2008

A Comparison of Ordinary Least Squares and Instrumental Variables Regression for High Intensity Disease Management Evaluation

Gerald A. Craver

School of Education
Virginia Commonwealth University

Dissertation Approval Certificate

This is to certify that the dissertation prepared by Mr. Gerald A. Craver entitled

A Comparison of Ordinary Least Squares and Instrumental Variables Regression for High Intensity Disease Management Evaluation

has been approved by his committee as satisfying completion of the dissertation requirement for the degree of Doctor of Philosophy.

[redacted] Director of Dissertation	✓ Pass	Fail
[redacted] Committee Member	✓ Pass	Fail
[redacted] Committee Member	✓ Pass	Fail
[redacted] Committee Member	✓ Pass	Fail
[redacted] Committee Member	✓ Pass	Fail
[redacted] Director, Ph.D. in Education	✓ Pass	Fail
[redacted] Dean, School of Education	✓ Pass	Fail
[redacted] Dean, Graduate Studies	✓ Pass	Fail

Friday, November 7, 2008
Date

A COMPARISON OF ORDINARY LEAST SQUARES AND INSTRUMENTAL VARIABLES REGRESSION FOR HIGH INTENSITY DISEASE MANAGEMENT EVALUATION

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Education at Virginia Commonwealth University

by
GERALD A. CRAVER

B.A. History, Virginia Commonwealth University, 1993
M.U.R.P., Virginia Commonwealth University, 1999

Director: James H. McMillan, Ph.D.
Professor, Foundations in Education

Virginia Commonwealth University
Richmond, Virginia

Acknowledgements

Earning a doctor of philosophy degree has been a very challenging, but personally rewarding experience for me. It is a goal that I have had for many years. Because my father was a professor, I grew up around a university. In fact, some of my earliest memories are of the university where my father worked and his colleagues and their families. Having grown up in such an environment, I did not feel finished with my schooling until earning a doctorate. As a result, I would like to begin by thanking my parents, Sam and Jeanie Craver, for instilling in me the importance of education and for their support during this endeavor. I would especially like to thank my father for the countless hours he has invested in my education over the years. I also would like to thank my wife, April, and children, Will and Avery, who supported me through this process, which I could not have completed without them. They sacrificed many days and nights over the last few years while I was locked away in a study toiling over school work. I have a lot of catching up to do on missed family time.

I would also like to thank all of my committee members for their help and support: Dr. Jim McMillan for his overall advice and guidance during both my doctoral program and dissertation; Dr. Darcy Mays for the conversations we had about regression and for reading drafts of my dissertation; Dr. Dan Longo for his insights about health services research, disease management, and diabetes; and Dr. Lisa Abrams and Dr. Kurt Stemhagen for their comments and suggestions. The advice and support I received from my committee members was invaluable during this process.

Finally, I would like to thank the Virginia Department of Medical Assistance Services (DMAS) for allowing me to use data on Medicaid recipients who were enrolled in the Virginia *Healthy Returns*^{SM} Disease Management Program. Without the agency's assistance, my dissertation would not have been possible. Nevertheless, it should be noted that the findings and conclusions from this study represent only the researcher's views and do not represent the views of DMAS. Allowing the researcher access to DM data in no way constitutes the agency's endorsement of the content, analysis, or findings of the present study.

Table of Contents

List of Tables

List of Figures

Abstract

A COMPARISON OF ORDINARY LEAST SQUARES AND INSTRUMENTAL VARIABLES REGRESSION FOR HIGH INTENSITY DISEASE MANAGEMENT EVALUATION

Gerald A. Craver

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Education at Virginia Commonwealth University

Virginia Commonwealth University, 2008

Dissertation Director: Jim McMillan, Professor
Foundations in Education

Purpose: Disease management (DM) programs are typically evaluated using study designs that are susceptible to selection bias and other internal validity threats because participants are often allowed to self-select into the programs. As a result, DM evaluation results are usually biased because researchers are unable to control for preexisting differences between the DM participants and non-participants. Linden and Adams (2006) offer an instrumental variables (IV) regression procedure as a means of deriving unbiased estimates of DM program effectiveness. However, IV regression relies upon the existence of one or more variables (or instruments) that produce considerable variation in the program participation variable, but have no direct effect on the outcome variable. Linden and Adams argue that participant three-digit zip codes meet these criteria and can be used as instruments in IV regression.

