivariate LISA is shown using the map in Figure 3.3a and 3.3b. The High–High and Low–Low locations (positive local spatial correlation) represent spatial clusters, while the High–Low and Low–High locations (negative local spatial correlation) represent spatial outliers. The significance was filtered at a 5% significance level based on 999 permutations.

Depictions in Figure 3.3a and 3.3b reveal that the Southern Plains, Southeastern, and Appalachian regions, where DRMR is high, have also low socioeconomic status, low social capital, high income disparity, high proportions of African American population (except Appalachian areas and Southwest Texas counties), high proportions of the population who lack independent access to grocery stores, and fast food dominated retail food environments. These counties are located in Ohio, West Virginia, Virginia (the Southwest part), North Carolina, South Carolina, Kentucky, Tennessee, Mississippi, Oklahoma, and Texas (Northeast, South, and West parts). High DRMR cluster counties in South and West Texas and California have additional distinct characteristics such as a high proportion of Hispanic population, high proportion of uninsured adults, and less density of hospital beds and physicians. Some DRMR cluster counties (for example, some of the Northern Ohio counties) have high DRMR as well as high socioeconomic and a low income disparity. The high DRMR in the Southern Plain region (Alabama and Mississippi) coincided with higher proportions of the African American population, while the high DRMR in Southern and Western Texas counties coincided with high proportions of the Hispanic population. High levels of residential segregations in some of parts of North Carolina, South Carolina, Oklahoma, Mississippi, Arkansas, Alabama, Texas, Ohio,
and Tulare county (California) overlap with high DRMR. On the contrary, low DRMR counties are described as having high social capital and high socioeconomic indicators (except in Montana), a low proportion of minority population, and a better retail food environment in neighboring counties. These counties are located in the Western part of the Midwest and mountain regions.

**Multivariate association of the DRMR clusters and explanatory variables**

Bivariate LISA analysis indicated that spatial dependence between DRMR and explanatory variables exhibits heterogeneity. There are two shortcomings with the bivariate LISA analysis. First, bivariate LISA analysis provided correlation between DRMR and neighboring county explanatory variables; that is, in a bivariate setting. What would happen to the associations between DRMR and a neighboring county’s explanatory variables in the presence of other explanatory variables is not deducible from bivariate LISA analysis. Second, the correlation statistics are too crude to present for every county. Results from bivariate LISA can be depicted only through maps. To overcome these shortcomings, multinomial logistic regression was used to identify explanatory variables that are contributing to the DRMR clusters (Table 3.3). Results from logistic regression helped to identify the socioeconomic, demographics, retail food access, and health service factors that are driving the observed High–High DRMR county clusters. The type of clusters that were identified using Moran’s *I a priori* are High–High, Low–Low, High–Low, Low–High, and insignificant (non-clustered). The result revealed that High–High DRMR cluster counties are characterized by: low socioeconomic index (OR = 0.48; 95% CI = 0.38, 0.60), fast food dominated retail food environment (OR = 4.52; 95% CI = 2.92, 7.01), a high proportion of white population (OR = 0.98; 95% CI = 0.96, 0.99), a high proportion of Hispanic population (OR = 0.98; 95% CI = 0.96, 0.99), a low proportion of adult population who meet the PA guideline (OR = 0.93; 95% CI = 0.91, 0.96), a high proportion of uninsured adults under age 65 (OR = 1.05; 95% CI = 1.03, 1.08), and higher segregation between subpopulation (OR = 0.12; 95% CI = 0.03, 0.57). On the other hand, Low–Low DRMR cluster counties are characterized by a high proportion of households who lack independent access to
grocery stores (OR = 0.76; 95% CI = 0.68, 0.86), a high percent of adults that meet the PA guideline, and a high proportion of white population (OR = 1.06; 95% CI = 1.02, 1.10).

### Spatial econometrics estimation results

#### Spatial regression diagnostics

The estimation started with OLS regression, followed by diagnostic tests on the residuals to identify the presence and type of spatial dependence. Global Moran’s $I$ and Lagrange Multiplier Lag (and its robust version) were used to identify the type of spatial dependence that may prevail in the data. Global Moran’s $I$ test indicates the presence or absence of spatial dependence in the residuals of OLS regression. A more-specific diagnostic test, Lagrange Multiplier Lag (and its robust version), indicates whether Spatial Autoregressive (SAR) and/or Spatial Error Model (SEM) are appropriate to model the spatially related data. Diagnostic tests in Table 3.4 show that county DRMR is spatially associated, both with Moran’s $I$ and Lagrange Multiplier Lag (and its robust version). Moran’s $I$ and Lagrange Multiplier Lag (and its robust version) statistics are statistically significant (p < 0.001) both for spatial lag and for spatial error models. These results reinforce the conclusions drawn from exploratory spatial analysis. From these diagnostic tests, it is obvious that there is spatial dependence in the residuals of OLS regression, which would violate the assumption of uncorrelated disturbances. As a result, the OLS estimation is both inconsistent and biased. Thus, both the Spatial Autoregressive (SAR) and Spatial Error Model (SEM) are justified to correct for the violated assumptions.

The SEM is motivated by the fact that the counties’ DRMR may not be totally explained by the factors included in this analysis, and the omitted independent variables may correlate across counties, which results in correlated error terms (Anselin, 1988; Baller et al., 2001). The SAR model closely represents a correlation between counties’ DRMR, implying that a county’s DRMR is influenced by a neighboring county’s DRMR. That is, DRMR is lagged across all the neighbors for an area. However, it does not represent the possibility that the neighboring counties’ characteristics could influence the DRMR of a county. The SDM is a generalized model that incorporates the effect of the original county’s socio-ecological characteristics,
neighboring county’s DRMR (lagged dependent variable), and neighboring county’s socio-ecological characteristics (lagged independent variables). SDM is justified because the socioeconomic, demographic, and resource environment in one county may not only influence the DRMR within a county but also the DRMR of neighboring counties because people may move around, crossing county borders to seek medical services, commute to work, and for grocery and shopping trips. In addition, spillover effects may arise from change in socioeconomic factors, services, and policies in a county or multiple counties.

**Spatial econometrics estimation results**

The estimated coefficients from SAR, SEM, and SDM models are presented in Table 3.5. Coefficients from SAR and SDM are not directly interpretable due to the presence of lagged dependent and/or explanatory variables in the regression. Therefore, discussion on coefficients from SDM is postponed to the next subsections. The coefficients from the SAR model are shown only for comparison purpose. The following paragraphs discuss the results from SEM and their implications.

The estimation result from the SEM indicated that there is a statistically significant autocorrelation among the residuals ($\lambda = 0.512, p < 0.001$). The spatial error model is statistically significant for both a lower order neighborhood (queen spatial weights matrix of order one) and a higher order neighborhood, but the magnitude of the correlation statistics is amplified with the higher order of the spatial matrix. This implies that error terms, either from omitted explanatory variables or from measurement errors, are spatially correlated. It is obvious from the result that not all factors that contribute to DRMR are known and accounted for in this model. In fact, the result from SEM estimation confirms that the omitted variables and/or measurement errors are spatially correlated. The SEM accounts for these correlation of errors among counties when it estimates coefficients. Thus, the coefficients from SEM are unbiased as compared to OLS estimates. The estimates from SEM (Table 5) shows that counties’ DRMR are a function of the proportion of Hispanic residents, socioeconomic factors, retail food environment, social capital, density of hospitals, proportion of uninsured adults, lack of access to grocery stores, and extent of urbanization of the county.
The coefficients from SEM can be interpreted as average direct effects of explanatory variables on the counties’ DRMRs. The results from the SEM model show that improvement in economic conditions and social capital of a county would reduce the DRMR by a significant amount. A unit increase in the socioeconomic index (higher index indicating better socioeconomic status) would reduce DRMR by 23.66%. Social capital did not have a statistically significant effect on DRMR. Once socioeconomic status is included in the model, income inequality is statistically significant but with opposite sign to the expectation based on literature. A unit increase of Gini coefficient would reduce DRMR by 70.03%. This result should not be construed as if income disparity actually reduces DRMR. It is simply that counties with population who earn high income have less DRMR. The often-mentioned psychological consequence of income disparity is limited at larger geographical areas such as county and if socioeconomic status difference between counties is incorporated in the model. Similarly, the generally held hypothesis is that counties with higher socioeconomically disadvantaged minority population (Hispanic and African American population) experience a higher mortality rate compared to mostly white population counties. In conformity with this hypothesis, a unit increase in proportion of African American population in a county would increase the DRMR by 20.08%. The effect of change in proportion of Hispanic population on the county’s DRMR is not statistically significant. Residential segregation was associated with increased DRMR. A unit increase in the spatial proximity index (SPI) would increase the DRMR by 20.44%.

The type of food available in a retail market is an important factor in diabetes self-management. To test this hypothesis, retail food environment and the proportion of populations who lack independent access to grocery stores were included in the model. The result shows that an increase in the food environment index (higher index indicating fast food dominated food retail environment) would increase DRMR by 7.79%. Access to grocery stores was measured by the proportion of a county population without vehicles and residing more than a mile (10 mile for rural residents) from a grocery store. The unit increase in the proportion of residents who lack independent access to grocery stores did not yield a statistically significant effect. However, it should be noted that the coefficient is the average of the spatially heterogeneous relationships, which might hide the true association between access to grocery stores and DRMR locally. Geographically Weighted Regression (GWR) revealed that in the Western, Southwestern, and
mountain regions of the US lack of access to grocery stores was negatively associated with DRMR. In the Northern, Southeast, and the Western part of the Midwest, Appalachia, and New England, lack of access to grocery stores was positively associated with DRMR.

Successful management of diabetes may require frequent contact with primary care facilities and hospitals (Willens et al., 2011). Increased density of healthcare facilities and physicians promote more access to care, thereby reducing the mortality rate. However, density of facilities and services by themselves do not increase the utilization rate. The results from SEM revealed that once socioeconomic index and social capital index are included, the hospital bed density and physician density coefficients were not statistically significant. However, a breakdown using GWR shows negative associations between DRMR and primary care facilities and hospitals in the Southwestern region (Texas, Nebraska, and Kansas), New England, and Montana, and a positive association in the Midwest, Northern, Southeastern, Western, mountain, and Appalachia regions. This result indicates that in the regions where DRMR clustered, the density of healthcare facilities and physicians is also high. This result simply reflects the fact that the density of healthcare facilities and physicians are higher in areas with a high demand for care. Diabetes being one of the diseases that necessitate continuous healthcare visits and education, counties with a high number of diabetic patients have a high demand for healthcare facilities and physicians.

The literature asserts that insurance coverage would increase access to medical care, thereby reducing mortality rate (Sommers, Baicker, & Epstein, 2012; Sommers, Long, & Baicker, 2014). When socioeconomic status was included in the model, the result from SEM and other regression models was contrary to the literature. The association between proportion of non-senior uninsured adults and DRMR shifted from positive to negative once socioeconomic and proportion of minority was included in the model. However, the bivariate association between DRMR and proportion of uninsured adults was positive. As more variables (except socioeconomic index) were included in the model, the coefficient of proportion of uninsured population changed signs. A unit increase in the proportion of uninsured adult (age 18–64) residents would decrease the mortality rate by 37.31%. Once the underlying factor for high proportion of uninsured was controlled for (lower socioeconomic status and proportion of African American and Hispanic populations), the proportion of the uninsured population
measures the proportion of adults who were not participating in the insurance market. Non-
participants in the insurance market tend to be healthy young adults (age 19–34) and tend to be
employed in agriculture and construction sectors and small business (Kaiser Family Foundation,
2013). Young adults are less concerned about health, and therefore might not participate in the
insurance market (Holahan & Kenney, 2008) prior to ACA. Agricultural workers and employees
of small businesses are not covered by the employer-based insurance market. Thus, counties
where agriculture is a major economic sector and there is a high proportion of young adults may
also have a high proportion of uninsured non-senior adults. This is very evident if we break
down the regression results by region. In the Western and mountain regions of the US, where the
proportion of adults who meet physical activity guidelines was high and mortality rate was low,
the proportion of the uninsured was also high. In addition, it should be noted that our outcome is
mortality rate. In this regard, the Hill-Burton Act established the requirement that hospitals
provide necessary services to stabilize a patient regardless of the patient’s ability to pay
(Melhado, 2006). Therefore, it makes intuitive sense that insurance coverage would have
influence on healthcare utilization as implied in the literature but would not be a significant
predictor of mortality rates.

**Spatial Durbin model: direct and indirect effect**

As the case with SAR model, the regression coefficients from SDM are not interpretable
directly. The estimated parameters need to be decomposed into direct and indirect effect (spatial
spillover effect), which provide substantial interpretations (Pace & LeSage, 2006). As discussed
previously, the direct impact is the average impact of change in an explanatory variable in the
same county, which also accounts for the feedback effect from all other counties in the sample.
The indirect effect is the measure of spillover effect, which is the average change in a specific
county’s DRMR due to change in explanatory variables in neighboring counties (LeSage & Pace,
2009; Pietrzak, 2013; LeSage & Pace, 2014). Neighboring counties in this case are determined
by a first order queen contiguity spatial matrix. The total impact is interpreted as the average
total impact of change in explanatory variables in all counties on the DRMR in a typical county
The direct, indirect, and total impact of change in independent variables on the DRMR are presented in Table 3.6.

Pertinent to the above interpretations, the impact of change in proportion of minority on the DRMR is limited to direct impact. However, the increase in proportion of population (African American and Hispanic) would have a negative spillover effect (increase in DRMR). A unit increase in the proportion of African American and Hispanic residents in neighboring counties would increase the DRMR in a specific county by 41.48% and 60.151%, respectively. If the proportion of African American and Hispanic populations increased by a unit simultaneously in all other counties, DRMR would increase by 43.62% and 47.55% in a typical contiguous US county, respectively. Residential segregation was associated with DRMR in the same county, but there was no spillover effect. An increase in the spatial proximity index would entail an increase of DRMR by 20.14% in the same county. Socioeconomic indicators such as education, employment and household income, and retail food environment and access usually follow the pattern of ethnic composition and residential segregation in a county. Therefore, the demographic composition results should be looked at in association with the socioeconomic status of the subpopulation and genetic predisposition of the subpopulation to diabetes.

The impact of improvement in socioeconomic status, at county and national levels, would reduce DRMR tremendously. A unit increase in the socioeconomic index of a particular county would decrease the DRMR by 25.02% in the same county. In addition, a unit increase in the socioeconomic index simultaneously in all counties would decreases DRMR by 24.87% in a typical contiguous US county. Socioeconomic status did not have immediate spillover effects to the neighboring counties. The direct impact of the social capital index (SCI) on DRMR is not statistically significant, but it has a spillover effect on the neighboring counties. If SKI increased by one unit in neighboring counties, DRMR would increase by 4.6% in a specific county. The positive association between increased neighboring counties’ SKI and increased DRMR in a specific county can be interpreted as the negative psychological effect of neighboring counties’ relatively high social capital on the health of the residents of a county with low social capital. Similarly, income inequality did not yield the expected result. As discussed previously, once economic characteristics and social capital are controlled in the model the income disparity measures the income differences between counties. The negative sign on the direct, indirect, and
total income disparity effects indicate that impact of increased income in neighboring counties and within the same county would have negative effect on the DRMR in a specific county as well as on any typical US county.

After controlling for socioeconomic status, hospital and physician density did not yield statistically significant results. The direct impact of the proportion of the uninsured population (age 18–64) has the opposite sign than expected, but no spillover effect from change in the proportion of uninsured in neighboring counties. Rural–Urban continuum index has a statistically significant impact. If the proportion of the urban population increases by a unit, DRMR in the same county would increase by 20.08%. If the proportion of the urban population increases simultaneously in all counties, DRMR would decreases by 27.51% in a typical contiguous US county. This is “rural–urban paradox”, which only appears after the socioeconomic index is included in the model.

The impact of change in the Retail Food Environment Index (RFEI) was not limited to direct impact. There was spillover effect to the neighboring counties. An increase in RFEI in a particular county would increase the DRMR by 9.53% in the same county. A unit increase in RFEI in neighboring counties would increase DRMR by 30.22% in a specific county. A unit increase in RFEI simultaneously in all other counties would increases DRMR by 42.62% in a typical contiguous US county. Similarly, an increase in proportion of residents who lack independent access to grocery stores would increase DRMR by 0.8% in the same county but have no spillover effect.

3.4 Discussions

This study was designed to test three hypotheses: 1) if diabetes-related mortality rates (DRMR) in US counties are spatially associated; 2) if DRMR is spatially associated with socioeconomic, demographic, and health service factors; and 3) if there is a spatial spillover effect from the association of socioeconomic, demographic, and health service factors. The first hypothesis (spatial dependence in county DRMR) was tested using exploratory spatial analysis (ESDA) tools and spatial econometrics estimations. Using ESDA tools, the study did not only prove that DRMR in US counties are spatially correlated but also identified clusters of counties
where DRMR is high and low. Four hundred and one counties were identified as hot spots, where unusually high DRMR counties were clustered. Three hundred and fourteen counties were identified as cold spot, where unusually low DRMR counties were clustered. Hot spot counties are located throughout the Southern Plains, Southeastern, and Appalachian regions of the US. States where high DRMR is clustered include Ohio, West Virginia, Virginia (the western part), Pennsylvania, North Carolina, South Carolina, Kentucky, Tennessee, Mississippi, Oklahoma, and Texas. The cold spots were observed in the mountain regions (Montana, Wyoming, Idaho, South Dakota, Utah, and Colorado), and Southwest states (Nevada and Northwestern California).

The exploratory and spatial econometric analysis results confirmed the second hypothesis that socioeconomic, demographic, and health service characteristics might show a spatial pattern of association with high and low DRMR clusters. The exploratory analysis and econometric regressions revealed that high DRMR regions are characterized by lower socioeconomic status, a fast food dominated retail food environment, a high percentage of the population who lack access to grocery stores, a high proportion of African American population, and a low percentage of adult population who meet the PA guideline. Whereas low DRMR cluster regions are characterized by a low proportion of African American population, higher socioeconomic status, a high percentage of adult population who meet the PA guideline, a non-fast-food dominated retail food environment, and a high proportion of residents who have independent access to grocery stores. This result helped to differentiate the underlying socioeconomic, demographic, and health service factors that are detrimental to the observed DRMR hot and cold spots.

Understanding of regional clustering of DRMR and associated ecological factors (socioeconomic factors, access to healthy food, and healthcare infrastructure) is vital to developing effective strategies and health delivery policies in order to reduce the burden of DRMR. In light of this, we should note that the counties where DRMR is high have distinct socioeconomic, demographic, and food environment characteristics. They have a lower median household income, lower level of education, and high unemployment rate. Fast food restaurants and lack of access to grocery stores and full-service restaurants dominate the food environment in these counties. A significant percent of residents in these counties lack access to healthy food due to distance from grocery stores and a lack of independent access to a vehicle. A large proportion of residents in high DRMR counties also did not meet the PA guideline compared to
counties where DRMR did not cluster. This is an indication that the majority of the population in high DRMR clusters lack awareness of the health benefits from engaging in PA and may also lack the opportunities to engage in physical activities. This result is similar to the findings of Shrestha et al. (2012) on the association between socio-ecological factors and clusters of diabetes.