Methods: To test the feasibility of their IV regression procedure, a series of ordinary least squares (OLS) and instrumental variables (IV) regression models were used to evaluate the effects of a high intensity Medicaid diabetes DM program on annual diabetes-related costs, emergency department visits, and hospital days. Program participation was the endogenous variable and age, gender, and propensity scores were the exogenous variable. Standard statistical tests were performed to assess the quality and validity of the IV regression models and zip code instruments.

Results: The study found that using propensity scores as covariates in the regression models appeared to offer a viable means of controlling for potential overt biases. However, the statistical tests performed to assess the quality and validity of the IV regression procedure using recipient three-digit zip codes as instruments indicated that it may not be appropriate due to various issues such as multicollinearity, lack of significant differences between the IV and OLS regression models, and weak instrument bias.

Conclusions: While the results of the present study do not support the use of participant three-digit zip codes as instruments in IV regression, the quality of the results obtained using this procedure may depend on the specific sample that is used in the analysis. Researchers may thus still wish to consider this procedure when evaluating DM programs because different samples may yield different results.

Chapter 1

Introduction

Two methods that can be used to estimate causal relationships using observational (i.e., quasi-experimental) data are the focus of this study. Causal relationships are cause-effect relationships that occur when one variable produces an effect in another variable. Effects are changes (or lack of changes) in outcome variables that can be attributed to program participation (Davidson, 2005). Social researchers examine causal relationships when they evaluate programs (or interventions) to determine if they caused certain outcomes (Shadish, Cook, & Campbell, 2002; Trochim, 2005).

The desire to estimate causal relationships between independent and dependent variables is an important objective in most quantitative social research (Winship & Morgan, 1999). Under the quantitative paradigm, experimental studies are viewed as the "gold standard" because they have the strongest internal validity due to random assignment of subjects to experimental (i.e., treatment, program, or intervention) and control groups. If executed correctly, randomization ensures that both groups are probabilistically equivalent (Trochim, 2005). During such studies, the groups are treated the same except that the treatment is administered to the experimental group. Ordinary least squares (OLS) regression is a statistical procedure that can be used to estimate the average effects of program participation for the treatment and control groups (Mohr, 1995). Differences that exist between the groups are viewed as unbiased estimates of the program's effects (Achen, 1986).

True experimental designs are relatively rare in quantitative social research (Achen, 1986). This is due to the fact that social experiments are often very expensive to conduct and subjects may be unwilling to cooperate with study requirements. Moreover, researchers are often unable to control which individuals receive the treatment. For these reasons, researchers usually rely on observational data generated through non-experimental processes such as censuses, surveys, or administrative activities when assessing causal effects of social interventions (Winship & Morgan, 1999). Relying on observational data is problematic, however, because it involves the creation of nonequivalent treatment and control groups, which can be a threat to the study's internal validity due to selection bias (Shadish et al., 2002).

Selection bias exists whenever subjects cannot be assigned randomly to treatment and control groups (McMillan & Schumacher, 2006). If present, it can interfere with the ability of researchers to make appropriate inferences about program causal effects. In the health care field, disease management (DM) programs are often evaluated using data contaminated by selection bias due to participant self-selection (Linden & Adams, 2006). The ability of DM evaluators to estimate program effects is therefore limited. Linden and Adams (2006) offer an econometric technique known as instrumental variables (IV) regression as a solution to the selection bias problem in DM evaluations. This study explored the utility of Linden and Adams' (2006) IV regression procedure by using it to analyze observational data on a group of Virginia Medicaid diabetes DM participants. The results of the IV regression models were compared against the results obtained using OLS regression to determine if Linden and Adams' (2006) procedure can generalize to this population.

Statement of the Problem

Since the early 1990s, many health plans and state Medicaid agencies have implemented disease management (DM) as a model of care for patients with chronic disease in an effort to reduce health care costs while improving the quality of care that these individuals receive. DM is a coordinated system of health care communications and interventions for populations that suffer from chronic diseases that require substantial patient self-care efforts (Buntin, 2006; Congressional Budget Office, 2004; Gillespie & Rossiter, 2003). Chronic diseases are prolonged illnesses that do not resolve spontaneously and are generally not curable by medication or vaccinations (Florida Department of Health, 2007). Examples of chronic diseases include kidney disease, diabetes, heart disease, HIV/AIDS, depression, traumatic brain injury, and multiple sclerosis (Johnson, 2003).