Further, the spatial econometris results confirmed the third hypothesis that changes in socio-ecological factors might exert spillover effect on the DRMR. The spatial spillover effect that may arise due to change in socio-ecological factors in one or more counties was tested using the Spatial Durbin Model (SDM). According to this model, in addition to the direct impact within the same counties, simultaneous improvement in socioeconomics index (SEI) in all other counties would reduce DRMR by about 29% in a typical contiguous US county. However, the spillover effect from change in SEI in immediate neighboring counties to a particular county was not statistically significant. These results are partially consistent with other spatial studies of the association between socioeconomic status and overall mortality (Yi-Ju Chen et al., 2012; Yang, Noah, & Shoff, 2013). However, unlike Yang, Noah, and Shoff (2013), this study did not find the spillover effect of socioeconomic change to the immediate neighboring counties. With regard to access to health, once socioeconomic, social capital, or proportion of minority in the county is controlled, hospital bed density and proportion of uninsured did not yield the expected results. The spillover effect from increased hospital bed density and the proportion of uninsured to other counties is statistically insignificant. The spillover effect from physician’s density (whether it is all physicians or primary care physicians) was the opposite of the expectation; it was positively associated with DRMR in other counties. This implies that physicians are located in places where there is greatest demand. In this case, diabetes cluster attracts more primary care physicians. The spatial spillover effect from change in the retail food environment in neighboring county would increase DRMR by 26% in a particular county and a simultaneous increase in RFEI in all other counties would increase DRMR by 35% to a typical contiguous US county. Therefore, improvement in the retail food environment in a particular county may have beneficial spillover effect to the immediate neighboring counties’ DRMR. The implication that can be drawn from these results is that economic, social capital, and retail food environments are
the most important factors associated with DRMR. Improvement in socioeconomic status and access to healthy food would significantly reduce DRMR in contiguous US counties.

This study has a few limitations that need to be interpreted cautiously. First, the use of counties as units of analysis for this study, and thus interpretations of the associations between DRMR and socio-ecological factors, should be made with caution. A well-known problem with aggregated data is the ecologic fallacy—an assumption that associations among variables assessed from aggregate data apply to analogous individual-level variables, which may lead to wrong inferences at an individual level. It hides the heterogeneity between individuals and joint distribution of explanatory variables at an individual level (Morgenstern, 1995). Aware of this problem, the variables in this study are mostly at the ecological level, and interpretations of the results strictly followed the ecological nature of the variables. Second, not all of the socio-ecological phenomena in the study are necessarily going on at the county-level; the conclusion from this study may encounter the so-called modifiable areal unit problem. Analysis from a county-level study may not hold for smaller (e.g., neighborhood or census tract) or larger (e.g., a state) geographical units. The dynamic related to the migration of people from county to county is also not accounted for in this study. On the other hand, a county is the smallest geographical unit for which mortality data is publicly available. Moreover, the planning and execution of much of the health services and economic activity is at county-level. Thus, the use of county as unit of analysis for mortality data is justified. Even with such limitations, ecologic analysis is often preferred when a phenomenon under study is public health implementation where a geographic boundary has a meaning and if variable selection is appropriate (Portnov, Dubnov, & Barchana, 2006). In addition, smaller mortality counts are either suppressed or flagged as unreliable in the CDC WONDER publicly available mortality data. The unreliability of this data for the smaller mortality counts would certainly introduce heteroscedasticity during parameter estimation.

Despite the inherent limitations of an ecological study, the findings from this study have strengths. First, the study not only confirmed contiguous US counties’ DRMR were spatially correlated but also were clustered in some regions. Second, our study showed how to combine exploratory spatial analysis and spatial econometric methods to identify high DRMR cluster counties and associated socio-ecological factors. It combined exploratory spatial statistics,
multinomial regression, and spatial regression models to overcome some of the issues posed by heterogeneity in the ecological studies. For example, combination of Moran’s I statistics and multinomial regressions helped in identifying socio-ecological factors that are associated with cold spots and hot spots. Identification of DRMR clusters is useful in health resource allocation and management. Health managers should target the resources to the areas identified as high DRMR clusters. Third, this study used spatial autoregressive and Spatial Durbin model to identify the most impactful socio-ecological factors if intervention has to be designed. The identification of the most impactful socio-ecological factors associated with DRMR encourages channeling of resources to those most impactful factors and most affected geographical areas in reducing DRMR and general mortality rate.

3.5 Conclusions

The policy implication from these findings is that efforts to reduce DRMR should invest in the most impactful ecological factors such as improving the socioeconomic gradients (income, education, and employment) and access to healthy food and social capitals (environment where residents engage in physical activities and social activities). Intervention programs should target geographies with a high cluster for counties while strategically considering the heterogeneous interrelationships between counties due to socioeconomic and policy interactions and shared resources. If policymakers wish to improve health, public funds are better spent on income transfers, education, and public health programs.
### 3.6 References


California Centre for Public Health Advocacy, Policy Link, the UCLA Center for Health Policy Research: *Designed for Disease: The Link Between Local Food Environments and Obesity and Diabetes*. Davis, CA. 2008.


Heath GW, Brownson RC, Kruger J et al. The effectiveness of urban design and land-use and transport policies and practices to increase physical activity: a systematic review. *J Phys Act Health* 2006; 3 (1): 55–76.


Kaiser Commission on Medicaid and the Uninsured/Urban Institute analysis of the 2013. ASEC Supplement to the CPS.


LeSage JP, Pace RK. Interpreting spatial econometric models. *Handbook of Regional Science* 2014; 1535-1552.


United States Department of Health and Human Services (US DHHS), Centers for Disease Control and Prevention (CDC), National Center for Health Statistics (NCHS), Compressed Mortality File (CMF) on CDC WONDER Online Database. The current release for years 1999 - 2010 is compiled from: CMF 1999-2010, Series 20, No. 2P, 2013.


Weng C, Coppini DV, Sönksen PH. Geographic and social factors are related to increased morbidity and mortality rates in diabetic patients. Diabetic Medicine 2000; 17(8): 612-617.


Chapter 4: What is the Value in Diabetes Prevention? Well-being Valuation Approach

Abstract

The prevention of chronic diseases such as diabetes is the focus of the affordable care act (ACA). Several prevention programs such as community and workplace based wellbeing programs are currently in place. The current preference based health valuation approach is not appropriate for prevention based programs valuation because they do not capture the social and economic value that an individual puts on a health condition. I utilize a recently developed subjective wellbeing valuation approach to quantify the monetary value of loss in wellbeing due to diabetes in the United States population. This approach assumes that individuals derive overall life satisfaction from wellbeing, which is a function of health and income. Health, in turn, is produced by the combined input of an individual’s behaviors and medical technology. Thus, a marginal trade-off between health and income is used to derive the monetary value of health. The Panel Study of Income Dynamics (PSID) data was utilized for this study. The results shows that the monetary value for diabetes prevention is approximately $37,000, which is less than the current implicit threshold for program implementation. This is the amount that individuals are willing to pay to not to have diabetes. The resulting monetary value can be interpreted as societal value for diabetes prevention, which can be used as benefit side of the in cost-benefit equation diabetes prevention evaluation.
4.1 Introduction

Prevention of chronic diseases such as diabetes is one of the provisions of the affordable care act (ACA). Currently, efforts are underway to scale-up the results from lifestyle modification interventions at local community organizations and workplace wellness programs. Young Men's Christian Association (YMCA) and CDC programs have established intensive lifestyle modification coach training centers (Preston & Alexander, 2010). In addition, a survey of large manufacturing companies showed that 77% of the companies offered workplace wellness programs (Anderko et al., 2012). At worksites, employees are encouraged to take self-administered health risk assessment (HRA) and participate in employer sponsored wellness programs, which include educational materials, counseling, fitness programs, and onsite clinics (Preston & Alexander, 2010). The 2010 Affordable Care Act (ACA) included several provisions to support the efforts of chronic disease (including diabetes) prevention programs, which allowed for screening and implementing chronic diseases preventative measures without additional fees and established funding for preventative programs (Preston & Alexander, 2010; Anderko et al., 2012).

Unfortunately, such preventative programs also need to compete for limited health care resources among other preventative and treatment programs. The prioritization of the resources allocation to the programs hinges on the valuation of the benefits and costs. The benefit-cost analysis for goods bought and sold in the marketplace is straightforward. Since health is not a good sold or bought in the market, quantifying the benefit part of the equation in benefit-cost analysis is difficult. The benefit side of the benefit-cost equation fundamentally raises the question of how much value people place on a particular health outcome. The most common method to quantify the value (benefit) of a particular health outcome is willingness to pay, which is usually measured using direct or indirect methods (Dolan & Kahneman, 2008).

In the direct method, a survey-based hypothetical choice of health technology and associated monetary value is presented to otherwise healthy people to elicit willingness to pay. This approach is known as stated preference approach which can take a form of contingent valuation (CV) and conjoint analysis methods. For example: van Gils et al. (2011) used conjoint analysis to measure willingness to pay for the combinations of program length, physical activity lessons, group sports activities, counseling, financial incentive, and out-of-pocket costs; further,
Veldwijk et al. (2013) employed conjoint analysis to evaluate the willingness to pay for the combinations of meal plans and physical activity (PA) schedule, consultation structure, expected program outcomes and out-of-pocket cost, and monetary incentive to elicit willingness to pay for a lifestyle modification program. A few of the limitations of the stated preference approach are: first, survey respondents are usually presented with hypothetical health choices or health technology of which they have never personally experienced or evaluated before (Dolan & Kahneman, 2008). Second, respondents may also find it difficult to fully comprehend the actual risks or consequence of choice involved in the situation hypothetically presented to them (Groot & Bring, 2006).

The Revealed Preference (RP) approach is the indirect method used to measure willingness to pay. This approach relies on the people’s revealed choices or behaviors in the real market to compute monetary values for non-market goods (e.g. different health conditions). The most common RP approach is observing a health risk that people are willing to trade off in exchange for a higher wage. This approach has several limitations (Groot & Bring, 2006). The workers who have chosen a specific type of occupation are not a random sample of the general population; rather, these are selected groups of individuals for whom the risk from the job has less weight than for those individuals who chose other occupations. Thus, the sample is not representative of the population, which might lead to selection bias.

An alternative approach, which this paper intends to pursue, is the wellbeing valuation. The Wellbeing Valuation approach uses subjective wellbeing measures to derive marginal rates of substitution (MRS) between a non-market good (such as health and unemployment) and income (Fujiwara, 2013). The approach relies on the assumption that subjective wellbeing can be used as a proxy for underlying utility. The marginal rate of substitution derived from the wellbeing approach can be used to measure compensating income variation (CIV) and equivalent income variation (EIV), which can be interpreted as willingness to accept and willingness to pay respectively. This approach overcomes the limitations of RP and CV (Groot & Bring, 2006). This approach is based on the respondents’ evaluation of their life satisfaction, individual level data on the incidence of a health condition, household income, and other socio-demographic characteristics.
This method has attracted researchers in various disciplines as a tool to evaluate the value of various life and social events. For instance: Clark and Oswald (2002) used the Wellbeing valuation method to calculate the values of illness, marriage or unemployment; Frey et al. (2004) quantified the trade-off between terrorism and income; Van Praag and Baarsma (2005) quantified the monetary value required to offset reduced wellbeing due to the noise caused by aircrafts; Welsch (2006) obtained value estimates for air pollution; and Powdthavee (2008) attached monetary values to social capital. It is important to acknowledge that this method has its own limitation including the interpersonal comparability of life satisfaction, state dependence of the response for ill and healthy persons and response to life satisfaction question may be mediated by culture.

Within the health discipline, similar attempts have been made to evaluate the monetary value that individuals attach to specific diseases. Ferrer-i-Carbonell and van Praag (2002) evaluated the monetary amount needed to compensate for multiple chronic diseases in a German population; Groot and Van den Brink (2004) evaluated the monetary value to satisfy people with headaches and migraines in a Dutch population; Groot and Van den Bring (2006) assessed the monetary value that may make people with cardiovascular diseases as well off as people without it in a Dutch population; and Powdthavee and Van den Berg (2011) used a multiple wellbeing measure to evaluate monetary value that people may attach to multiple diseases in a British population (See Table C.4 for summary of these studies). The studies calculated compensating income variation (CIV), which can be interpreted as the monetary value that people are willing to accept in exchange for poorer health.

The CIV computed from wellbeing valuation methodology thus far has provided biased estimates of the value of health conditions (Dolan & Metcalfe, 2008; Dolan, Fujiwara & Metcalfe, 2011; Fujiwara, 2013). The main reason for the exaggerated compensating income variation value is that the estimated effect of income on subjective wellbeing is biased downward. The previous studies, except for Ferrer-i-Carbonell and van Praag (2002), have only investigated the direct effect of income on subjective wellbeing, while controlling for socio-demographics, health conditions, and the economic variable. The indirect effect of income on subjective wellbeing through the factors that are controlled in the regressions (for example, health, unemployment, and marital status) (Dolan, Fujiwara, & Metcalfe, 2011) has been
ignored. However, controlling for these factors without accounting for indirect effects of income may bias the effect of income on subjective wellbeing downward (Dolan, Fujiwara, & Metcalfe, 2011). This may be why the income effect appears small when other associated factors are controlled (Dolan & Peasgood, 2006). Thus, the CIV calculated from such biased estimates is highly exaggerated. The current study contributes to the empirical validation of the method by using the recently developed step approach estimation procedure by Dolan, Fujiwara, and Metcalfe (2011), to account for the indirect effect of income to correct the biases and derive accurate CIV for diabetes.

**Research questions**

1. How much monetary value are individuals with diabetes willing to pay/accept not to have diabetes/to feel as well off as individuals without diabetes?
2. Does willingness to pay/accept or place a monetary value for diabetes differ by age group, gender, and ethnicity?

**Objectives**

1. To compute the monetary value that patients with diabetes are willing to pay/accept not to have diabetes/in order to feel as well off as individuals without diabetes.
2. To evaluate the variation in the monetary value of diabetes among age group, gender, and racial/ethnic groups.

**Hypotheses**

1. Reasonable monetary value can be quantified for loss of wellbeing due to diabetes using the subjective wellbeing method.
2. The monetary value assigned to the loss of wellbeing varies by age, gender, and race/ethnicity.

**Subjective wellbeing**
Neoclassical economists use inferred utility from individuals’ preference based on the assumptions that individuals make rational decisions, are fully informed about their choice, and seek to maximize utility (Dolan, Peasgood, & White, 2008). Over the years, however, economists and psychologists casted a doubt on the use of preference methods. The concern is that preferences may not reflect the change in wellbeing as a consequence of one’s choice. Behavioral economists have shown that people make inconsistent and rationally-bound choices, which is inconsistent with the neoclassical model of rational economic agents (Kahneman & Krueger, 2006).

The subjective wellbeing approach is a detour from neoclassical economic assumptions and assumes that utility, or wellbeing, can directly be measured. Thus, the subjective wellbeing approach takes the individuals’ wellbeing to be their overall assessment of their welfare. It measures human perception of their wellbeing by simply asking how their life is going. Layard (2005) argued that human perception is vital to understanding an individual’s wellbeing, as no one knows how a person is feeling except the person himself/herself. Subjective wellbeing includes both emotional reactions and cognitive judgments, and therefore encompasses moods and emotions as well as evaluations of satisfaction with general and specific areas of life (Diener, 2009). It involves such concepts as positive and negative affect, happiness, and life satisfaction (Diener, 2009).

**Measuring subjective wellbeing**

Subjective wellbeing is commonly measured with life satisfaction questions. Life satisfaction approach asks people to assess their life or circumstances of their life in a reflective judgment manner (Diener, 2009). Life satisfaction measures may assess overall life satisfaction or a certain domain of life such as satisfaction with their health, job, and relationships. Examples of overall life satisfaction measures use questions such as “All things considered, how satisfied are you with your life as a whole these days?” in The World Values Survey conducted in 81 countries, and “Taken all together, how would you say things are these days? Would you say that you are very happy, pretty happy, or not too happy?” in The General Social Survey (GSS) in the United States (Kahneman & Krueger, 2006). The Cantril ladder of life in which respondents rate
their current life on a ladder scale for which 0 is ‘the worst possible life for you’ and 10 is ‘the best possible life for you’ is an alternative measure of life satisfaction (Kahneman & Krueger, 2006).

This paper measures subjective wellbeing with a general life satisfaction question which is available in the Panel Survey of Income Dynamics (PSID). Subjects answer the question: “Please think about your life as a whole. How satisfied are you with it? Are you completely satisfied, very satisfied, somewhat satisfied, not very satisfied, or not at all satisfied?”

Life satisfaction is appealing because of its close relation to the utilitarian notion of the subjective wellbeing and it reflects the individual assessment of their overall life (Veenhoven, 1993). In addition, Campbell et al. (1976) point out that life satisfaction is a less vague concept which can be defined as discrepancy between aspiration and achievement. Furthermore, Pavot and Diener (1993) noted that life satisfaction is a more stable measure of subjective wellbeing compared to positive and negative affects.

**Determinants of subjective wellbeing**

Subjective wellbeing depends on a range of factors. Determinants of wellbeing can be individual level factors such as personality; socio-demographic factors such as age, gender, ethnicity, education, and social relationships (e.g. marriage); economics such as income and employment; and situational factors such as health status. In this section, some of the well-researched correlates of subjective wellbeing are reviewed in order to motivate the selection of variables in our analysis.

**Health**

A positive relationship holds between life satisfaction (subjective wellbeing) and self-assessed health (Dolan & Kahneman, 2008; Dolan et al., 2008; Easterlin, 2003). Causality is bi-directional because health status affects wellbeing and wellbeing affects health, but with a delay of two to three years. (Binder & Coad, 2010; Lyubomirsky, King, & Diener, 2005). The stronger
relationship even after controlling for such reverse causality, however, shows that the effect of health on subjective wellbeing is greater (Dolan, Peasgood, & White, 2008).

In a study on specific conditions, Shields and Price (2005) reported decreased subjective well-being for individuals with problems with muscular-arthritis-rheumatism, stomach problems and respiratory problems using a cross-sectional analysis of the Health Survey for England (HSE) data. The authors did not find significant associations between cancer and diabetes and well-being. They reported differences in specific conditions on psychological well-being among men and women. Heart attacks, strokes, migraines, and epilepsy were associated with decreased psychological well-being in males; while hypertension and blood pressure problems were associated with decreased psychological well-being in females (Shields & Price, 2005).