DM has evolved over the years from a model that focused on specific patients identified as having chronic illnesses through retrospective reviews of medical claim records to a population-based model focusing on all patients identified as having chronic illnesses based on predictive modeling algorithms. These modeling algorithms typically classify all patients with chronic illnesses into various high and low risk categories based on their potential for future high costs and poor health outcomes (Sprague, 2003).

Despite their popularity, DM programs have been one of the least rigorously evaluated developments in the health services area (Matheson, Wilkins, & Psacharopoulos, 2006; Mattke, Seid, & Ma, 2007). In fact, Buntin, (2006) refers to many DM evaluations as nothing more than marketing disguised as research. One difficulty surrounding rigorous DM evaluation is that patients often self-select into the programs. Thus, comparisons between participants and non-participants on costs and other outcome variables will probably be skewed because participants may be healthier or more active in managing their own care than non-participants (Congressional Budget Office, 2004).

Another obstacle to rigorous evaluation has been DM's shift to population-based care, where all eligible individuals are automatically enrolled unless they specifically request to be excluded. The emphasis on population-based care means that fewer control subjects are available for comparison purposes in DM evaluations (Linden & Adams, 2006). In fact, this obstacle was encountered in the present study because all of the Medicaid diabetes recipients were enrolled in the DM program. A way to overcome this obstacle, however, is to use DM patients who receive the high intensity intervention as the treatment group and DM participants who do not receive this intervention as the control group, which was the approach adopted in this study. (Additional details on the treatment and control groups are provided later in this chapter and in Chapter 3.)

The inability to evaluate DM programs rigorously calls into question the value of disease management, which is under increasing scrutiny from a variety of stakeholders including government agencies, managed care organizations, and self-insured employers because they want to know how well these programs have performed. However, the inability to conduct quality evaluations makes it extremely difficult to separate DM program participation effects on outcomes from other external factors. Some observers argue that the lack of credible evaluations is the most important issue currently facing disease management (MacDowell & Wilson, 2002).

IV regression, which is a member of a family of statistical models known as selection bias models, can provide social researchers with a tool for evaluating DM programs with more rigor. These statistically complex models were primarily developed by economists and are often difficult to implement. The models are intended to adjust for selection bias due to nonequivalent group designs by deriving unbiased estimates of program participation effects (Shadish, et al. 2002). In order to implement IV regression, researchers must identify one or more variables (referred to as instruments) that: 1) have a causal effect on the program variable (a dichotomous variable coded as 1 = participant and 0 = non-participant), 2) affect the outcome variable only through the program variable, and 3) are independent of common causes of the outcome variable (Newhouse, 2005; Hernan & Robins, 2006).

Identifying such variables can be very difficult, especially in DM evaluations that rely mostly on administrative claims data, which limits the number of variables for researchers to select (Linden & Adams, 2006).[1] Linden and Adams (2006) argue that recipient three-digit zip code (i.e., the first three digits of an individual's residential zip code) can be used as an instrument in IV regression. Their argument is based on the theoretical supposition that a valid instrument will function as a randomizer by assigning

[1] In this study, administrative claims data represent records of the health services that Medicaid recipients received from providers. More specifically, claims data represent the electronic versions of bills submitted by health care providers for providing services to Medicaid recipients (Piecoro, Wang, Dixon, & Crovo, 1999; Wyant & Parente, n.d.).

subjects to the treatment and control groups irrespective of the outcome variable (Leigh & Schembri, 2004). While Linden and Adams' (2006) analysis suggests that three-digit zip codes are feasible, there may be reason for some skepticism. For instance, the IV effect estimate will be biased if the zip code instruments fail to meet conditions 1 and 3, which are generally not statistically verifiable. The effect will also be biased if conditions 1 and 3 fail and the association between zip codes and program participation is weak (Hernan & Robins, 2006). Due to the importance of identifying rigorous statistical methods that can be applied to DM evaluations, research needs to be conducted on the zip code instruments to determine if they can generalize to other DM populations while meeting the IV regression assumptions.

Purpose of the Study

The problem addressed in the present study is whether IV regression using the three-digit zip code instrumental variables procedure can generalize to a group of Virginia Medicaid diabetes DM patients. In other words, the main purpose of this study is not to evaluate the effectiveness of the Virginia *Healthy Returns*^{SM} DM Program or to determine the appropriateness of IV regression in general, but rather to assess the feasibility of an instrumental variables procedure developed by Linden and Adams (2006) for DM evaluations. The study is motivated by the problem of using observational data in lieu of experimental data to draw causal inferences about program effects. Specifically, the purpose of the study was to compare OLS regression and IV regression using patient three-digit zip codes as instrumental variables, to use these procedures to estimate the effects of high intensity DM program participation on three outcome variables, and then to determine which procedure offered the best means for estimating program participation effects when analyzing observational data.[2] Implications of the analytical findings are provided in Chapter 5.

Rationale and Significance of the Study

The article by Linden and Adams (2006) provides the rationale for this study. In the article, Linden and Adams argue that IV regression using patient three-digit zip codes as instrumental variables can be used to provide unbiased estimates of the effects of program participation in DM evaluations.[3] A key aspect of IV regression for program evaluation is that the researchers must be able to find one or more variables that are predictive of an individual's program participation, but not associated with any unobserved confounding variables that influence the outcome. This task can be very difficult when dealing with administrative claims data due to the limited number of variables that are available. Linden and Adams (2006) hypothesize that patient zip codes are appropriate instruments to use when employing IV regression for DM evaluations for two reasons: 1) residing in a DM covered area makes the individual eligible for participation, but does not guarantee that the individual will participate, and 2) living in a given zip code may be independent of unobserved covariates. While these covariates are often unmeasured or

[2] Generally, researchers test the feasibility of IV regression models by comparing them to OLS regression models which are usually viewed as being more efficient. This is often accomplished by testing whether the treatment variable is correlated with the error term. Rejecting the null hypothesis of no correlation (or endogeniety) suggests that there is sufficient difference between the OLS and IV coefficients to reject the OLS model in favor of the IV model (Hadley et al., 2003).

[3] According to Linden and Adams (2006), IV regression can be used to derive an unbiased estimate of program participation effects. However, Angrist and Krueger (2001) and Wooldridge (2002) argue that while IV estimators are consistent, they are not unbiased because they represent a ratio of random quantities. Nevertheless, because this study sought to assess the feasibility of Linden and Adams' procedure, their terminology was adopted.

even unimagined, examples may include motivation, illness level, health status, and self-care diligence (Diamond, 1999; Greenland, 2000; Fetterolf & Olson, 2008).

To test their hypothesis, Linden and Adams (2006) collected one year's worth of claims data (excluding pharmacy claims) for a group of diabetes DM patients and then estimated a series of OLS and IV regression models for different outcome variables.[4] They compared parameter estimates produced by both models and concluded that IV estimation using patient zip codes as instruments can be useful when OLS estimates are influenced by selection bias. However, the authors' note that they failed to perform a sensitivity analysis on their results and that the zip code IV procedure may not generalize to other DM programs. Sensitivity analyses are important in observational studies because they can provide insights into the extent to which the empirical results are sensitive to proxies for the treatment variable (Kennedy, 2003; Rosenbaum, 2005). It should be noted that the article by Linden and Adams (2006) is the only study that the author has found to date that uses IV regression to evaluate DM program effectiveness. This finding is not surprising since IV methods, which are commonly used in econometrics, are still relatively new to medical and epidemiological research (Greenland, 2000; Leigh & Schembri, 2004; Newhouse, 2005).

This study extended Linden and Adams' (2006) work by examining the feasibility of using OLS and IV regression analyses to evaluate DM programs, testing to determine if the zip code IV procedure generalized to other DM program populations, and performing a sensitivity test on the program indicator variable.