**Income**

Studies predominantly show that the relationship between income and subjective wellbeing is positive (Diener, 2009), even after controlling for variables such as education (Dolan, Peasgood, & White, 2008). The relationship between income and subjective wellbeing can also be seen from the reverse direction. Income is not necessarily an exogenous determinant that can improve a person’s wellbeing, as people with higher life satisfaction (subjective wellbeing) may generate more income than people with low life satisfaction (Diener, 2009).

The effect of income is often small when other factors are controlled (Dolan & Peasgood, 2006). However, income may indirectly affect subjective wellbeing through these controlled factors (example: health, unemployment, social relationships, marital status and place of residence) (Dolan, Fujiwara, & Metcalfe, 2011). Controlling for these factors could cause downward bias on the effect of income on subjective well-being (Dolan, Fujiwara, & Metcalfe, 2011).
Age

The effects of age on the life satisfaction are complex. Evidence shows that younger and older people are happier than middle-aged people both for men and for women (Dolan, Peasgood & White, 2008). Empirical evidence shows that life satisfaction reaches its minimum between mid-30s and early 50s (Clark, 2007; Blanchflower & Oswald, 2008; Van Landegghem, 2012). A meta-analysis by Frey and Stutzer (2002) indicated that the relationship between age and happiness is convex even after controlling for other demographics, income, health status, marital status, and employment. This is now dubbed a U-shaped relationship in the literature. The possible reason for this relationship is that middle age people have higher expectations that are un-met compared to older people, while older people have realistic expectations and adapt to their situation (Ullola, Moller, & Sousa-Poza, 2013). Kahneman and Krueger (2006) find that people with teenagers at home experienced lowest life satisfaction and reported satisfaction improvement thereafter.

Gender

Some studies report gender difference in life satisfaction (Fujita et al., 1991) while others report no difference between men and women (Okun & George 1984). According to Tesch-Römer, Motel-Klingebiel, & Tomasik (2008), women are less satisfied with life compared to their male counterparts because they are disadvantaged in terms of opportunity structures and action resources. On the contrary, Alesina, Di Tella, & MacCulloch (2004) reported that women tend to report higher levels of happiness.

Race

Ethnicity-related difference in life satisfaction in a multiracial society like United States is an inevitable matter due to discrimination (Blanchflower & Oswald, 2004). In the US, Whites have higher life satisfaction than African Americans (Thoits & Hewitt, 2001).
Education

Some studies find a positive relationship between each additional level of education and subjective wellbeing (Blanchflower & Oswald, 2004). Additional years of schooling are positively correlated with life satisfaction even after controlling for the possibility of reverse causation (Oreopoulos, 2003; Kahneman & Krueger, 2006). Education may reflect some unobservable personal variables such as motivation, intelligence or family background (Dolan, Peasgood & White, 2008). In addition to the direct effects, education contributes to life satisfaction through income and health (Dolan, Peasgood & White, 2008).

Unemployment

Several studies have found negative effects of individual unemployment on life satisfaction (Dolan, Peasgood & White, 2008). Blanchflower (2001), Clark and Oswald (1994), Di Tella et al. (2001) found a negative correlation between wellbeing and unemployment. Clark (2006) used the British Household Panel Survey (BHPS), European Community Household Panel (ECHP) and German Socio-Economic Panel (G-SOEP) to show unemployment negatively affects a person even after years and that adaptation to the situation is very limited. Similarly, Lucas et al. (2004) use GSOEP data to conclude that adaptation to unemployment is incomplete.

Marriage (relationship)

Studies show that being in a relationship (fulfilling and stable partnership) is positively associated with a higher life satisfaction and being single (separated, divorced or widowed or alone) is associated with a lower life satisfaction (Helliwell, 2003; Stutzer & Frey, 2006). However, the direction of causation is a matter of further research; it is not clear whether people who are satisfied with life get married or people who are married get life satisfaction (Stutzer & Frey, 2006).
Wellbeing and health production

This section briefly discusses the theoretical relationship between subjective wellbeing and health productions at the individual level. The assumption here is that an individual’s ultimate objective is to maximize subjective wellbeing (life satisfaction) through consumption of market goods (and services) and health. Subjective well-being (SWB), a measure of how people evaluate their lives in general (incorporating physical, social and psychological dimensions of wellbeing (Diener, Christie & Richard, 2009), can be used as proxy for utility to be maximized by individuals (Stutzer & Frey, 2004). Since we do not observe quantity goods consumed but income, the consumption of composite market goods and services can be represented by income (Welsch & Kuhling, 2011).

Provided that the ultimate goal of individuals is to maximize wellbeing (utility), they produce and consume health along with other goods and services. In this regard, the seminal work of Grossman (1972) is the first model to formalize the production and consumption of health. In Grossman’s initial health production model, households acts both as consumers and as producers of health. In Grossman’s model, individuals invest in health by allocating time and consuming medical care (medical goods and services). The demand for health production inputs is a derived demand, that is, individuals invest in health production inputs because they derive wellbeing (utility) from the resulting commodity (i.e. good health). On the other hand, individuals consume health because health provides utility and production benefits (i.e. healthy individuals have higher earnings).

The modified health production function can be written as

\[ H_t = f(B, L, X, M, E, H_{t-1}, S) \]

where \( H_t \) is the current stock of health, \( B \) is biological factors (e.g. age and genetics), \( L \) is a vector of lifestyle choices or health-related behaviors, \( X \) represents nonmedical purchased inputs to improve or maintain health, \( M \) is purchased medical inputs, \( E \) is environmental factors (natural, built and social environments), \( H_{t-1} \) is existing health stock, and \( S \) is education level. The function \( f \) is a biological function that shows how determinants of health are related to health stock.

This model views health as a durable capital stock that yields an output of healthy period in life. Individuals inherit an initial amount of health stock, which is determined by heredity, environment, and chance and it depreciates with age (Grossman, 2002). Individuals can increase
their initial health stock by behavioral and medical care investments. The return to investment in health inputs (positive health behaviors like exercise and healthy eating or negative health behaviors such as smoking) is not seen immediately. The individual’s level of investment in health production inputs partially depends on the time preference. For instance; behaviors such as smoking may provide immediate satisfaction, but the health stock of individuals may decline over the long run. On the contrary, behaviors as healthy dietary habits and regular exercise may decrease current utility, but increase health stock. The optimal investment is determined by the trade-offs that individuals face between choices that provide direct satisfaction and other behaviors, such as physical activity that are expected to improve health (Grossman, 2002).

4.2 Methods
Theoretical Model

The central assumption in the subjective wellbeing valuation model is subjective wellbeing is a good approximation for individual utility (Dolan, Fujiwara, & Metcalfe, 2011). Subjective wellbeing is a monotonic transformation of utility only observable by the individual. Utility in this model refers to the experienced utility (not decision utility) as disentangled in Dolan and Kahneman (2008). Subjective wellbeing can be measured by global evaluations of overall life (e.g., life satisfaction). Let us assume a generalized subjective wellbeing function, which is measurable by overall life satisfaction and depends on the income, health status, and socio-demographic factors. The subjective wellbeing function can be expressed as:

\[ W = u(H, Y, Z) \]  

Where: \( W \) is subjective wellbeing, \( u \) is utility observable only by the individual, \( H \) is health condition (diabetes), \( Y \) is income, and \( Z \) is a set of socio-demographic and economic variables.

Now, let the change in health status of an individual be \( \Delta H \) and let the change in income that make the individual as well off as before the health status change be \( \Delta Y \). Assuming that health changes from good to bad, the subjective wellbeing of the individual would decrease. Since subjective wellbeing depends on both the income and the health status of an individual, it is possible to derive the amount of income change that would be needed to compensate for the loss
of subjective wellbeing (due to change in health status) to keep the individual at the same
subjective wellbeing level as before the change in health status. Formally,

\[ W(H, Y, Z) = W(H + \Delta H, Y + \Delta Y, Z) \] (2)

Keeping \( W \) constant and taking differentiation with respect to \( H \) and \( Y \)

\[ \frac{\partial w}{\partial H} + \frac{\partial w}{\partial Y} = \frac{\partial w}{\partial H}(1 + \Delta H) + \frac{\partial w}{\partial Y}(1 + \Delta Y) \]

Rearranging and canceling out terms

\[ -\frac{\partial W}{\partial H} \Delta H = \frac{\partial W}{\partial Y} \Delta Y \]

\[ \Delta Y / \Delta H = -\frac{\partial W}{\partial H} / \frac{\partial W}{\partial Y} \]

As \( \Delta H \) tend to zero, the income deeded to compensate for loss of health status is

\[ \frac{dY}{dH} = -\frac{\partial W}{\partial H} / \frac{\partial W}{\partial Y} \] (3)

The value estimated from this method is equivalent to the compensating income variation (CIV).

Accounting for endogeniety of income, where income indirectly affects subjective wellbeing
through health and socio-demographic and economic variables the subjective wellbeing function

\[ W = u(H, H(Y), Y, Z, Z(Y)) \] (4)

The associated compensating income variation can be derived as in (3):

\[ \frac{dY}{dH} = -\frac{\partial w}{\partial H} / \frac{\partial w}{\partial Y} \] (5)
Data and descriptive statistics

The estimation for this study is conducted using 2009 and 2011 waves of the Panel Study of Income Dynamics (PSID) data. The study began in 1968 with a nationally representative sample of over 18,000 individuals living in 5,000 families in the United States. Information on cohort individuals and their descendants has been collected continuously covering a range of socioeconomic information. Recently information on the health status, health condition and limitation, and emotional well-being and life satisfaction questions have been added to the survey. The health questions have been asked only for head of the household and spouse (e.g., wife). The life satisfaction question was added to the PSID survey only in 2009 and 2011. The exact wording of the question is as follows: “Please think about your life as a whole. How satisfied are you with it? Are you completely satisfied, very satisfied, somewhat satisfied, not very satisfied, or not at all satisfied?” with response choices “1= completely satisfied,….5= not at all satisfied.” Information on Annual family income before tax, diabetes diagnosis, date diabetes diagnosed, and other health conditions and disabilities are also recorded.

A total of 26,521 individuals were included in the sample. The unbalanced two-wave panel data has an average of 1.8 years for each variable record. Table C.2 and Table C.3 shows the descriptive statistics for variables in the study. The average age of individuals was 43 (range from 20-80). About 54% of the sample data are female and 45% are male. PSID data is based on family units; about 61% of the sample were married and about 39% were not married (about 22% never married, 11% widowed, 3% separated, and 3% widowed). There were no changes in the sample individual’s marital status between 2009 and 2011 survey period. About 56% of the individuals are white, 31% are African American, 11% are Hispanic, and 3% are other (Asians, American Indian or Alaska Natives, and Native Hawaiian or Pacific Islanders). The average family size is about three people, with maximum of 14 person per family unit. Respondents were relatively highly educated; about 57% completed some type of higher education, another 35% completed only high school, and about 8% did not complete high school. Overall, about 69% of the sample individuals were employed while 31% were not employed (retired, homemakers students, out of work force or temporarily unemployed). Of those who were employed in 2009, 91% were also employed on 2011. About 82% of those who were unemployed in 2009 were also unemployed in 2011. Average total family income is about $76,000, with maximum being about
six million and minimum being family business loss of about -$70,000. About 9% of respondents reported being diagnosed with diabetes, similar to the US population prevalence statistics (CDC, 2012). Of those people who reported no diabetes diagnosis during 2009, about 3% reported being diagnosed with diabetes in the 2011 survey wave. The majority of the respondents reported being satisfied with their life; about 22% were completely satisfied, about 45% very satisfied, about 29% not very satisfied, about 4% somewhat satisfied, and about 1% not satisfied at all.

The distribution of life satisfaction with diabetes status, gender, and ethnicity is presented in Table (4.1). The percentage of those who reported being diagnosed with diabetes decreased as the reported level of life satisfaction increased. Among those who said that they were not satisfied with their life at all, about 17% reported being diagnosed with diabetes, while only about 8% of very satisfied and completely satisfied respondents reported having diabetes. Males and females had a similar distribution of life satisfaction. A majority of the Whites (about 71%) were completely or very satisfied with life, while only 66% of Hispanics and 67% of African Americans reported being completely or very satisfied with life.

**Empirical Model Specification**

Following Dolan, Fujiwara & Matcalfe (2011) transform equation (4) into empirical econometric model as follows

\[ W = \beta_0 + \beta_1 \ln Y + \beta_2 H(\ln Y) + \beta_3 Emp(\ln Y) + \beta_4 Edu(\ln Y) + \beta_5 Marr(\ln Y) + \gamma Z + \epsilon \] (6)

\( W \) represents an individual’s subjective well-being (proxy for unknown utility function) which is measured by life satisfaction. \( H \) is dummy of 1 for presence of diabetes and 0 for absence in each period. \( \ln Y \) is log of per capita income (household income divided by household size). Income is in log to account for diminishing utility from income. As in Groot (2006) we assume income is linearly related to subjective well-being (\( W \)). \( Emp \) represents employment dummies. \( Edu \) represents highest completed education level. \( Marr \) represents dummies marriage. \( H(\ln Y), Emp(\ln Y), Edu(\ln Y), \) and \( Marr(\ln Y) \) denote the indirect effect of income through health, employment, education and marriage respectively. The presence of indirect effect of income is ascertained based on the auxiliary test of multicollinearity (see the estimation
strategy below for details). Z represents a matrix of demographic characteristics (i.e. age, sex, ethnicity(race) and number of children), which are assumed to be exogenous in the model.

The incidence of diabetes increases with age, thus age-squared was included in the model. Employment is included in the model with dummy (1=employed and 0=unemployed, combining the retired, homemakers, student, and unemployed together). Marriage is modeled as dummy 1 for married and 0 for non-married (divorced, separated, widowed and never married). Education included in the model as continuous variable, which is years of schooling.

It is shown that subjective wellbeing can be approximated with various instruments (Powdthavee & Van den Berg, 2011). These include life satisfaction, self-assessed general health, and mental well-being scores (such as General Health Questions, GHQ-12). This study utilizes life satisfaction as a dependent variable, measured as an ordered response with 5 scales (see Table C.1). Ferrer-i-Carbonell & Frijters (2004) showed that ethologically using the life satisfaction scale as cardinal or ordinal does not change the robustness of the estimated coefficients. Thus, the study used life satisfaction scale as cardinal (continuous dependent variable).

Finally, the monetary value of health loss (onset of diabetes) is calculated as the change in income that is required to retain a person’s life satisfaction at the similar position as before the onset of diabetes. The compensating income variation shown in equation (5) can be derived from empirical equation (6) as follows

\[ CIV = \frac{dY}{dH} = -\frac{\beta_2}{\beta_1} \left( \frac{1}{\gamma} + \beta_2 \frac{\partial H}{\partial Y} + \beta_3 \frac{\partial Emp}{\partial Y} + \beta_4 \frac{\partial Edu}{\partial Y} + \beta_5 \frac{\partial Marr}{\partial Y} \right) \]

Where \( \beta_2 \frac{\partial H}{\partial Y}, \beta_3 \frac{\partial Emp}{\partial Y}, \beta_4 \frac{\partial Edu}{\partial Y}, \) and \( \beta_5 \frac{\partial Marr}{\partial Y} \) are indirect effect of income and accounted for in the model during estimation (the steps are detailed in the estimation strategy section).

**Estimation Strategy**
The estimation strategy for the empirical model specified in the equation (6) is pursued using the Dolan, Fujiwara & Matcalfe (2011) “step” approach. The step approach estimation strategy helps to account for full instrumental value (direct and indirect effect) of income. The estimation follows three steps:

**Step1.**

Estimate the full subjective wellbeing function with income, health status and other control variables

\[ W_{it} = \beta_0 + \beta_1 \ln Y_{it} + \beta_2 H_{it} + \beta_3 Emp_{it} + \beta_4 Edu_{it} + \beta_5 Marr_{it} + \gamma_1 Age_{it} + \gamma_2 Age^2_{it} + \gamma_3 Sex_i + \gamma_4 Race_i + \gamma_5 NumChild_{it} + \epsilon_{it} \]  

(8)

Identify possible indirect effect of income on subjective wellbeing by testing the correlation between income and other independent variables (health status, employment status, education, and marriage). The auxiliary regression technique is used where income is regressed on the health status, employment status, education, and marriage as follows

\[ \ln Y_{it} = \beta_0 + \beta_2 H_{it} + \beta_3 Emp_{it} + \beta_4 Edu_{it} + \beta_5 Marr_{it} + \epsilon_{it} \]  

(9)

Compare \( R^2 \) from (8) and (9) to determine the collinearity between income and health status and other control variables. This technique commonly employed to detect multicollinearity between independent variables. The test for multicollinearity in auxiliary regression is Klein’s rule of thumb that if \( R^2 \) from auxiliary regression is larger than the original main regression (equation [8]), then income is correlated with health and other control variables. Thus, estimating the subjective wellbeing equation would not account for indirect effects of income. \( \beta_1 \) would therefore be biased downward. If no multicollinearity problem was found, then the estimation of subjective wellbeing specification in equation (8) is robust and we assume there is no indirect effect of income.

**Step2.**

If multicollinearity problem was found in step1, the estimation has to account for the indirect effect of income. This is accomplished by estimating K+1 subjective wellbeing regression; where K+1 is number of control variables influenced by income plus variable of
interest (health status) which is also influenced by income. The one regression for each of the control variables and one for health status will be estimated; while dropping one control variable from each regression at a time. The variables that are dropped are those that are correlated with income (employment, education, marriage, and health), therefore, dropping them would increase the instrumental value of income ($\beta_1$). Variables that are not affected by income (age, sex, and race) retained in the regressions. The four regressions (one for each variable that are related to income) are specified as follows:

i. Employment dropped

$$W_{it} = \beta_0 + \beta_1 \ln Y_{it} + \beta_2 H_{it} + \beta_4 Edu_{it} + \beta_5 Marr_{it} + \gamma_1 Age_{it} + \gamma_2 Age^2_{it} + \gamma_3 Sex_{i} + \gamma_4 Race_{i} + \gamma_5 NumChild_{it} + \varepsilon_{it}$$  

(10)

ii. Education dropped

$$W_{it} = \beta_0 + \beta_1 \ln Y_{it} + \beta_2 H_{it} + \beta_3 Emp_{it} + \beta_5 Marr_{it} + \gamma_1 Age_{it} + \gamma_2 Age^2_{it} + \gamma_3 Sex_{i} + \gamma_4 Race_{i} + \gamma_5 NumChild_{it} + \varepsilon_{it}$$  

(11)

iii. Marriage dropped

$$W_{it} = \beta_0 + \beta_1 \ln Y_{it} + \beta_2 H_{it} + \beta_3 Emp_{it} + \beta_4 Edu_{it} + \gamma_1 Age_{it} + \gamma_2 Age^2_{it} + \gamma_3 Sex_{i} + \gamma_4 Race_{i} + \gamma_5 NumChild_{it} + \varepsilon_{it}$$  

(12)

iv. Health status dropped

$$W_{it} = \beta_0 + \beta_1 \ln Y_{it} + \beta_2 H_{it} + \beta_3 Emp_{it} + \beta_4 Edu_{it} + \beta_5 Marr_{it} + \gamma_1 Age_{it} + \gamma_2 Age^2_{it} + \gamma_3 Sex_{i} + \gamma_4 Race_{i} + \gamma_5 NumChild_{it} + \varepsilon_{it}$$  

(13)

The indirect effect of income can be calculated by subtracting coefficient of income ($\beta_1$) in equation (8) from coefficients of income ($\beta_{1j}$) in equation (10) to (13) and added together.

$$\beta_1^* = \sum_{j=1}^{4} (\beta_{1j} - \beta_1)$$  

j= equation (10) to (13)  

(14)

Adding the indirect effect of income to the direct effect of income ($\beta_1^T = \beta_1^* + \beta_1$) gives the total instrumental value (effect) of income on subjective wellbeing.