Overview of the Literature

Four relevant bodies of literature provide important background information on the study topic: general disease management, patient self-management education and self-efficacy theory, diabetes disease management evaluation, and instrumental variables regression. While a detailed review of these bodies is beyond the scope of this chapter, a cursory review will be provided. The general disease management literature provides the foundation for the relevancy of the present study, while the patient self-management literature provides the theoretical justification for the education component of DM.[5] The diabetes disease management evaluation literature provides insights into the designs that researchers have employed to evaluate these programs, the variables they have analyzed, and their study findings. The IV regression literature provides information on how to estimate IV regression models as well as procedures for checking the adequacy of the models. It also provides information on how other researchers have employed IV regression to reduce the influence of selection bias in observational studies. Additional information on these bodies of literature is provided below.

[4] Linden and Adams (2006) did not indicate why they excluded pharmacy claims from their analysis.

[5] While not discussed in the present study, the counterfactual model of causality provides the theoretical justification for the statistical methods used in this analysis. Briefly stated, the counterfactual is what researchers would ideally like to determine because it represents what would have happened to individuals who participated in the program if they had never received the intervention. However, the counterfactual state cannot be observed because it is not possible for individuals who participated in a treatment to go back in time in order to not participate. As a result, researchers have employed a variety of statistical methods, including OLS, IV, and propensity score regression procedures, in an effort to approximate counterfactual states. Additional information on the counterfactual model of causality can be found in Winship and Morgan (1999), Gelman and Hill (2007), Morgan and Winship (2007), and Schneider et al. (2007).

Overview of Disease Management

Disease management has become one of the latest fads in government and health foundations due to its promise of lowering health care costs by reducing emergency department visits and hospital stays for patients with costly chronic illnesses (Fireman, Bartlett, & Selby, 2004). In fact, approximately 97 percent of US health plans and at least 43 state Medicaid agencies are developing or have implemented disease management programs (Center on an Aging Society, 2004; Afifi, Morisky, Kominski, & Kotlerman, 2007). Chronic diseases are the leading cause of death in the United States, accounting for approximately seven out of ten deaths in 2004. During that time, care for patients with chronic diseases accounted for approximately 75 percent of the $1.4 trillion spent on health care and almost 80 percent of all Medicaid expenditures (Department of Medical Assistance Services [DMAS], 2005a).

Disease management seeks to improve outcomes for individuals with chronic diseases by promoting prevention of disease-related complications and exacerbations using patient empowerment tools and evidence-based guidelines. It strives to improve the overall health of selected populations with chronic conditions by supporting the patient-clinician relationship. Major DM program components include the identification and enrollment of patient populations, the use of evidence-based practice guidelines by patients and health care providers, the provision of services that enhance patient self-care, and patient-provider communication and collaboration (Gillespie & Rossiter, 2003).

Private and public health insurance organizations have developed DM programs in an effort to ease individuals and society of the social, psychological, physical, and economic pressures related to chronic illnesses. The intent behind these programs is to reduce health care costs while improving the quality of care that patients receive. When considering that people who suffer from chronic diseases account for approximately 72 percent of all physician office visits, 76 percent of all inpatient hospital stays, and 88 percent of all prescription drug fills, DM programs offer a means of containing, or even reducing, health care costs by focusing on prevention rather than acute care services for these individuals (DMAS, 2006).

Due to the potential to achieve positive outcomes, many state Medicaid agencies have implemented DM programs for their respective Medicaid populations, which tend to be less educated, poorer, and more likely to suffer from disabilities than the general U.S. population (Gillespie & Rossiter, 2003). In Virginia, the Department of Medical Assistance Services (DMAS) provides DM services to Medicaid recipients who are not enrolled in managed care organizations through the *Healthy Returns*^{SM} Disease Management Program, which became operational in January 2006. As of July 2008, the *Healthy Returns*^{SM} program provided DM services to 5,308 Virginia Medicaid recipients who suffered from either asthma, diabetes, coronary artery disease, congestive heart failure, or chronic obstructive pulmonary disorder (DMAS, 2006; Health Management Corporation [HMC], 2008).

Substantial resource investments have been made in the development of DM programs since the early 1990s. For example, the nation invested more than $1 billion in the development of DM programs and related activities in 1999 alone. Despite the considerable investments that have been made in these interventions, however, the availability of information on their effectiveness and safety is limited. According to some critics, the poor quality of DM outcome information stems from statistical biases, abbreviated study periods, and small patient universes, which may cast some doubt on the success claims of many DM programs (Ofman, Badamgarav, Henning, Knight, Gano, Levan, Gur-Arie, Richards, Hasselbald, & Wingarten, 2004).

Patient Self-Management Education and Self-Efficacy Theory

Chronic diseases often result in debilitating conditions for individuals who suffer from them. However, research shows that certain lifestyle changes such as eliminating tobacco use, improving diet and nutrition, and participating in regular exercise regimens can enhance the physical and mental health of chronically ill patients. Research further suggests that adopting improved health behaviors can delay or even prevent the onset of some chronic illnesses (Lorig, Stewart, Ritter, Gonzalez, Laurent, & Lynch, 1996). Because chronic diseases cannot be cured, conventional medical care that is focused on diagnosing and treating acute health care problems is often inappropriate for chronically ill individuals. Consequently, many individuals with chronic diseases may fail to receive the level of care needed to achieve optimal health outcomes (Funnell & Anderson, 2002).

Due to the inability of the traditional medical system to provide the support and education that individuals with chronic diseases require to effectively care for and live with their illnesses, the responsibility for the daily management of these conditions often falls entirely upon the individuals who have them (Funnell & Anderson, 2002). Disease management (DM) programs attempt to prepare participants for this task by emphasizing patient self-management education (Disease Management Association of America, n.d.).

Self-management education differs from traditional patient health education that focuses on providing individuals with disease-specific information and technical skills (i.e., learning how to monitor blood glucose levels). Self-management education complements traditional patient health education by focusing on developing individual problem-solving skills. In particular, self-management education teaches patients how to identify problems, make decisions, implement appropriate actions, and even alter actions as their diseases or personal circumstances change. The development of patient short-term action plans is a central feature of self-management education. Action plans are short plans that identify behavior changes that patients can realistically make in order to effectively manage their diseases (i.e., walk three times a week before lunch) (Bodenheimer, Lorig, Holman, & Grumbach, 2002; Nuovo, 2007).

An important concept in DM self-management education is self-efficacy theory, which refers to the confidence that individuals have in their ability to change behaviors in order to meet specific goals. Self-efficacy is derived from four sources of information: performance attainment, vicarious experiences, verbal persuasion, and physiological states. Of the four sources of information, performance attainment is the most influential because it is derived from individual experiences of success. Because people tend to avoid tasks that they believe they cannot successfully perform and undertake tasks that they believe they can successfully perform, self-management education uses self-efficacy theory to assist patients with developing tools and strategies to successfully change their behaviors (Lorig et al., 1996).

Action plans are an example of one such tool that DM programs use. The plans are based in part on the role that performance attainment plays in self-efficacy. Self-efficacy theory posits that the successful achievement of the action plans is more important than the content of the actual plans because success will motivate individuals to undertake behavioral changes. An important characteristic of the action plans is that they are not dictated by physicians. Instead, they are developed by the patients as something that they want to achieve. In other words, the plans are intended to give individuals the confidence and motivation needed to manage their diseases effectively (Bodenheimer et al., 2002).

While not a central focus of the present study, self-efficacy theory can provide a means of understanding how self-management education promotes healthy behavioral changes (or disease management behavior) among DM participants. Organizations that purchase DM programs that promote patient self-management education may achieve certain outcomes over time, such as reduced health care costs and utilizations, if patients are knowledgeable, motivated, and confident in their abilities to change unhealthy behaviors that exacerbate their chronic conditions. However, this can only occur if the self-management education curriculum is appropriate for the DM target population. One way to evaluate the effectiveness of DM programs may be to examine the effectiveness of their education interventions in terms of self-efficacy theory (Lorig, Sobel, Rittner, Laurent, & Hobbs, 2001; Siu, Chan, Poon, Chui, & Chan, 2007).

Diabetes-Related Disease Management Evaluations

Because the present study only examined data on a group of diabetes DM patients, this section is limited to diabetes-specific DM program evaluations and/or evaluations of

DM programs that provide services on several chronic illnesses including diabetes. (For clarity, these two program types will be referred to as diabetes-related DM programs.)

The review found that a variety of methods and statistical procedures have been employed in both published and unpublished studies to evaluate diabetes-related DM interventions. For instance, Meigs, Cagliero, Dubey, Murphy-Sheehy, Gildesgame, Chueh, Barry, Singer, and Nathan (2003), Choe, Mitrovich, Dubay, Haywood, Krein, and Vijan (2005), and Landon et al. (2006) used experimental designs to study various diabetes DM elements. In these studies, several statistical procedures were employed to analyze data including significance tests, parametric and non-parametric analysis of variance, and linear and logistic regression models.