Step 3.
Referring back to equation (7), basically $\beta_2 \frac{\partial H}{\partial Y}$, $\beta_3 \frac{\partial Emp}{\partial Y}$, $\beta_4 \frac{\partial Edu}{\partial Y}$ and $\beta_5 \frac{\partial Marr}{\partial Y}$ are incorporated in $\beta_1^T$. Thus, the compensating income variation is reduced to:

$$CIV = \frac{dY}{dH} = -\frac{\beta_2}{\beta_1^T} \cdot Y$$ (15)

**Choice of econometric method**

As discussed earlier, this study used two waves of the PSID longitudinal data set, since life satisfaction (the dependent variable) is only available only for two survey waves. This section identifies the appropriate panel data econometrics that would take advantage of the short panel data setting. Starting with a general panel data model:

$$w_{it} = \alpha_i + \beta x_{it} + \gamma z_i + \mu_i + \epsilon_{it}$$ (16)

Where $w_{it}$ is the dependent variable life satisfaction; $x_{it}$ is the vector of time-varying explanatory variables such as age, health status, employment status and income; $z_i$ is the vector of time-invariant explanatory variables such as sex and race; $\alpha$ is the intercept; $\beta$ is the vector of parameters for time varying explanatory variables; $\gamma$ is the vector of parameters for time-invariant explanatory variables; $\mu_i$ is random *individual-specific effect*; and $\epsilon_{it}$ is an *idiosyncratic error term*. The unobserved individual-specific effects are time-invariant and distributed randomly.

The method for estimating the above model entirely depends on the assumptions about individual-specific time-invariant effect. If $\mu_i$ is assumed to be correlated with observed explanatory variables, which is often the case in empirical analysis, Pooled OLS and Random Effect (RE) models estimations would be biased and inconsistent due to the potential problem of omitted variable (endogeneity). It violates the assumption that $\sum \eta_{it} = 0$ where $\eta_{it} = \mu_i + \epsilon_{it}$. A model with such property is estimated with Fixed Effect (FE) model, which assumes that unobserved heterogeneity ($\mu_i$) are constant overtime and correlated with observed explanatory variables (Mundlak, 1978).
However, the data at hand does not render itself to fixed effect (FE) model estimation for the following reasons. First, as can be seen from the descriptive statistics (in Table C.2) most of the explanatory variables have little or no variation over time. Second, the estimation involves time-invariant variables. As discussed earlier, gender and ethnicity are time-invariant variables in the wellbeing model. Although marital status is time-variant by definition, none of the individuals had status changes between the two waves. This study is interested in estimating the parameters for variables that did not change over time such as marital status to be used in subsequent analysis (note that the estimation involves three steps to derive total effect of income on life satisfaction). This is not possible through FE model. Third, some of the variables in the model are slow-changing by nature. The change in employment status, education level, number of children, and diabetes status over time was very small. The numbers of people whose state changed from non-diabetic to diabetic within three years period were small, thus, the power to detect the effect of diabetes on the life satisfaction from two wave data using FE model is slim. Though the time is very short, the only true time-variant variables are family income and age.

Therefore, this study had to make a tradeoff between the ability of FE model to control for omitted variable to estimate unbiased parameter for the power of RE model to detect significant statistical effect and being able to estimate parameter for time invariant observed explanatory variables. Hausman and Taylor (1981) offer a compromise between the FE and RE model. However, the Hausman-Taylor (HT) model forces the specifications of endogenous and exogenous variables. Reviewing relevant literature shows that most of the variables in this study might have reverse causality with life satisfaction. There is not enough justification to designate any of the variables as exogenous. Therefore, this study consciously chose to estimate the RE model with all its caveats.

4.3 Results

As discussed in the methodology section, the first-step of the estimation three-step estimation procedure is to estimate the wellbeing model with log per capita family income, status of diabetes, and all control variables as outlined in equation (8). As shown in Table (4.2), self-reported diabetic status is negatively associated with life satisfaction and log of per capita family income is positively associated with life satisfaction. Females are more satisfied with life
compared to their male counterparts. As expected, age demonstrated a U-shape relationship with life satisfaction. Age is negatively associated with life satisfaction, while age square is positively associated with life satisfaction. That is, older adults are more satisfied with life as compared to younger adults. Employment had the expected positive impact on the life satisfaction. Married adults are more satisfied with life compared to unmarried adults. Number of children is positively associated with life satisfaction. Education and ethnicity were not statistically significantly associated with life satisfaction.

According to Klein’s rule of thumb, if the $R^2$ from auxiliary regression is larger than the wellbeing regression, we know that income affects wellbeing indirectly through control variable such that we need to account for the indirect effects income via health, employment, education, and marriage. From Table (4.2), we can see that the $R^2$ from auxiliary regression is indeed larger than the full wellbeing regression. This shows that income is related to life satisfaction indirectly through health, employment, education, and marriage. Therefore, it is necessary to account for the indirect effect of income in the second-step estimations.

The second-step estimation involves estimation of series of wellbeing functions by dropping control variables that are correlated with income one at a time (equation 10 to 13). The results of the second-step estimations are displayed in Table (4.3). We can see that the magnitudes of the coefficients of income are improved when the control variables are dropped one at a time. This implies that instrumental value of income was improved by dropping variables through which household income influences wellbeing.

As outlined in equation (14), the indirect effect of income on the wellbeing can be calculated by subtracting the income coefficient in full wellbeing model (in Table 4.2, column 2) from the income coefficients in the models with one less control variable (in Table 4.3) and summing up the differences as shown below in Table (4.4).

The resulting total indirect effect of income on wellbeing through health, employment, education, and marriage is $\beta_1^* = 0.034$. Therefore, the full instrumental value of income (the total income effect) is obtained by adding the total indirect effect of income to the direct effect of income as follows:

$$\beta_1^\tau = \beta_1^* + \beta_1 = 0.034 + 0.033 = 0.067$$
After, the indirect effects of income through employment, education, marital status, and health are incorporated into the total effect, the instrumental value of income effect on subjective wellbeing has increased by about 103%. The third-step is computing the compensating income variation as in equation (15). The coefficient for health ($\beta_2$) is from Table (4.3) and total income effect is ($\beta_1^T$) as calculated above.

$$CIV = \frac{dY}{dH} = -\frac{\beta_2}{\beta_1^T} = -\left(-\frac{0.099}{0.067}\right) * Y = 1.459 * Y$$

At the sample median per capita income of $22,700, the CIV for diabetes is $33,110.

Following the same estimation procedure, CIV was also estimated by gender and age category. The detailed procedure is not reported here. Table (4.5) shows the estimated direct, indirect, and total income coefficients, health coefficient, and estimated CIV by gender and age class (less than 65 years and age greater than 65 years). The CIV for loss of health due to diabetes is larger for females compared to males. The CIV for females is $35,500 while it is $23,338 for males at the median per capita family income of $21,300 and $24,400 respectively. The effect of relative income on the older (age greater than 65) sample population is insignificant unlike younger sample population (age 20 to 65). On the other hand, health is significant for both younger and older populations. The CIV of the younger sample population at median per capita family income of $22,570 is $29,605 and CIV of the older sample population at median per capita family income of $24,046 is $60,616. This indicates that older populations value health at much more than younger populations. Alternatively, this result indicates that the effect of income on life satisfaction decreases with age and is relatively less valued among females.

To test the second hypothesis that the value of health (in this case diabetes) differs among racial/ethnic groups, the same three-step estimation procedure above was repeated for each racial/ethnic group. The estimation was repeated only for Whites, African Americans, and Hispanics since the sample of Asians, American Indian or Alaska Natives, and Native Hawaiian or Pacific Islanders was too small to run the estimation. The results of full wellbeing model estimations for the each Racial/ethnic group are presented in Table (4.6). Marital status has a strong and consistent positive impact on life satisfaction across the three racial/ethnic groups. Gender is not significant among Whites and Hispanics and age is not significant among African Americans. Surprisingly, number of children is not significant among African Americans, while
it is highly significant among Whites and Hispanics. Employment status is statistically significant in all three groups but the magnitude is larger for Black and Hispanic respondents. Diabetes has negative impact on life satisfaction of in all three groups, but not statistically significantly for Hispanics. Diabetes induces strong negative life satisfaction among the African Americans than Whites, but not significantly for Hispanics. Income is insignificant and small in magnitude for African Americans.

The results from second-step estimation, where control variables are dropped to obtain the estimation of the indirect income effects, are not presented here for brevity. Income coefficients from the models in the second-step of estimations and their difference from full wellbeing model are presented (Table 4.7). The second part of the table shows the calculated total indirect income effect, CIV multiplier and dollar value CIV. From Table (4.7), it is obvious that there is a large difference in CIV between African Americans and the other two racial/ethnic groups (Whites and Hispanics). The CIV value for Whites and Hispanics is similar and less than the general sample. As shown above the negative impact of diabetes on life satisfaction among African Americans is large, while the positive impact of income is insignificant. Therefore, the CIV for African Americans is very large.

4.4 Discussions

This study assessed the impact of diabetes on the subjective wellbeing and estimated monetary valuation for loss of good health due to diabetes for United States population. The study also compared the monetary valuation across gender and major ethnic groups in the United States. This study used the new “subjective wellbeing” method, as alternative to preference based approaches, with a three-step estimation approach (Fujiwara and Dolan, 2013). Accordingly, the monetary valuation for diabetes for the total population at median income is $33,109 annually. That is, a diabetic person needs a monetary compensation of $33,109 annually to feel as well off as a non-diabetic person. There is no established method to validate if the monetary value computed using SW method is appropriate or better than the one computed using preference methods. However, we can definitively report that the subjective wellbeing approach relies on the utility of the person experiencing disutility from loss of health due to diabetes. On the other hand, preference-based approaches such as quality adjusted life rely on the ex-ante
evaluation of the imagined disutility from health outcomes. It should be noted that willingness to pay using both approaches does not necessarily reflect the ability to pay, which is affordability of the service and bounded by household budget.

In addition, this study finds that females place higher valuation on loss of health due to diabetes as compared to their male counterparts. The finding is similar to Latif (2012) results for cardiovascular diseases among Canadian population. However, the finding is contradictory to the results of Groot and Maassen van den Brink (2006), where monetary valuation for loss of health due to cardiovascular diseases for males is greater than for females. The effect of income on the life satisfaction of older populations is not significant while the effect of both relative income and diabetes status is significant on the life satisfaction of older people. Thus, the valuation of diabetes increases at older age because social relationships (marriage) and health are more important than employment and relative income at this age, while adults under age 45 are more concerned about their income, employment, marriage and children as compared to their health.

Further, the analysis revealed variation in monetary valuation of diabetes across major ethnic groups in United States. African Americans have the highest monetary valuation followed by whites and Hispanics. The variation among the ethnic groups stems from the relative importance of income and diabetes for the life satisfaction of the individuals in the particular ethnic group. The evaluation of the relative importance of income and health is the direct result of socioeconomic and cultural differences among these ethnic groups. For example, log per capita family income is insignificant and small in magnitude for African Americans, while health, employment status, and marital status are relatively large and highly significant. This is not surprising that in a community with high unemployment and a high prevalence of diabetes, being employed, married, and having good health have high value. The effect of relative income on life satisfaction relative and if one lives in a poor community the effect is not significant (Firebaugh & Schroeder, 2009). On the other hand, African Americans not only have high prevalence of diabetes, but also suffer more from the disease complications. Diabetes in African Americans causes higher rates of blindness, amputation, End Stage Renal Diseases (ESRD) and death compared to both whites and Hispanics (Chow et al., 2012). Thus, fearing the complications of diabetes is accentuated during life evaluation.
On the contrary, the monetary valuation of diabetes among Hispanics is the smallest of the three ethnic groups. Although the prevalence of diabetes among Hispanics is just below that in African Americans, diabetes is not significantly associated with changes in life satisfaction for Hispanics. Rather, life satisfaction is associated with marital status, employment, and log per capita family income. The reason behind this result might be related to both socioeconomic status and culture. One study showed that Hispanic diabetic diabetes exhibit less knowledge and denial about diabetes complications due to lower education and English literacy (Bhargava et al., 2014). On the other hand, the strong family support system could mediate the impact of diabetes on life satisfaction (Phelps et al., 2012).

This study has several limitations. First, only two rounds of PSID survey data were available to assess the impact of health status on life satisfaction. Thus, the panel data is too short to control for fixed effects using the fixed effects panel data econometrics model. The econometric analysis applied for this study is the random effects model, which does not control for fixed individual effects such as genetics and personality differences. Second, the study utilizes a self-reported diabetes diagnosis. There may be measurement error in relation to the accuracy of the self-reported health status. Third, the direction of causality between income and life satisfaction and health and life satisfaction could be converse or third variables could be mediating the relationship. There is potential shortcoming since this study did not account for this endogeneity and confounding factors. Therefore, the inferences from this study should be interpreted with caution. Additional robust studies need to be conducted with longer panel data and with confirmed diabetes diagnosis status to confirm the results from this study. The methodology itself also need to be refined in finding good instrumental variables to resolve potential endogeneities of income and health status with life satisfaction.

A strength of this study is that it is not only the first study to use the SWB approach to quantify the monetary value of diabetes in the US, but it also highlights the heterogeneity among various ethnic groups within the same country. The study also improved the estimation of CIV by utilizing three-step approach of incorporating the indirect effect of income on life satisfaction. Note that one of the criticisms of the SWB approach to monetary valuation is that the monetary values tend to be too large to be considered for practical policy implications. The main reason for exaggerated monetary value estimation is that the direct effect of income on life satisfaction is
small when other variables are controlled for in the regression. This study overcomes the problem by extracting the indirect income effect on life satisfaction to improve the magnitude of total income effect on life satisfaction. There would be about 51% increase in the estimated monetary value (about $34,715 difference) if the three-step estimation was not followed to incorporate the indirect effect of income on subjective wellbeing through control variables.

4.5 Conclusions

Policy-makers routinely prioritize or make choices among programs to allocate scarce healthcare resources. In making such choices, decision-makers can benefit by knowing the monetary value of the diseases being prevented or cured by implementing effective intervention programs. Economic evaluations provide this information by comparing the benefit (value) and cost of public health interventions. While cost can be quantified by accounting for the costs incurred to develop and implement an intervention, including direct costs, indirect costs, and intangible costs, the value side of the equation is difficult to determine. Preference-based valuation methods, such as benefit-cost analysis, fail to reflect the lost utility of the person with the disease experience. The subjective wellbeing method was initiated in several European countries as alternative to estimate the value of particular health outcome including diabetes. Ferrer-i-Carbonell & van Praag (2002) and Groot & Van den Brink (2004) proposed monetary valuation from subjective wellbeing to be used for ranking the disease for prioritization and resource allocation in health systems as alternatives to monetization of Quality Adjusted Life Year (QALY). This study generates the empirical evidence of subjective wellbeing valuation of diabetes diseases in a United States population, particularly emphasizing the value of diabetes prevention. The study also provided evidence for the hypothesis that the monetary valuation of loss of wellbeing due to diabetes varies by gender, age group (young and older adults), and race/ethnic groups. The main reason for heterogeneity is the difference in relative importance of income, health and other domains of life satisfaction across culture, age and gender in evaluating one’s life satisfaction. Economic valuation of health needs to include these heterogeneous factors in different culture and strata of the society in determining the value that society attaches to a particular disease. The monetary value generated through this method should be viewed as total value for prevented diabetes cases. These values can be used as benefit side of the benefit-cost
analysis of diabetes prevention interventions as an alternative to preference-based willingness to pay approaches. Moreover, these results show that valuation of disease prevention can vary by ethnicity; therefore, it is crucial to develop culturally appropriate interventions tailored to the specified values and unique issues that face Americans of various ethnicities.
4.6 References


Finkelstein EA, Linnan LA, Tate DF, Birken BE. A pilot study testing the effect of different levels of financial incentives on weight loss among overweight employees. *J Occup Environ Med.* 2007; 49(9):981-989.


### Tables and Figures

#### Table 2.1. Distribution of the categorical variables for the study sample. (Based on NHANES, 2005—2012).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Category label</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Male</td>
<td>2,510</td>
<td>45.80</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>2,961</td>
<td>54.12</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>5,471</td>
<td></td>
</tr>
<tr>
<td>Race/Ethnicity</td>
<td>Mexican-American</td>
<td>984</td>
<td>17.99</td>
</tr>
<tr>
<td></td>
<td>Other Hispanics</td>
<td>572</td>
<td>10.46</td>
</tr>
<tr>
<td></td>
<td>White</td>
<td>2,668</td>
<td>48.77</td>
</tr>
<tr>
<td></td>
<td>Black</td>
<td>994</td>
<td>18.17</td>
</tr>
<tr>
<td></td>
<td>Others</td>
<td>253</td>
<td>4.62</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>5,471</td>
<td></td>
</tr>
<tr>
<td>Marital Status</td>
<td>Never married</td>
<td>1,072</td>
<td>19.67</td>
</tr>
<tr>
<td></td>
<td>Married or living with partner</td>
<td>3,181</td>
<td>58.14</td>
</tr>
<tr>
<td></td>
<td>Separated, divorced, or widowed</td>
<td>1,214</td>
<td>22.19</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>5,471</td>
<td></td>
</tr>
<tr>
<td>Household food security category</td>
<td>Full food secure</td>
<td>3,818</td>
<td>69.79</td>
</tr>
<tr>
<td></td>
<td>High marginal food security</td>
<td>592</td>
<td>10.82</td>
</tr>
<tr>
<td></td>
<td>Low food security</td>
<td>675</td>
<td>12.34</td>
</tr>
<tr>
<td></td>
<td>Very low food security</td>
<td>386</td>
<td>7.06</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>5,471</td>
<td></td>
</tr>
<tr>
<td>Education level</td>
<td>Less than high school</td>
<td>1,523</td>
<td>27.84</td>
</tr>
<tr>
<td></td>
<td>High school</td>
<td>1,326</td>
<td>24.24</td>
</tr>
<tr>
<td></td>
<td>College and higher</td>
<td>2,622</td>
<td>47.93</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>5,471</td>
<td></td>
</tr>
<tr>
<td>Cigarette smoking</td>
<td>Never smoker</td>
<td>3,804</td>
<td>69.53</td>
</tr>
<tr>
<td></td>
<td>Smoker</td>
<td>1,667</td>
<td>30.47</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>5,471</td>
<td></td>
</tr>
<tr>
<td>Alcohol consumption</td>
<td>Abstainers</td>
<td>841</td>
<td>15.37</td>
</tr>
<tr>
<td></td>
<td>Occasional drinkers</td>
<td>2,545</td>
<td>46.52</td>
</tr>
<tr>
<td></td>
<td>Moderate drinkers</td>
<td>901</td>
<td>16.47</td>
</tr>
<tr>
<td></td>
<td>Heavy drinkers</td>
<td>1,184</td>
<td>21.64</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>5,471</td>
<td></td>
</tr>
<tr>
<td>Do you have close relative who</td>
<td>No</td>
<td>3,321</td>
<td>60.70</td>
</tr>
<tr>
<td>have/had diabetes?</td>
<td>Yes</td>
<td>2,150</td>
<td>39.30</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>5,471</td>
<td></td>
</tr>
</tbody>
</table>
Table 2.2. Distribution of the continuous variables in the study sample. (Based on NHANES, 2005—2012).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Sample size</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>5471</td>
<td>46.75</td>
<td>17.40</td>
<td>20</td>
<td>80</td>
</tr>
<tr>
<td>BMI(Kg/m$^2$)</td>
<td>5471</td>
<td>27.51</td>
<td>5.57</td>
<td>13.18</td>
<td>57.93</td>
</tr>
<tr>
<td>Waist circumference (cm)</td>
<td>5471</td>
<td>95.04</td>
<td>13.98</td>
<td>59.70</td>
<td>166.4</td>
</tr>
<tr>
<td>Subscapular skinfold (mm)</td>
<td>5471</td>
<td>21.32</td>
<td>8.22</td>
<td>4.40</td>
<td>42</td>
</tr>
<tr>
<td>Ratio family income to poverty</td>
<td>5471</td>
<td>2.46</td>
<td>1.63</td>
<td>0.00</td>
<td>5</td>
</tr>
<tr>
<td>Depression mean score</td>
<td>5471</td>
<td>2.59</td>
<td>3.72</td>
<td>0.00</td>
<td>52</td>
</tr>
<tr>
<td>Moderate to vigorous physical activity (MVPA)</td>
<td>5471</td>
<td>101.70</td>
<td>157.12</td>
<td>0.00</td>
<td>1371.43</td>
</tr>
<tr>
<td>Diastolic blood pressure</td>
<td>5471</td>
<td>69.58</td>
<td>11.70</td>
<td>10.00</td>
<td>132</td>
</tr>
<tr>
<td>Systolic blood pressure</td>
<td>5471</td>
<td>121.84</td>
<td>18.31</td>
<td>78.00</td>
<td>228</td>
</tr>
<tr>
<td>Sleep hours</td>
<td>5471</td>
<td>6.81</td>
<td>1.45</td>
<td>1.00</td>
<td>12</td>
</tr>
</tbody>
</table>

Std.Dev= standard deviation
Table 2.3. Results of Additive MARS Model after Backward Selection. N=5471. (Based on NHANES, 2005—2012).

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Basis transformation</th>
<th>Variable description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5016</td>
<td>Intercept</td>
<td></td>
</tr>
<tr>
<td>0.05615</td>
<td>MAX(Age - 69, 0)</td>
<td>Age in years</td>
</tr>
<tr>
<td>-0.09302</td>
<td>MAX(69 - Age, 0)</td>
<td>Age in years</td>
</tr>
<tr>
<td>0.05310</td>
<td>MAX(WC - 69.9, 0)</td>
<td>Waist circumference (cm)</td>
</tr>
<tr>
<td>-0.9643</td>
<td>Diabetes Family = No</td>
<td>Family diabetes history (Yes/NO)</td>
</tr>
<tr>
<td>0.3683</td>
<td>Education = high school or less</td>
<td>Education level (less, equal, more than high school)</td>
</tr>
<tr>
<td>-0.04133</td>
<td>MAX(33 - SubSF, 0)</td>
<td>Subscapular skinfold</td>
</tr>
<tr>
<td>-0.04197</td>
<td>MAX(11 - DP score, 0)</td>
<td>Depression scale score (PHQ-9)</td>
</tr>
<tr>
<td>-0.3761</td>
<td>Ethnicity = White</td>
<td>Ethnicity (White, Black, Hispanic, Others)</td>
</tr>
<tr>
<td>0.05902</td>
<td>MAX(DBP - 78, 0)</td>
<td>Diastolic blood pressure (mmHg)</td>
</tr>
<tr>
<td>-0.07069</td>
<td>MAX(BMI - 21.59, 0)</td>
<td>Body mass index (kg/m²)</td>
</tr>
<tr>
<td>-0.02355</td>
<td>MAX(125 - SBP, 0)</td>
<td>Systolic blood pressure (mmHg)</td>
</tr>
<tr>
<td>-0.04447</td>
<td>MAX(DBP - 60, 0)</td>
<td>Diastolic blood pressure (mmHg)</td>
</tr>
<tr>
<td>-0.05006</td>
<td>MAX(Age - 51,0)</td>
<td>Age in years</td>
</tr>
<tr>
<td>0.002619</td>
<td>MAX(132.86 - MVPA, 0)</td>
<td>Moderate to vigorous physical activity (min/day)</td>
</tr>
<tr>
<td>0.08359</td>
<td>MAX(7 - Sleep Hours, 0)</td>
<td>Sleep duration (Hours)</td>
</tr>
<tr>
<td>0.3716</td>
<td>MAX(BMI - 49.82, 0)</td>
<td>Body mass index (kg/m²)</td>
</tr>
<tr>
<td>0.01692</td>
<td>MAX(SBP - 157.33, 0)</td>
<td>Systolic blood pressure (mmHg)</td>
</tr>
<tr>
<td>-0.05680</td>
<td>MAX(RFItpov - 0.87, 0)</td>
<td>Ratio family of income to poverty</td>
</tr>
</tbody>
</table>
Table 2.4. Results of Two-way interaction MARS Model after Backward Selection (N=5471, based on NHANES 2005—2012).

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Basis transformation</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.9982</td>
<td>Intercept</td>
</tr>
<tr>
<td>0.05965</td>
<td>MAX(WC - 69.9, 0)</td>
</tr>
<tr>
<td>-0.4956</td>
<td>Diabetes Family = No</td>
</tr>
<tr>
<td>0.3681</td>
<td>Education = High school or less</td>
</tr>
<tr>
<td>-0.00187</td>
<td>MAX(WC - 69.9, 0) * MAX(27.8 - SubSF, 0)</td>
</tr>
<tr>
<td>0.02721</td>
<td>MAX(69 - Age, 0) * MAX(SBP - 192, 0)</td>
</tr>
<tr>
<td>-0.00097</td>
<td>MAX(69 - Age, 0) * MAX(192 - SBP, 0)</td>
</tr>
<tr>
<td>-0.8134</td>
<td>Ethnicity = White</td>
</tr>
<tr>
<td>-0.07918</td>
<td>MAX(11 – DP score, 0)</td>
</tr>
<tr>
<td>-0.06775</td>
<td>MAX(BMI - 21.88, 0)</td>
</tr>
<tr>
<td>-0.2599</td>
<td>MAX(21.88 - BMI, 0)</td>
</tr>
<tr>
<td>0.004003</td>
<td>MAX(11 – DP score, 0) * MAX(77 - DBP, 0)</td>
</tr>
<tr>
<td>-0.08892</td>
<td>(Ethnicity = White) * MAX(57 - DBP, 0)</td>
</tr>
<tr>
<td>0.2324</td>
<td>(Ethnicity = White) * MAX(7 - Sleep Hours, 0)</td>
</tr>
<tr>
<td>-0.9036</td>
<td>(Education = college) * (Alcohol Consumption = Heavy)</td>
</tr>
<tr>
<td>0.002487</td>
<td>(Diabetes Family = Yes) * MAX(233.57 - MVPA, 0)</td>
</tr>
<tr>
<td>-0.02466</td>
<td>(Ethnicity not = White) * MAX(Age - 53, 0)</td>
</tr>
<tr>
<td>-0.03044</td>
<td>(Ethnicity not = White) * MAX(53 - Age, 0)</td>
</tr>
<tr>
<td>0.007556</td>
<td>MAX(Age - 69, 0) * MAX(SBP - 169, 0)</td>
</tr>
<tr>
<td>-0.01306</td>
<td>MAX(BMI - 21.88, 0) * MAX(SBP - 175, 0)</td>
</tr>
</tbody>
</table>
Table 3.1. Descriptive statistics summarizing diabetes-related mortality rate and explanatory variables (N=3109 contiguous US counties, aggregated data span from 2003 to 2010).

<table>
<thead>
<tr>
<th>Variables</th>
<th>Mean (Std. Err.) [95% Conf. Interval]</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRMR per 100,000</td>
<td>60.62 (0.57)[60.01, 62.25]</td>
<td>31.86</td>
<td>5.84</td>
<td>229</td>
<td>NA</td>
</tr>
<tr>
<td>Spatial Proximity Index (SPI)</td>
<td>1.06(0.00)[1.05,1.06]</td>
<td>0.09</td>
<td>0.00</td>
<td>1.68</td>
<td>1.27</td>
</tr>
<tr>
<td>Proportion of Hispanic Population</td>
<td>0.08(0.00)[0.07,0.08]</td>
<td>0.13</td>
<td>0.00</td>
<td>0.99</td>
<td>2.03</td>
</tr>
<tr>
<td>Proportion of Black Population</td>
<td>0.09(0.00)[0.08,0.09]</td>
<td>0.14</td>
<td>0.00</td>
<td>0.87</td>
<td>1.59</td>
</tr>
<tr>
<td>Socioeconomic index</td>
<td>0.00(0.02)[0.03,0.03]</td>
<td>0.94</td>
<td>-2.90</td>
<td>4.62</td>
<td>2.21</td>
</tr>
<tr>
<td>GINI coefficient (Income Disparity)</td>
<td>0.43(0.00)[0.43,0.43]</td>
<td>0.04</td>
<td>0.21</td>
<td>0.65</td>
<td>1.67</td>
</tr>
<tr>
<td>No. of Hospital beds per 1,000</td>
<td>3.56 (0.09)[3.38,3.73]</td>
<td>4.96</td>
<td>0.00</td>
<td>68.64</td>
<td>1.37</td>
</tr>
<tr>
<td>No. of Active MD per 10,000</td>
<td>1.25(0.03)[1.20,1.31]</td>
<td>1.49</td>
<td>0.00</td>
<td>26.94</td>
<td>1.70</td>
</tr>
<tr>
<td>Proportion of uninsured Adults</td>
<td>0.64(0.00)[0.64,0.64]</td>
<td>0.04</td>
<td>0.00</td>
<td>0.73</td>
<td>2.19</td>
</tr>
<tr>
<td>Social Capital Index (SCI)</td>
<td>0.00(0.02) [-0.05,0.05]</td>
<td>1.34</td>
<td>-3.86</td>
<td>15.97</td>
<td>1.87</td>
</tr>
<tr>
<td>Proportion of Adults meets PA guideline</td>
<td>0.21(0.00)[0.21,0.21]</td>
<td>0.06</td>
<td>0.07</td>
<td>0.51</td>
<td>1.25</td>
</tr>
<tr>
<td>Retail food Environment Index (RFEI)</td>
<td>0.63(0.01)[0.62,0.64]</td>
<td>0.30</td>
<td>0.00</td>
<td>2.29</td>
<td>1.59</td>
</tr>
<tr>
<td>Percent Low Access to Grocery Stores</td>
<td>3.03(0.04)[2.95,3.11]</td>
<td>2.18</td>
<td>0.00</td>
<td>29.51</td>
<td>1.85</td>
</tr>
<tr>
<td>Proportion of urban residents</td>
<td>0.40(0.01)[0.39,0.41]</td>
<td>0.31</td>
<td>0.00</td>
<td>1.00</td>
<td>2.84</td>
</tr>
</tbody>
</table>

VIF=Variance Inflation Factor
Table 3.2. Global bivariate Moran’s- $I$: global association between DRMR and neighboring county socioecological factors (N=3109 contiguous US counties, aggregated data span from 2003 to 2010).

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Lagged independent variables</th>
<th>Bivariate Moran’s $I$</th>
<th>Pseudo p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log of DRMR per 100, 000</td>
<td>Log of DRMR per 100, 000</td>
<td>0.373</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>Log of DRMR per 100, 000</td>
<td>Spatial Proximity Index (SPI)</td>
<td>0.036</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>Log of DRMR per 100, 000</td>
<td>Proportion of Hispanic Population</td>
<td>0.021</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>Log of DRMR per 100, 000</td>
<td>Proportion of Black Population</td>
<td>0.219</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>Log of DRMR per 100, 000</td>
<td>Socioeconomic index</td>
<td>-0.305</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>Log of DRMR per 100, 000</td>
<td>GINI coefficient (Income Disparity)</td>
<td>0.161</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>Log of DRMR per 100, 000</td>
<td>No. of Hospital beds per 1,000</td>
<td>-0.029</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>Log of DRMR per 100, 000</td>
<td>No. of Active MD per 10,000</td>
<td>-0.079</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>Log of DRMR per 100, 000</td>
<td>Proportion of uninsured Adults</td>
<td>0.147</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>Log of DRMR per 100, 000</td>
<td>Social Capital Index (SCI)</td>
<td>-0.159</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>Log of DRMR per 100, 000</td>
<td>Proportion of Adults meets PA guideline</td>
<td>-0.214</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>Log of DRMR per 100, 000</td>
<td>Retail food Environment Index (RFEI)</td>
<td>0.122</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>Log of DRMR per 100, 000</td>
<td>Percent Low Access to Grocery Stores</td>
<td>0.224</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>Log of DRMR per 100, 000</td>
<td>Rural-Urban Continuum Index</td>
<td>-0.059</td>
<td>&lt;0.05</td>
</tr>
</tbody>
</table>
Table 3.3. Results of multinomial logistic regression to depict the association between DRMR clusters and socioecological factors (N=3109 contiguous US counties, aggregated data span from 2003 to 2010).

<table>
<thead>
<tr>
<th>Variables</th>
<th>Type of clusters</th>
<th>High-High</th>
<th>Low-Low</th>
<th>Low-High</th>
<th>High-Low</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social Capital Index (SKI)</td>
<td></td>
<td>1.06</td>
<td>0.97</td>
<td>1.19*</td>
<td>1.19*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.94,1.19]</td>
<td>[0.85,1.10]</td>
<td>[1.00,1.42]</td>
<td>[1.01,1.41]</td>
</tr>
<tr>
<td>Socioeconomic index (SEI)</td>
<td></td>
<td>0.48***</td>
<td>2.63***</td>
<td>1.01</td>
<td>1.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.38,0.60]</td>
<td>[2.19,3.17]</td>
<td>[0.66,1.55]</td>
<td>[0.79,1.55]</td>
</tr>
<tr>
<td>Income disparity(GINI coefficient)</td>
<td></td>
<td>0.03</td>
<td>7.23</td>
<td>4798.23*</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.00,2.25]</td>
<td>[0.10,545.87]</td>
<td>[3.20,7.19×10^6]</td>
<td>[0.00,3.25]</td>
</tr>
<tr>
<td>Percent with low access to grocery store</td>
<td></td>
<td>1.06</td>
<td>0.79***</td>
<td>0.96</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[1.00,1.13]</td>
<td>[0.70,0.89]</td>
<td>[0.83,1.10]</td>
<td>[0.79,1.06]</td>
</tr>
<tr>
<td>Retail Food Environment Index (RFEI)</td>
<td></td>
<td>4.52***</td>
<td>0.48*</td>
<td>2.54*</td>
<td>0.28*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[2.92,7.01]</td>
<td>[0.27,0.87]</td>
<td>[1.03,6.26]</td>
<td>[0.10,0.82]</td>
</tr>
<tr>
<td>Spatial Proximity Index (SPI)</td>
<td></td>
<td>0.12**</td>
<td>3.46</td>
<td>1.92</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.03,0.57]</td>
<td>[0.42,28.48]</td>
<td>[0.10,36.52]</td>
<td>[0.02,1.02]</td>
</tr>
<tr>
<td>Percent of uninsured adults</td>
<td></td>
<td>1.05***</td>
<td>1.04**</td>
<td>1.07*</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[1.02,1.08]</td>
<td>[1.01,1.08]</td>
<td>[1.02,1.13]</td>
<td>[0.95,1.05]</td>
</tr>
<tr>
<td>No. of hospital beds per 1000</td>
<td></td>
<td>0.98</td>
<td>1.02</td>
<td>1.01</td>
<td>1.04*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.95,1.01]</td>
<td>[1.00,1.05]</td>
<td>[0.96,1.06]</td>
<td>[1.00,1.07]</td>
</tr>
<tr>
<td>No. of active physicians per 1000</td>
<td></td>
<td>0.85</td>
<td>0.94</td>
<td>0.64</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.50,1.46]</td>
<td>[0.63,1.39]</td>
<td>[0.25,1.67]</td>
<td>[0.45,1.64]</td>
</tr>
<tr>
<td>Percent Whites</td>
<td></td>
<td>0.97***</td>
<td>1.06**</td>
<td>0.98</td>
<td>0.95***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.96,0.99]</td>
<td>[1.02,1.11]</td>
<td>[0.95,1.01]</td>
<td>[0.93,0.98]</td>
</tr>
<tr>
<td>Percent of African Americans</td>
<td></td>
<td>1.00</td>
<td>1.00</td>
<td>0.99</td>
<td>0.97*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.98,1.01]</td>
<td>[0.95,1.04]</td>
<td>[0.96,1.02]</td>
<td>[0.94,1.00]</td>
</tr>
<tr>
<td>Percent of Hispanics</td>
<td></td>
<td>0.98**</td>
<td>1.04</td>
<td>0.98</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.96,0.99]</td>
<td>[1.00,1.09]</td>
<td>[0.95,1.02]</td>
<td>[0.95,1.00]</td>
</tr>
<tr>
<td>Percent adults meet PA guideline</td>
<td></td>
<td>0.93***</td>
<td>1.08***</td>
<td>0.95</td>
<td>1.08*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.91,0.96]</td>
<td>[1.04,1.12]</td>
<td>[0.89,1.01]</td>
<td>[1.01,1.15]</td>
</tr>
<tr>
<td>percent urban population</td>
<td></td>
<td>1.28</td>
<td>1.55</td>
<td>0.38</td>
<td>1.51</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.66,2.46]</td>
<td>[0.78,3.08]</td>
<td>[0.11,1.37]</td>
<td>[0.50,4.60]</td>
</tr>
<tr>
<td>Constant</td>
<td></td>
<td>409.08***</td>
<td>0.00***</td>
<td>0.02</td>
<td>3.64</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[9.83,17032.28]</td>
<td>[0.00,0.00]</td>
<td>[0.00,29.08]</td>
<td>[0.00,2971.29]</td>
</tr>
</tbody>
</table>

* p < 0.05, ** p < 0.01, *** p < 0.001. 95% confidence intervals in brackets. The multinomial logistic odds ratio is in comparison to non-significant cluster (base model).
Table 3.4. Diagnostic tests to determine if Spatial Autoregressive (SAR) and/or Spatial Error Model (SEM) are better representations of the underlying spatial dependence in DRMR data (N=3109 contiguous US counties, aggregated data span from 2003 to 2010).

<table>
<thead>
<tr>
<th>Models</th>
<th>Spatial Weight matrix</th>
<th>Test Type</th>
<th>Test Statistics</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatial Autoregressive (SAR)</td>
<td>Queen Order 1</td>
<td>Lagrange Multiplier</td>
<td>431.95</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Robust Lagrange Multiplier</td>
<td>10.21</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Spatial error Model (SEM)</td>
<td>Queen Order 1</td>
<td>Lagrange Multiplier</td>
<td>624.28</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Robust Lagrange Multiplier</td>
<td>202.54</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>
Table 3.5. Estimation results from SAR, SEM and SDM regression models for association between socioecological factors and DRMR (N=3109 contiguous US counties, aggregated data span from 2003 to 2010)

<table>
<thead>
<tr>
<th>Explanatory Variables</th>
<th>SEM</th>
<th>SAR</th>
<th>SDM</th>
<th>SDM Lag</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(Std. error)</td>
<td>(Std. error)</td>
<td>(Std. error)</td>
<td>(Std. error)</td>
</tr>
<tr>
<td>Social capital index (SCI)</td>
<td>0.002</td>
<td>0.010</td>
<td>0.006</td>
<td>0.023*</td>
</tr>
<tr>
<td></td>
<td>(0.006)</td>
<td>(0.005)</td>
<td>(0.007)</td>
<td>(0.010)</td>
</tr>
<tr>
<td>Socioeconomic index (SEI)</td>
<td>-0.270***</td>
<td>-0.216***</td>
<td>-0.289***</td>
<td>0.137***</td>
</tr>
<tr>
<td></td>
<td>(0.010)</td>
<td>(0.009)</td>
<td>(0.012)</td>
<td>(0.018)</td>
</tr>
<tr>
<td>Income disparity (Gini coefficient)</td>
<td>-1.205***</td>
<td>-1.245***</td>
<td>-1.064***</td>
<td>-1.011**</td>
</tr>
<tr>
<td></td>
<td>(0.193)</td>
<td>(0.194)</td>
<td>(0.194)</td>
<td>(0.358)</td>
</tr>
<tr>
<td>Retail food environment index (RFEI)</td>
<td>0.075**</td>
<td>0.084***</td>
<td>0.076**</td>
<td>0.113**</td>
</tr>
<tr>
<td></td>
<td>(0.024)</td>
<td>(0.023)</td>
<td>(0.024)</td>
<td>(0.043)</td>
</tr>
<tr>
<td>% HH with low access to grocery stores</td>
<td>0.006</td>
<td>0.006</td>
<td>0.008*</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>(0.003)</td>
<td>(0.003)</td>
<td>(0.003)</td>
<td>(0.007)</td>
</tr>
<tr>
<td>No. of hospital beds per 1000 pop</td>
<td>0.0004</td>
<td>0.001</td>
<td>0.001</td>
<td>-0.002</td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
<td>(0.001)</td>
<td>(0.001)</td>
<td>(0.003)</td>
</tr>
<tr>
<td>No. of active physicians per 1000</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.001)</td>
</tr>
<tr>
<td>Proportion of uninsured Adults</td>
<td>-0.467*</td>
<td>-0.441**</td>
<td>-0.609**</td>
<td>0.575</td>
</tr>
<tr>
<td></td>
<td>(0.181)</td>
<td>(0.130)</td>
<td>(0.234)</td>
<td>(0.288)</td>
</tr>
<tr>
<td>Proportion of black population</td>
<td>0.183**</td>
<td>0.179***</td>
<td>-0.005</td>
<td>0.197</td>
</tr>
<tr>
<td></td>
<td>(0.070)</td>
<td>(0.048)</td>
<td>(0.099)</td>
<td>(0.114)</td>
</tr>
<tr>
<td>Proportion of hispanic population</td>
<td>0.055</td>
<td>0.153*</td>
<td>-0.110</td>
<td>0.316*</td>
</tr>
<tr>
<td></td>
<td>(0.086)</td>
<td>(0.061)</td>
<td>(0.114)</td>
<td>(0.135)</td>
</tr>
<tr>
<td>Spatial proximity index (SPI)</td>
<td>0.186**</td>
<td>0.195**</td>
<td>0.198**</td>
<td>-0.214</td>
</tr>
<tr>
<td></td>
<td>(0.070)</td>
<td>(0.069)</td>
<td>(0.071)</td>
<td>(0.127)</td>
</tr>
<tr>
<td>Proportion of adults meets PA guideline</td>
<td>-0.461*</td>
<td>-0.126</td>
<td>0.358</td>
<td>-0.923***</td>
</tr>
<tr>
<td></td>
<td>(0.211)</td>
<td>(0.139)</td>
<td>(0.266)</td>
<td>(0.262)</td>
</tr>
<tr>
<td>% of urban population</td>
<td>0.178***</td>
<td>0.182***</td>
<td>0.179***</td>
<td>-0.051</td>
</tr>
<tr>
<td></td>
<td>(0.031)</td>
<td>(0.030)</td>
<td>(0.031)</td>
<td>(0.057)</td>
</tr>
<tr>
<td>Intercept</td>
<td>4.583***</td>
<td>2.883***</td>
<td>3.156***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.174)</td>
<td>(0.168)</td>
<td>(0.203)</td>
<td></td>
</tr>
<tr>
<td>Lambda (λ)</td>
<td>0.512***</td>
<td>0.366***</td>
<td>0.469***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.022)</td>
<td>(0.019)</td>
<td>(0.021)</td>
<td></td>
</tr>
</tbody>
</table>

Significance levels: ***< 0.001, ** <0.01, & * <0.05.
Table 3.6. Direct and indirect impact of change in independent variables on county DRMR (N=3109 contiguous US counties, aggregated data span from 2003 to 2010).

<table>
<thead>
<tr>
<th>Explanatory variables</th>
<th>SDM impacts</th>
<th></th>
<th></th>
<th>% of indirect impacts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>Indirect</td>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>Social capital index (SCI)</td>
<td>0.009</td>
<td>0.045**</td>
<td>0.054**</td>
<td>84.08</td>
</tr>
<tr>
<td>Socioeconomic index (SEI)</td>
<td>-0.288***</td>
<td>0.003</td>
<td>-0.286***</td>
<td></td>
</tr>
<tr>
<td>Income disparity (Gini coefficient)</td>
<td>-1.222***</td>
<td>-2.687***</td>
<td>-3.908***</td>
<td>68.75</td>
</tr>
<tr>
<td>Retail food environment index (RFEI)</td>
<td>0.091***</td>
<td>0.264***</td>
<td>0.355***</td>
<td>74.37</td>
</tr>
<tr>
<td>Percent HH with low access to grocery stores</td>
<td>0.008*</td>
<td>0.013</td>
<td>0.022</td>
<td></td>
</tr>
<tr>
<td>No. of hospital beds per 1000 pop</td>
<td>-0.001</td>
<td>-0.004</td>
<td>-0.004</td>
<td></td>
</tr>
<tr>
<td>No. of active physicians per 1000</td>
<td>0.001</td>
<td>0.005*</td>
<td>0.006*</td>
<td>82.62</td>
</tr>
<tr>
<td>Proportion of uninsured Adults</td>
<td>-0.579*</td>
<td>0.515</td>
<td>-0.064</td>
<td></td>
</tr>
<tr>
<td>Proportion of African Americans</td>
<td>0.015</td>
<td>0.347*</td>
<td>0.362**</td>
<td>95.85</td>
</tr>
<tr>
<td>Proportion of Hispanics</td>
<td>-0.082</td>
<td>0.471**</td>
<td>0.389**</td>
<td>85.17</td>
</tr>
<tr>
<td>Spatial proximity index (SPI)</td>
<td>0.186**</td>
<td>-0.215</td>
<td>-0.030</td>
<td></td>
</tr>
<tr>
<td>Proportion of adults meets PA guideline</td>
<td>0.279</td>
<td>-1.344***</td>
<td>-1.066***</td>
<td>82.81</td>
</tr>
<tr>
<td>Percent of urban population</td>
<td>0.183***</td>
<td>0.060</td>
<td>0.243*</td>
<td></td>
</tr>
</tbody>
</table>

Significance levels: ***< 0.001, ** <0.01, & * <0.05.
Table 4.1. Distribution Life satisfaction across diabetes status, gender, and Ethnicity (based on PSID data, 2009 and 2011).

<table>
<thead>
<tr>
<th>Life satisfaction</th>
<th>Diabetes status</th>
<th>Gender</th>
<th>Ethnicity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Nondiabetic</td>
<td>Diabetic</td>
<td>Male</td>
</tr>
<tr>
<td>Not at all satisfied</td>
<td>5,200 (21.81%)</td>
<td>473 (20.18%)</td>
<td>120</td>
</tr>
<tr>
<td>Somewhat satisfied</td>
<td>10,730 (45.00%)</td>
<td>956 (40.78%)</td>
<td>410</td>
</tr>
<tr>
<td>Not very satisfied</td>
<td>6,810 (28.56%)</td>
<td>738 (31.48%)</td>
<td>3,030</td>
</tr>
<tr>
<td>Very satisfied</td>
<td>851 (3.57%)</td>
<td>127 (5.42%)</td>
<td>5,159</td>
</tr>
<tr>
<td>Completely satisfied</td>
<td>252 (1.06%)</td>
<td>50 (2.13%)</td>
<td>2,398</td>
</tr>
</tbody>
</table>

Number in parentheses is column percentage.
Table 4.2. Estimation results of wellbeing function using RE panel data models (based on PSID data, 2009 and 2011).

<table>
<thead>
<tr>
<th>Variables</th>
<th>Full wellbeing model (Random Effect)</th>
<th>Auxiliary regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender (Female)</td>
<td>0.0336**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0124)</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>-0.0236***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.00279)</td>
<td></td>
</tr>
<tr>
<td>Age Square</td>
<td>0.000281***</td>
<td>0.468***</td>
</tr>
<tr>
<td></td>
<td>(0.00003)</td>
<td>(0.0184)</td>
</tr>
<tr>
<td>Marital status (Married)</td>
<td>0.366***</td>
<td>0.407***</td>
</tr>
<tr>
<td></td>
<td>(0.0146)</td>
<td>(0.0192)</td>
</tr>
<tr>
<td>Employment status (employed)</td>
<td>0.112***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0142)</td>
<td></td>
</tr>
<tr>
<td>Ethnicity (Black)</td>
<td>-0.0297</td>
<td>0.0117</td>
</tr>
<tr>
<td></td>
<td>(0.0155)</td>
<td>(0.0218)</td>
</tr>
<tr>
<td>Ethnicity (Hispanic)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ethnicity (all others)</td>
<td>-0.0744</td>
<td>0.0171</td>
</tr>
<tr>
<td></td>
<td>(0.0412)</td>
<td>(0.0218)</td>
</tr>
<tr>
<td>Diabetes (Yes)</td>
<td>-0.0986***</td>
<td>0.0414</td>
</tr>
<tr>
<td></td>
<td>(0.0227)</td>
<td>(0.0270)</td>
</tr>
<tr>
<td>Log per capita family income</td>
<td>0.0332***</td>
<td>0.0494</td>
</tr>
<tr>
<td></td>
<td>(0.00706)</td>
<td>(0.0461)</td>
</tr>
<tr>
<td>Education level (High school graduate)</td>
<td>0.0243</td>
<td>0.555***</td>
</tr>
<tr>
<td></td>
<td>(0.0326)</td>
<td>(0.0461)</td>
</tr>
<tr>
<td>Education level (College graduate or more)</td>
<td>0.0561</td>
<td>1.035***</td>
</tr>
<tr>
<td></td>
<td>(0.0324)</td>
<td>(0.0447)</td>
</tr>
<tr>
<td>Number of Children in the household</td>
<td>0.0391***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.00602)</td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>2.545***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0866)</td>
<td></td>
</tr>
<tr>
<td>Overall R²</td>
<td>0.0667</td>
<td>0.1790</td>
</tr>
<tr>
<td>N</td>
<td>24380</td>
<td>24516</td>
</tr>
</tbody>
</table>

Standard errors in parentheses; * p < 0.05, ** p < 0.01, *** p < 0.001
Table 4.3 Results from wellbeing model estimation with one control variable dropped at a time (Based on PSID data, 2009 and 2011).

<table>
<thead>
<tr>
<th>Variables</th>
<th>Wellbeing model with a control variable dropped one at a time (random effect coefficients)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Employment dropped</td>
</tr>
<tr>
<td>Gender (Female)</td>
<td>0.0230</td>
</tr>
<tr>
<td></td>
<td>(0.0124)</td>
</tr>
<tr>
<td>Age</td>
<td>-0.0199***</td>
</tr>
<tr>
<td></td>
<td>(0.00275)</td>
</tr>
<tr>
<td>Age Square</td>
<td>0.000231***</td>
</tr>
<tr>
<td></td>
<td>(0.0000293)</td>
</tr>
<tr>
<td>Marital status (Married)</td>
<td>0.367***</td>
</tr>
<tr>
<td></td>
<td>(0.0146)</td>
</tr>
<tr>
<td>Employment status (employed)</td>
<td>0.115***</td>
</tr>
<tr>
<td></td>
<td>(0.0142)</td>
</tr>
<tr>
<td>Ethnicity (Black)</td>
<td>-0.0311*</td>
</tr>
<tr>
<td></td>
<td>(0.0155)</td>
</tr>
<tr>
<td>Ethnicity (Hispanic)</td>
<td>0.0148</td>
</tr>
<tr>
<td></td>
<td>(0.0218)</td>
</tr>
<tr>
<td>Ethnicity (all others)</td>
<td>-0.0767</td>
</tr>
<tr>
<td></td>
<td>(0.0411)</td>
</tr>
<tr>
<td>Diabetes (Yes)</td>
<td>-0.106***</td>
</tr>
<tr>
<td></td>
<td>(0.0227)</td>
</tr>
<tr>
<td>Log of per capita family income</td>
<td>0.0430***</td>
</tr>
<tr>
<td></td>
<td>(0.00704)</td>
</tr>
<tr>
<td>Education level (High school graduate)</td>
<td>0.0359</td>
</tr>
<tr>
<td></td>
<td>(0.0325)</td>
</tr>
<tr>
<td>Education level (College graduate or more)</td>
<td>0.0743*</td>
</tr>
<tr>
<td></td>
<td>(0.0323)</td>
</tr>
<tr>
<td>Number of Children in the household</td>
<td>0.0386***</td>
</tr>
<tr>
<td></td>
<td>(0.00603)</td>
</tr>
<tr>
<td>Constant</td>
<td>2.464***</td>
</tr>
<tr>
<td></td>
<td>(0.0864)</td>
</tr>
<tr>
<td>N</td>
<td>24380</td>
</tr>
</tbody>
</table>

Standard errors in parentheses; * p < 0.05, ** p < 0.01, *** p < 0.001
Table 4.4. Computing total indirect income effect (Based on PSID data, 2009 and 2011).

<table>
<thead>
<tr>
<th>Variable dropped</th>
<th>Income coefficient</th>
<th>Indirect income effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Employment status</td>
<td>0.04</td>
<td>0.009</td>
</tr>
<tr>
<td>Education level</td>
<td>0.04</td>
<td>0.004</td>
</tr>
<tr>
<td>Marital status</td>
<td>0.05</td>
<td>0.02</td>
</tr>
<tr>
<td>Health (Diabetes status)</td>
<td>0.03</td>
<td>0.0007</td>
</tr>
<tr>
<td>Full model</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>Total indirect income effect ($\beta_1^*$)</td>
<td></td>
<td>0.03</td>
</tr>
</tbody>
</table>
Table 4.5. The estimated direct, indirect, and total income coefficients, health coefficient, and estimated CIV by gender and age class (less than 65 years and age greater than 65 years) (based on PSID data, 2009 and 2011).

<table>
<thead>
<tr>
<th>Dropped variable</th>
<th>Gender</th>
<th>Age</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male</td>
<td>Female</td>
<td>65 and younger</td>
<td>Above 65</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Direct income effect</td>
<td>Indirect income effects</td>
<td>Direct income effect</td>
<td>Indirect income effects</td>
<td>Direct income effect</td>
<td>Indirect income effects</td>
<td>Direct income effect</td>
</tr>
<tr>
<td>Employment status</td>
<td>0.05</td>
<td>0.02</td>
<td>0.04</td>
<td>0.006</td>
<td>0.04</td>
<td>0.007</td>
<td>0.03</td>
</tr>
<tr>
<td>education level</td>
<td>0.04</td>
<td>0.006</td>
<td>0.03</td>
<td>0.002</td>
<td>0.04</td>
<td>0.002</td>
<td>0.04</td>
</tr>
<tr>
<td>Marital status</td>
<td>0.05</td>
<td>0.01</td>
<td>0.07</td>
<td>0.04</td>
<td>0.07</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>Health (diabetes)</td>
<td>0.04</td>
<td>0</td>
<td>0.03</td>
<td>0.001</td>
<td>0.03</td>
<td>0</td>
<td>0.03</td>
</tr>
<tr>
<td>Indirect income effect</td>
<td>0.04</td>
<td>0.05</td>
<td>0.04</td>
<td>0.004</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>Direct income effect</td>
<td>0.04</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>Total income effect</td>
<td>0.08</td>
<td>0.08</td>
<td>0.08</td>
<td>0.07</td>
<td>0.07</td>
<td>0.10</td>
<td>0.10</td>
</tr>
<tr>
<td>Health effect</td>
<td>0.07</td>
<td>0.13</td>
<td>0.13</td>
<td>0.13</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
</tr>
<tr>
<td>CIV multiplier</td>
<td>0.95</td>
<td>1.67</td>
<td>1.67</td>
<td>1.31</td>
<td>1.31</td>
<td>1.31</td>
<td>1.31</td>
</tr>
<tr>
<td>Sample median income ($)</td>
<td>24,400</td>
<td>21,300</td>
<td>22,570</td>
<td>24,046</td>
<td>24,046</td>
<td>24,046</td>
<td>24,046</td>
</tr>
<tr>
<td>CIV</td>
<td>23,338</td>
<td>35,500</td>
<td>29,605</td>
<td>60,616</td>
<td>60,616</td>
<td>60,616</td>
<td>60,616</td>
</tr>
</tbody>
</table>
Table 4.6. Wellbeing model estimation by race/ethnicity (PSID, 2009 and 2011).

<table>
<thead>
<tr>
<th>Variables</th>
<th>Whites</th>
<th>African Americans</th>
<th>Hispanics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender (Female)</td>
<td>0.01</td>
<td>0.07***</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.03)</td>
<td>(0.04)</td>
</tr>
<tr>
<td>Age</td>
<td>-0.04***</td>
<td>-0.0006</td>
<td>-0.02*</td>
</tr>
<tr>
<td></td>
<td>(0.004)</td>
<td>(0.006)</td>
<td>(0.009)</td>
</tr>
<tr>
<td>Age Square</td>
<td>0.0004***</td>
<td>0.00008</td>
<td>0.0002*</td>
</tr>
<tr>
<td></td>
<td>(0.00004)</td>
<td>(0.00006)</td>
<td>(0.0001)</td>
</tr>
<tr>
<td>Marital status (Married)</td>
<td>0.39***</td>
<td>0.31***</td>
<td>0.39***</td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.03)</td>
<td>(0.05)</td>
</tr>
<tr>
<td>Employment status (employed)</td>
<td>0.09***</td>
<td>0.14***</td>
<td>0.12**</td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.03)</td>
<td>(0.04)</td>
</tr>
<tr>
<td>Diabetes (Yes)</td>
<td>-0.09**</td>
<td>-0.14**</td>
<td>-0.10</td>
</tr>
<tr>
<td></td>
<td>(0.03)</td>
<td>(0.04)</td>
<td>(0.07)</td>
</tr>
<tr>
<td>Log per capita family income</td>
<td>0.07***</td>
<td>0.003</td>
<td>0.06**</td>
</tr>
<tr>
<td></td>
<td>(0.01)</td>
<td>(0.009)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>Education level (High school graduate)</td>
<td>-0.04</td>
<td>0.05</td>
<td>-0.002</td>
</tr>
<tr>
<td></td>
<td>(0.05)</td>
<td>(0.06)</td>
<td>(0.06)</td>
</tr>
<tr>
<td>Education level (College graduate or more)</td>
<td>0.006</td>
<td>0.006</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>(0.05)</td>
<td>(0.06)</td>
<td>(0.06)</td>
</tr>
<tr>
<td>Number of Children in the household</td>
<td>0.06**</td>
<td>0.02</td>
<td>0.05**</td>
</tr>
<tr>
<td></td>
<td>(0.008)</td>
<td>(0.01)</td>
<td>(0.018)</td>
</tr>
<tr>
<td>Constant</td>
<td>2.57***</td>
<td>2.25***</td>
<td>2.28***</td>
</tr>
<tr>
<td></td>
<td>(0.12)</td>
<td>(0.15)</td>
<td>(0.26)</td>
</tr>
<tr>
<td>N</td>
<td>14,166</td>
<td>7,027</td>
<td>2,538</td>
</tr>
</tbody>
</table>

Standard errors in parentheses; * p < 0.05, ** p < 0.01, *** p < 0.001
Table 4.7. Indirect, direct, and total income effects and compensating income variation for by race/ethnicity (PSID, 2009 and 2011).

<table>
<thead>
<tr>
<th>Dropped variable</th>
<th>Whites</th>
<th>African Americans</th>
<th>Hispanics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct income</td>
<td>Indirect income</td>
<td>Direct income</td>
</tr>
<tr>
<td></td>
<td>effects</td>
<td>effects</td>
<td>effects</td>
</tr>
<tr>
<td>Employment status</td>
<td>0.08</td>
<td>0.009</td>
<td>0.01</td>
</tr>
<tr>
<td>education level</td>
<td>0.08</td>
<td>0.004</td>
<td>0.002</td>
</tr>
<tr>
<td>Marital status</td>
<td>0.09</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>Health (diabetes)</td>
<td>0.07</td>
<td>0.001</td>
<td>0.003</td>
</tr>
<tr>
<td>Indirect income effect</td>
<td>0.04</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>Direct income effect</td>
<td>0.07</td>
<td>0.003</td>
<td></td>
</tr>
<tr>
<td>Total income effect</td>
<td>0.11</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>Health effect</td>
<td>0.09</td>
<td>0.12</td>
<td></td>
</tr>
<tr>
<td>CIV multiplier</td>
<td>0.88</td>
<td>4.61</td>
<td></td>
</tr>
<tr>
<td>Sample median income ($)</td>
<td>28,600</td>
<td>14,833</td>
<td>15,000</td>
</tr>
<tr>
<td>CIV ($)</td>
<td>25,139</td>
<td>68,335</td>
<td>13,082</td>
</tr>
<tr>
<td>N</td>
<td>14,166</td>
<td>7,027</td>
<td>2,538</td>
</tr>
</tbody>
</table>
Figure 2.1. Mechanism and synergetic effect of T2D risk factors. SNS= Sympathetic Nervous System and T2D= Type 2 diabetes
Figure 2.2: Plots of nonlinear contribution of the explanatory variables to the MARS basis transformation functions
Figure 2.3. Area under operating characteristics curve for additive MARS model and Logistic regression model
Figure 2.4: Sample surface plots of interactions between basis of the explanatory variables as contribution to the MARS basis transformation functions
Figure 2.5. Area under operating characteristics curve for two-way interactions MARS model.
Figure 3.1. A Conceptual framework showing pathways in which socio-ecological factors influences diabetes related health behavior and outcome.
Figure 3.2. Spatial cluster of diabetes-related mortality rate (DRMR) in contiguous US counties
Figure (3.3a). Bivariate LISA cluster map for DRMR and neighboring county (A) Socioeconomic index (SEI), (B) Social capital index (SCI), (C) Income disparity (Gini coefficient), (D) Proportion of urban resident, (E) Retail food environment index (RFEI), and (F) Proportion of households who lack access to grocery stores.
Figure (3.3b). Bivariate LISA cluster map for DRMR and neighboring county (H) medical doctor’s density, (I) hospital beds density, (J) proportion of uninsured residents, (K) spatial proximity index (SPI) for segregation, (L) proportion of Black population, and (M) proportion of Hispanic population.
Appendix A: Summary of literature rated to chapter 2

Table A.1: Summary of diabetes prediction models based on United States population (chapter 2).

<table>
<thead>
<tr>
<th>Author, date</th>
<th>Population</th>
<th>Study type and sample size</th>
<th>Diabetes diagnosis criteria</th>
<th>Final model risk factors</th>
<th>Statistical method and Limitations</th>
<th>Type of model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kanaya et al., 2005</td>
<td>Affluent, white, older adults in southern California</td>
<td>Cohort of Rancho Bernardo study (RBS), (1984–1987), N=1,549</td>
<td>Diabetes: 2h-OGTT</td>
<td>Age, sex, FPG, and triglycerides.</td>
<td>Logistic regression ROC=0.73 : Sample is not representative for the US population (by ethnicity), no interaction between variables, model lacks important factors (diet, physical activity and psychological factors).</td>
<td>Invasive</td>
</tr>
<tr>
<td>Kahn et al., 2009</td>
<td>Age 45-64</td>
<td>Cohort of Atherosclerosis Risk in Communities (ARIC) study, from 1987-1989 to 1996-1998, N=9,587</td>
<td>Diabetes: reported diagnosed diabetes or FPG or 2h-OGTT or Random Blood Glucose (RBG)</td>
<td>Waist circumference, maternal diabetes, paternal diabetes, height, ethnicity, age, weight, hypertension, pulse rate, and smoking history.</td>
<td>Accelerated failure time (Weibull) ROC=0.71 : No interaction between variables, arbitrary segmentation of continuous variables, lacks important factors (diet, physical activity, psychological factors, BMI, west circumference) and sample is not representative for US population (by age, and ethnicity).</td>
<td>Invasive + Noninvasive</td>
</tr>
<tr>
<td>Schmidt et al., 2005</td>
<td>Age 45–64</td>
<td>Cohort study of the Atherosclerosis Risk in Communities (ARIC) study, from 1987-1989 to 1996-1998, N=7,915</td>
<td>Diabetes: Clinical diagnosis or 2 h-OGTT</td>
<td>Waist circumference, height, hypertension, blood pressure, family history of diabetes, ethnicity, and age. Additional models included triglycerides, HDL cholesterol, and FPG</td>
<td>Logistic regression ROC=0.71 : No interaction between variables, arbitrary segmentation of continuous variables, model lacks important factors (diet, physical activity, psychological factors, BMI, west circumference) and sample is not representative for US population (by age, and ethnicity)</td>
<td>Invasive + Noninvasive</td>
</tr>
<tr>
<td>Stern et al., 2002</td>
<td>Mexican American and non-Hispanic whites</td>
<td>San Antonio Heart Study participant cohort study N=2903</td>
<td>Diabetes: FPG 2-h OGTT or reported history of diabetes diagnosed by physician and reported current use of insulin or oral antidiabetic</td>
<td>Age, sex, ethnicity, FPG, 2h-OGTT, systolic blood pressure, HDL cholesterol, LDL cholesterol, BMI, family history of diabetes, triglycerides</td>
<td>Logistic regression ROC=0.85 : Identical variable used as dependent and predictor (fasting glucose used for diagnosis and for predicting), model lacks important factors (diet, physical activity, psychological factors, waist circumference) and sample is not representative for US population (by age, and ethnicity)</td>
<td>Invasive</td>
</tr>
</tbody>
</table>
Table A.1: Summary of diabetes prediction model based on United States population (chapter 2). (Continued)

<table>
<thead>
<tr>
<th>Author, date</th>
<th>Population</th>
<th>Study type and sample size</th>
<th>Diabetes diagnosis criteria</th>
<th>Final model risk factors</th>
<th>Statistical method and limitations</th>
<th>Limitations</th>
<th>Type of model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wilson et al., 2007</td>
<td>99% white</td>
<td>The Framingham Offspring Prospective cohort study, from 1995 - 1998 to 1998 – 2001 N=3140</td>
<td>Diabetes: FPG or if started using oral hypoglycemic medication or insulin</td>
<td>Age, sex, FPG, BMI, HDL cholesterol, parental history of diabetes, triglyceride level, blood pressure</td>
<td>Logistic regression ROC=0.85</td>
<td>Sample is not representative for US population (by ethnicity), no interaction between variables, segmentation of continuous variables, and model lacks important factors (diet, physical activity and psychological factors)</td>
<td>Invasive</td>
</tr>
<tr>
<td>Bang et al., 2009</td>
<td>Adults 20 years and older</td>
<td>Crosssectional study (NHANES, 1999 to 2004)</td>
<td>Diabetes: FPG</td>
<td>Age, sex, family history of diabetes, history of hypertension, BMI and physical activity</td>
<td>Logistic regression ROC=0.79</td>
<td>Segmentation of continuous variables (BMI &amp; age), and model lacks important factors (diet, and psychological factors)</td>
<td>Non invasive</td>
</tr>
<tr>
<td>Heikes et al., 2008</td>
<td>Adults 20 years and older</td>
<td>CROSSECTIONAL STUDY (NHANES III) N=7,092</td>
<td>Diabetes: FPG or 2-h OGTT Prediabetes: FPG or 2-h OGTT Undiagnosed diabetes: FPG and/or 2-h OGTT and absence of having been told that he or she has diabetes</td>
<td>Age, waist circumference, history of gestational diabetes, family history of diabetes, ethnicity, high blood pressure, weight, height, parental diabetes, physical activity</td>
<td>Logistic regression + Classification and Regression Tree(CART) ROC=0.85</td>
<td>Model lacks important factors (diet, and psychological factors) and no interaction between variables</td>
<td>Non invasive</td>
</tr>
<tr>
<td>Borrell et al., 2007</td>
<td>Adults 20 years and older</td>
<td>Cross-sectional study (NHANES III) N=4,830</td>
<td>Diabetes: FPG</td>
<td>Age, sex, ethnicity, family history of diabetes, self-reported hypertension and Self-reported hypercholesterolemia. and periodontitis</td>
<td>Conditional Logistic regression ROC=0.74</td>
<td>Model lacks important factors (diet, physical activity and psychological factors) and no interaction between variables</td>
<td>Invasive</td>
</tr>
</tbody>
</table>
## Appendix B: Summary of data source and literature related to chapter 3

Table B.1: Candidate variables for spatial statistics and spatial econometrics analysis of diabetes-related mortality rate.

<table>
<thead>
<tr>
<th>Factors</th>
<th>Evidence from literature</th>
<th>Variables (measurement)</th>
<th>Data source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diabetes specific mortality</td>
<td>Dependent variable</td>
<td>Diabetes specific mortality</td>
<td>CDC WONDER (suppressed mortality), Data available from 1999 to 2010</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><a href="http://wonder.cdc.gov/controller/datarequest/D91">http://wonder.cdc.gov/controller/datarequest/D91</a></td>
</tr>
<tr>
<td>Socioeconomic)</td>
<td>Diabetes-related mortality could increase due to poverty that economic deprivation may limit people’s ability to adhere to lifestyle (diet and physical activity) and medication recommendations to delay or manage type-2 diabetes. Counties with high proportion of family earning below federal poverty line have high overall mortality rate.</td>
<td>Percentage of county population living under poverty median household income Education: percent of adult population with four year college education</td>
<td>Health Resources and Services Administration: area health resource, 1999 – 2010</td>
</tr>
<tr>
<td>Poverty/income</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Household income inequality</td>
<td>Income inequality affects diabetes-related mortality through psychological process. It is argued that income inequality affect health of individuals through perceptions social hierarchy based on relative socioeconomic position based on income. Such perceptions produce negative emotions that trigger psycho-neuro-endocrine imbalance and stress induced behaviors such as over eating and smoking (Lynch, et al. 2000).</td>
<td>Gini coefficient</td>
<td>Spatial Impact Factor Data, RTI International, Version 5, May 2012 (<a href="https://rtispatialdata.rti.org/">https://rtispatialdata.rti.org/</a>), Data available for year 2000</td>
</tr>
<tr>
<td>Availability and of healthy food</td>
<td>Availability of healthy food is associated with the cardiovascular related mortality (Alter &amp; Eny, 2005)</td>
<td>Density of full super market in a county Density of fast food restaurant per 1000 population</td>
<td>US census County Business Patterns, Data available 1999 - 2010</td>
</tr>
</tbody>
</table>
Table B.1: Definition of variables for spatial statistics and spatial econometrics analysis of diabetes-related mortality rate. (continued)

<table>
<thead>
<tr>
<th>Factors</th>
<th>Evidence from literature</th>
<th>Variables (measurement)</th>
<th>Data source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical access to healthy food</td>
<td>Percentage of county population living more than 1 mile (10 mile for rural population) from grocery store and have no vehicle</td>
<td>Percentage of county population living more than 1 mile (10 mile for rural population) from grocery store and have no vehicle</td>
<td>USAD-ERS food atlas. Data available for 2010</td>
</tr>
<tr>
<td>Access to physical activity</td>
<td>Suitability of built environment for physical activity is related to incidence and diabetes-related mortality (Wilcox et al., 2000; Deshpande et al., 2005; Gordon-Larsen et al., 2006). Attractive (inviting) open areas, walkability of the residential and workplace, neighborhood crime rate, proximity and affordability of commercial physical activity facilities are some of the natural and built environment that encourages or discourages daily physical activity in relation to diabetes management.</td>
<td>Density of commercial physical activity related facilities per 1000 population</td>
<td>US census County Business Patterns, data available 1999 – 2010 Spatial Impact Factor Data, RTI International, Version 5, May 2012 (<a href="https://rtispatialdata.rti.org/">https://rtispatialdata.rti.org/</a>), Data available for year 2001</td>
</tr>
<tr>
<td>Demographic</td>
<td>Disparity among ethnic mortality rate remains an issue in the US (Hoyert, 2012). Diabetes prevalence is higher among African American and Hispanics, and the mortality rate from diabetes is also higher among these groups (Carter JS, Pugh JA, Monterrosa, 1996). This need to be controlled for in the model.</td>
<td>Percentage of African Americans the county Percentage of African Americans the county</td>
<td>US census Spatial Impact Factor Data, RTI International, Version 5, May 2012 (<a href="https://rtispatialdata.rti.org/">https://rtispatialdata.rti.org/</a>)</td>
</tr>
<tr>
<td>Racial segregation</td>
<td>According to Williams &amp; Collins (2001), racial residential segregation is a fundamental source of racial disparities in health outcome. Enforced physical separation of races through norms limiting interactions between races and institutional racism may contribute to the socioeconomic disparity, which is a root cause for health outcome disparity (Williams &amp; Collins, 2001).</td>
<td>spatial proximity index</td>
<td>Spatial Impact Factor Data, RTI International, Version 5, May 2012 (<a href="https://rtispatialdata.rti.org/">https://rtispatialdata.rti.org/</a>), Data available for year 2000</td>
</tr>
<tr>
<td>Access to health services</td>
<td>Positive association between healthcare resources and mortality rate has been established (Shi et al., 2003). Rural counties with low density of physicians and hospital beds have higher overall mortality rate. In case of diabetes-related mortality the statistics should be higher due to frequent follow up needed for diabetes disease management (Brown et al., 2004).</td>
<td>Proportion of population under 65 with insurance Density of primary care physicians in the county Density of hospital beds in the county</td>
<td>Area resource file (ARF), Department of Health and Human Services (DHHS); Density of hospital beds in the county, data available 2000, 2005-2010</td>
</tr>
</tbody>
</table>
Table B.2: Summary of literature related to chapter 3.

<table>
<thead>
<tr>
<th>Study</th>
<th>Unit of Analysis</th>
<th>Statistical Method</th>
<th>health outcome studied</th>
<th>Explanatory variable included</th>
<th>Result</th>
<th>Limitation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sparks &amp; Sparks (2010)</td>
<td>County (all continental US counties)</td>
<td>spatial lag and spatial error models</td>
<td>All-cause mortality</td>
<td>percentage of the county population that is rural; percentage of the county population that is black; percentage of the county population that is Hispanic; percentage of the county population that lives below the federally designated poverty threshold; percentage of households in the county with a female head; the county unemployment rate; the median household income in the county; the county’s median house value; and the population density per square mile in the county.</td>
<td>The results show that after controlling for spatial structure in the data, several key socio-ecological variables become insignificant in the analysis. In addition, the authors conclude that the spatial pattern is largely the result of the existing autocorrelation among the omitted variables in the empirical model (spatial error explained the model).</td>
<td>-Mortality not specific to any disease</td>
</tr>
<tr>
<td>Yi-Ju Chen et al., (2012)</td>
<td>County (all continental US counties)</td>
<td>geographically weighted quantile regressions</td>
<td>All-cause mortality</td>
<td>racial/ethnic composition, income inequality, poverty rate, percentage of persons receiving public assistance, unemployment rate, percentage of families that have incomes of more than $75,000, percentage of workers in managerial professional positions, percentage of adults with at least a bachelor’s degree, percentage of female-headed households with children, and crimes rate.</td>
<td>The result from GWQR analysis show that the associations between county overall mortality rate and socioeconomics determinants mortality vary spatially.</td>
<td>-Mortality not specific to any disease -the possibility of spatial dependence among independent variables not modeled</td>
</tr>
<tr>
<td>Massing et al. (2004)</td>
<td>County (all continental US counties)</td>
<td>Poisson regression analyses</td>
<td>CVD specific mortality rate</td>
<td>County median income and county income inequality</td>
<td>County income inequality was negatively associated with CVD, coronary heart disease, and stroke mortality</td>
<td>-the possibility of spatial dependence among dependent and independent -might be biased by missing variable problem (few variables included)</td>
</tr>
<tr>
<td>Study</td>
<td>Unit of Analysis</td>
<td>Statistical Method</td>
<td>health outcome studied</td>
<td>Explanatory variable included</td>
<td>Result</td>
<td>Limitation</td>
</tr>
<tr>
<td>-------</td>
<td>------------------</td>
<td>--------------------</td>
<td>------------------------</td>
<td>-------------------------------</td>
<td>--------</td>
<td>------------</td>
</tr>
<tr>
<td>Saydah S &amp; Lochner (2010)</td>
<td>County (all continental US counties)</td>
<td>geographically weighted regressions</td>
<td>Prevalence of diabetes</td>
<td>Percent of population under poverty (federal threshold) in counties, percent of Latino population counties, and percent of African American population in counties.</td>
<td>The statistical relationship between diabetes and percent of population under poverty in a county varies as a function of location. In particular, prevalence of diabetes in a county is not always positively associated with percent of population under poverty.</td>
<td>-the possibility of spatial dependence among independent variables not modeled</td>
</tr>
<tr>
<td>Shrestha et al. (2012)</td>
<td>County (all continental US counties)</td>
<td>Global and local (LISA) Moran’s I and multinomial regression</td>
<td>Prevalence of diabetes</td>
<td>Percent of population under poverty (federal threshold) in counties, percent of Latino population in counties, percent of African-American population in counties, age-adjusted estimates of the percentage of obese adults in counties.</td>
<td>Clusters of counties with high diabetes prevalence also have higher percent of population under poverty, high proportion of African American, high rate of obesity and physical inactivity.</td>
<td>-the possibility of spatial dependence among independent variables not modeled</td>
</tr>
<tr>
<td>Gebreab et al., (2012)</td>
<td>County (all continental US counties)</td>
<td>geographically weighted regression</td>
<td>Black and white CHD specific mortality</td>
<td>percentage of persons with income below poverty line separately for African Americans and whites at the county-level and county-level residential segregation</td>
<td>The GWR results showed significant spatial heterogeneity in black-white differences in CHD mortality. However, after controlling for county and race-specific poverty and segregation, significant race differences in CHD mortality were no longer present.</td>
<td>-the possibility of spatial dependence among independent variables not modeled</td>
</tr>
</tbody>
</table>
## Appendix C: summary of data and literature related to chapter 4

Table C.1: Description of PSID questionnaire as related to candidate variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>PSID questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>General health</td>
<td>Now I have a few questions about your health, including any serious limitations you might have. Would you (HEAD) say your health in general is … 1. Excellent 2. Very Good 3. Good 4. Fair 5. Poor</td>
</tr>
<tr>
<td>Diabetes</td>
<td>“Has a doctor ever told you that you have diabetes or high blood sugar?” 1. Yes 2. No If yes, “In what month and year was that first diagnosed?”</td>
</tr>
<tr>
<td>Age</td>
<td>Age in years</td>
</tr>
<tr>
<td>Race</td>
<td>What is your race? Are you white, black, American Indian, Alaska Native, Asian, Native Hawaiian or other Pacific Islander? 1. white 2. African Americans 3. American Indian or Alaska Native 4. Asian, Native Hawaiian or Pacific Islander 5. Other *There is also separate question Hispanic origin</td>
</tr>
<tr>
<td>Sex</td>
<td>Sex 1. Male 2. Female</td>
</tr>
<tr>
<td>Education level</td>
<td>“How many grades of school did you finish?”</td>
</tr>
<tr>
<td>Employment status</td>
<td>Employment status every year 1. employed 2. Unemployed 3. Out of labor force 4. Inactive service</td>
</tr>
</tbody>
</table>
Table C.2: Descriptive statistics for categorical variables, PSID, 2009 and 2011.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Overall Frequency (%)</th>
<th>Between variation Frequency (%)</th>
<th>Within variation %</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sex</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>12232 (45.48)</td>
<td>6762 (45.91)</td>
<td>100</td>
</tr>
<tr>
<td>Female</td>
<td>14663 (54.52)</td>
<td>7968 (54.09)</td>
<td>100</td>
</tr>
<tr>
<td>Total</td>
<td>26895 (100.00)</td>
<td>14730 (100)</td>
<td>100</td>
</tr>
<tr>
<td><strong>Marital Status</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Married</td>
<td>16631 (61.85)</td>
<td>8898 (60.42)</td>
<td>100</td>
</tr>
<tr>
<td>Never married</td>
<td>5519 (20.52)</td>
<td>3233 (21.95)</td>
<td>100</td>
</tr>
<tr>
<td>Widowed</td>
<td>777 (2.89)</td>
<td>435 (2.95)</td>
<td>100</td>
</tr>
<tr>
<td>Divorced</td>
<td>3100 (11.53)</td>
<td>1694 (11.50)</td>
<td>100</td>
</tr>
<tr>
<td>Separated</td>
<td>864 (3.21)</td>
<td>466 (3.16)</td>
<td>100</td>
</tr>
<tr>
<td>Total</td>
<td>26891 (100)</td>
<td>14726 (100)</td>
<td>100</td>
</tr>
<tr>
<td><strong>Ethnicity/Race</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>14899 (56.37)</td>
<td>8147 (56.19)</td>
<td>100</td>
</tr>
<tr>
<td>African Americans</td>
<td>7974 (30.17)</td>
<td>4442 (30.64)</td>
<td>100</td>
</tr>
<tr>
<td>Hispanic</td>
<td>2837 (10.73)</td>
<td>1615 (11.14)</td>
<td>100</td>
</tr>
<tr>
<td>Others</td>
<td>720 (2.72)</td>
<td>403 (2.78)</td>
<td>100</td>
</tr>
<tr>
<td>Total</td>
<td>26430 (100)</td>
<td>14607 (100)</td>
<td>100</td>
</tr>
<tr>
<td><strong>Education level</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not completed high school</td>
<td>1888 (7.52)</td>
<td>1197 (8.65)</td>
<td>94.40</td>
</tr>
<tr>
<td>Completed high school</td>
<td>8776 (34.94)</td>
<td>4926 (35.61)</td>
<td>98.63</td>
</tr>
<tr>
<td>Higher education</td>
<td>14452 (57.54)</td>
<td>7914 (57.21)</td>
<td>99.13</td>
</tr>
<tr>
<td><strong>Employment status</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unemployed</td>
<td>8113 (30.76)</td>
<td>5457 (37.69)</td>
<td>82.11</td>
</tr>
<tr>
<td>Employed</td>
<td>18262 (69.24)</td>
<td>10973 (75.79)</td>
<td>91.11</td>
</tr>
<tr>
<td>Total</td>
<td>26375 (100)</td>
<td>16430 (113.48)</td>
<td>88.12</td>
</tr>
<tr>
<td><strong>Diabetes status</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not diagnosed</td>
<td>24115 (90.88)</td>
<td>13508 (92.66)</td>
<td>98.33</td>
</tr>
<tr>
<td>Diagnosed</td>
<td>2420 (9.12)</td>
<td>1521 (10.43)</td>
<td>85.17</td>
</tr>
<tr>
<td><strong>Life satisfaction</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not at all satisfied</td>
<td>308 (1.16)</td>
<td>281 (1.93)</td>
<td>60.85</td>
</tr>
<tr>
<td>Somewhat satisfied</td>
<td>1014 (3.82)</td>
<td>914 (6.26)</td>
<td>62.96</td>
</tr>
<tr>
<td>Not very satisfied</td>
<td>7671 (28.92)</td>
<td>5926 (40.60)</td>
<td>72.27</td>
</tr>
<tr>
<td>Very satisfied</td>
<td>11792 (44.46)</td>
<td>8569 (58.71)</td>
<td>74.90</td>
</tr>
<tr>
<td>Completely satisfied</td>
<td>5736 (21.63)</td>
<td>4594 (31.47)</td>
<td>68.54</td>
</tr>
<tr>
<td>Total</td>
<td>26521 (100)</td>
<td>20284 (138.97)</td>
<td>71.96</td>
</tr>
</tbody>
</table>
Table C.3: Descriptive statistics for continuous variables, PSID, 2009 and 2011.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Variation</th>
<th>Mean</th>
<th>St. deviation</th>
<th>Min</th>
<th>Max</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>overall</td>
<td>43.83</td>
<td>14.41</td>
<td>20</td>
<td>80</td>
<td>N=26895</td>
</tr>
<tr>
<td></td>
<td>between</td>
<td>14.64</td>
<td>20</td>
<td>80</td>
<td></td>
<td>n=14730</td>
</tr>
<tr>
<td></td>
<td>within</td>
<td>0.97</td>
<td>37.83</td>
<td>49.83</td>
<td></td>
<td>T-bar=1.83</td>
</tr>
<tr>
<td>Total family income</td>
<td>overall</td>
<td>76743.72</td>
<td>105049.50</td>
<td>-70000</td>
<td>6317099</td>
<td>N=26895</td>
</tr>
<tr>
<td></td>
<td>between</td>
<td>95535.32</td>
<td>37000</td>
<td>0</td>
<td>4035852</td>
<td>n=14730</td>
</tr>
<tr>
<td></td>
<td>within</td>
<td>39300.93</td>
<td>-2204504</td>
<td>2357991</td>
<td></td>
<td>T-bar=1.83</td>
</tr>
<tr>
<td>Number of Children</td>
<td>overall</td>
<td>0.90</td>
<td>1.20</td>
<td>0</td>
<td>11</td>
<td>N=26895</td>
</tr>
<tr>
<td></td>
<td>between</td>
<td>1.16</td>
<td>0</td>
<td>10</td>
<td></td>
<td>n=14730</td>
</tr>
<tr>
<td></td>
<td>within</td>
<td>0.29</td>
<td>-2.10</td>
<td>3.90</td>
<td></td>
<td>T-bar=1.83</td>
</tr>
<tr>
<td>Family size</td>
<td>overall</td>
<td>2.85</td>
<td>1.46</td>
<td>1</td>
<td>14</td>
<td>N=26895</td>
</tr>
<tr>
<td></td>
<td>between</td>
<td>1.41</td>
<td>1</td>
<td>13</td>
<td></td>
<td>n=14730</td>
</tr>
<tr>
<td></td>
<td>within</td>
<td>0.41</td>
<td>-1.15</td>
<td>6.85</td>
<td></td>
<td>T-bar=1.83</td>
</tr>
</tbody>
</table>
Table C.4: Summary of related studies for chapter 4 (continued).

<table>
<thead>
<tr>
<th>Study</th>
<th>Data used</th>
<th>Well-being measure used</th>
<th>Independent variables included</th>
<th>Assumption</th>
<th>Diseases</th>
<th>Result</th>
<th>Limitation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ferrer-i-Carbonell &amp; van Praag, 2002</td>
<td>German socioeconomic panel data (GSOEP)</td>
<td>Six life satisfaction domains were cardinalized (converted from ordinal measure to scale to make them continuous variable) and their effect on the general life satisfaction was estimated. Then the effect of income on each domain of life satisfaction was estimated. The income change needed to keep the welfare constant was calculated by multiplying change in health with multiple effect of income (health domain effect on general life satisfaction divided by sum of effect of income on each life domain). But did not have health event data to calculate the actual income change within their data, rather used estimates from other studies to calculate percentage change in income.</td>
<td>job satisfaction, financial satisfaction, house satisfaction, health satisfaction, leisure satisfaction, environmental satisfaction, income age and sex</td>
<td>Individual well-being or life satisfaction depends on 6 domain of life (job satisfaction, financial satisfaction, house satisfaction, health satisfaction, leisure satisfaction, environmental satisfaction).</td>
<td>Multiple chronic diseases including diabetes</td>
<td>Diabetes would decrease individuals' well-being as much as reducing income by 59%. Hearing impediments are on average equivalent to an income reduction of about 20%, and heart or blood difficulties are equivalent to a 47% income reduction.</td>
<td>CIV not estimated from the data. Only prevalence of not the frequency or severity of these conditions was used in the estimation.</td>
</tr>
<tr>
<td>Groot &amp; Van den Brink, 2004a</td>
<td>Cross sectional study of Dutch survey called Center for Economic Research on Retirement and Ageing, wave I (CERRA-I). Sample=2786</td>
<td>The life-satisfaction question is phrased as follows: “Here is a picture of a ladder, representing the ladder of life. The bottom of this ladder, step 0, represents the worst possible life, while the top of this ladder, step 10, represents the best possible life. Where on this ladder do you feel you personally stand at present?” Quality of life weight (QoLW) calculated based on the coefficient from estimation of income and health variable on the life satisfaction equation. Compensating income variation was also calculated.</td>
<td>Age, income, bronchitis, asthma, high blood pressure, stroke, stomach ulcer, gallstones, intestine problems, kidney stones, nephritic disease, prostrate, diabetes, back pain, hernia, ischia’s, artroses, rheumatism, Parkinson disease, multiple Scleroses, epilepsy, overstrain, depression, severe nervousness, cancer, chronic skin disease, eczema, prolapse, varicose veins, accident injury</td>
<td>Life satisfaction is determined by income and health status and other individual characteristics (including co-morbidity)</td>
<td>Headache and migraine</td>
<td>Individuals need 1920–2680 guilders per month to compensate for the loss of life satisfaction due to severe headache or migraine</td>
<td>Only prevalence of severe headache or migraine and not the frequency or severity of these conditions was used in the estimation. Indirect effect of income not accounted for (biased estimation)</td>
</tr>
<tr>
<td>Study</td>
<td>Data used</td>
<td>Well-being measure used</td>
<td>Independent variables included</td>
<td>Assumption</td>
<td>Diseases</td>
<td>Result</td>
<td>Limitation</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>---------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------</td>
<td>---------------------------</td>
<td>------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Groot &amp; Van den Brink, 2004b</td>
<td>Cross-sectional data from 1995 wave of the Supplementary Provision Survey (SPS) of the Dutch Social and Cultural Planning Bureau</td>
<td>The Leyden function is based on the income valuation of individuals, where individual asked to rate monthly household after tax income from bad, Bad, Insufficient, Sufficient, Good, Very good</td>
<td>Age, age-square, sex, household size, household income, education level, and dummy for disease</td>
<td>Cardinal utility was assumed and the Leyden welfare function was used to develop the scale between 0 and .99.</td>
<td>Cardiovascular disease</td>
<td>Combined direct and indirect welfare effects of CVD at age 25 requires income to compensate with a 100% income increase to maintain welfare level</td>
<td>CIV not estimated from the data. Indirect effect of income not accounted for (biased estimation). Only prevalence of not the frequency or severity of these conditions was used in the estimation.</td>
</tr>
<tr>
<td>Groot &amp; Van den Brink, 2006</td>
<td>Panel data from the British Household Panel Survey</td>
<td>Life satisfaction question: “How dissatisfied or satisfied are you with your life overall?” scale ranges from 1 = not satisfied at all to 7 = completely satisfied. Compensating income variation calculated directly from the estimated confidents (with ordered probit)</td>
<td>Age, per capita household income, cardiovascular diseases</td>
<td>It is assumed that life satisfaction is determined by income and health status and other individual characteristics (including co-morbidity)</td>
<td>Cardiovascular diseases</td>
<td>Average willingness to accept (Compensating income variation) for heart diseases is nearly three times higher for men than for women: 49,564 pound for men and 17,503 pound for women. For men the CIV ranges from 93,532 pound for a 25 year old to 1,808 pound for a 75 year old.</td>
<td>Indirect effect of income not accounted for (biased estimation). Only prevalence of not the frequency or severity of these conditions was used in the estimation.</td>
</tr>
<tr>
<td>Study</td>
<td>Data used</td>
<td>Well-being measure used</td>
<td>Independent variables included</td>
<td>Assumption</td>
<td>Diseases</td>
<td>Result</td>
<td>Limitation</td>
</tr>
<tr>
<td>-----------------------</td>
<td>----------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------</td>
<td>--------------------------------</td>
<td>------------</td>
<td>---------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Powdthavee &amp; Van den Berg, 2011</td>
<td>The panel data set comes from the British Household Panel Survey (BHPS). Sample 22,169</td>
<td>The well-being was measured based on three valuations: Overall life satisfaction (“All things considered, how satisfied or dissatisfied are you with your life overall using a 1-7 scale? 1=very dissatisfied, . . . , 7 = very satisfied”) The usual states of mental well-being (derived from General Health Questionnaire (GHQ-12) score, a measure of mental stress and strain) Health satisfaction (“How satisfied or dissatisfied are you with your health overall using a 1-7 scale? 1 = very dissatisfied, . . . , 7 = very satisfied”) Self-assessed health (“Please think back over the last 12 months about how your health has been. Compared to people of your own age, would you say that your health has on the whole been excellent, good, fair, poor, and very poor?” The responses are coded so that 1 = very poor health, …. , 5 = excellent health) age, age-squared, gender, employment, education, marital status, number of dependent children</td>
<td>Utility is assumed to be unobserved. The authors used continuous scale measure of wellbeing in the regression, assuming cardinal utility. It is assumed that life satisfaction is determined by income and stock of health status and other individual characteristics</td>
<td>Problems connected with arms, legs, hand, feet, and back, Difficulty in seeing (other than needing glasses to read), difficulty in hearing, skin conditions/allergies, chest/breathing problems, asthma, bronchitis, heart/blood pressure or blood circulation problems, stomach/liver/kidneys or digestive problems, diabetes, depression and anxiety, alcohol or drug related problems, epilepsy, migraine or frequent headaches, other health problems not, Health limits daily activities, health limits amount or type of work</td>
<td>Person with diabetes need £6000, £2000, £132x1011, and £1482000 based on the life satisfaction, mental well-being, health satisfaction and self-assessed health. Note that the monitory value based on the health satisfaction question is extremely large. Which may show unreliability of this measure? Mental-well-being based measure provide least monetary value.</td>
<td>Indirect effect of income not accounted for (biased estimation). Only prevalence of not the frequency or severity of these conditions was used in the estimation.</td>
<td></td>
</tr>
</tbody>
</table>