While these studies all found that DM interventions tended to produce positive results for the enrolled diabetes patients, they were impacted by a number of shortcomings that may limit their applicability. In particular, Meigs et al. (2003) noted that their study was limited due to inadequate patient self-care data and because some of the participating physicians were unable to consistently integrate the intervention into their normal patient encounters. Choe et al. (2005) reported that their results may not generalize to other sites due to their small sample size and because the intervention was performed in a single suburban university-affiliated clinic. Finally, Landon et al. indicated that their findings are limited because they were forced to rely on subject matching rather than pure random assignment, which limited their ability to control for unobserved confounding variables.

While experimental studies are preferred for determining the effects of health care-related interventions, their widespread use is limited. This results from the fact that employing true experimental designs in health care settings is often very difficult due to various financial, ethical, and practical issues (Harris & Remler, 1998). Thus, the review found that many researchers have used quasi-experimental and non-experimental designs to study DM programs.

The quasi-experimental studies employed a variety of statistical procedures similar to the procedures used in the experimental studies to evaluate DM programs. For example, Christakis, Connell, Richardson, and Maciejewski (2004) used linear, logistic, and Poisson regression models to evaluate Washington State's Medicaid DM program, while Afifi et al. (2007) used the two-part model and propensity scores to evaluate Florida's Medicaid DM program. The two-part model is similar to IV regression and was developed by econometricians in an effort to obtain statistically unbiased outcome estimates for programs in which participation decisions may be related to unobserved variables that influence outcomes (Wendel & Dumitras, 2005). The propensity score procedure is used to balance out the treatment and control groups in terms of their observed covariates when patients are allowed to self-select into programs (Ettner, 2004). Finally, Berg and Wadhwa (2007) used propensity score matching and non-parametric analysis of variance procedures to evaluate a telephonic DM program for elderly diabetic patients in three states using a matched-cohort study design.

While these studies tended to find that DM program participation had positive effects on outcome variables, their findings may also be limited. Christakis et al. (2004) found that the DM program resulted in improved outcomes for patients with kidney disease, asthma, and diabetes, but not for congestive heart failure. Their findings may be skewed because they were unable to control for many of the preexisting differences that existed between the treatment and control groups. Afifi et al. (2007) found that Florida's DM program appeared to reduce hospital stays and emergency department visits for individuals with chronic conditions; however, their findings may be limited if their statistical analysis did not address all sources of unobserved bias adequately. The analysis by Berg and Wadhwa (2007) revealed that a commercially delivered DM program can reduce hospitalizations, while improving self-management activities for diabetes patients. However, their study could be limited because propensity score matching only adjusts for observed differences.

The non-experimental studies used statistical procedures that were similar to the ones used in the experimental and quasi-experimental designs. These studies also produced similar results. Coberly et al. (2007) used trend lines to examine the association between the number of phone calls patients received and healthy behaviors measured as the frequency with which patients obtained required hemoglobin and lipoprotein tests. They argued that these findings suggest that telephonic strategies should be used in order to promote healthy behaviors among diabetics. Mangione, Gerzoff, Williamson, Steers, Kerr, Brown, Waitzfelder, Marrero, Dudley, Kim, Herman, Thompson, Safford, and Selby (2006) used hierarchical mixed-effects regression models to examine the association between quality of care and the intensity of diabetes disease management. They randomly surveyed 8,661 diabetics who were receiving DM services through 63 physician groups located within seven health plans. The analysis revealed that increased use of DM services was significantly related to enhanced diabetes care.

As with the experimental and quasi-experimental designs, the findings from the non-experimental studies may also be limited. Coberly et al. (2007) reported that their outcome variables (number of tests received) do not necessarily proxy for healthy behaviors and that these variables may contain errors because they were developed using administrative claims data and not medical chart reviews which are the preferred source for this information. Mangione et al. (2006) reported that their findings were also limited due to the short participation time period for many of the subjects in the disease management programs (less than three years) and because they did not randomly select the physician groups from which they selected their subjects.

Instrumental Variables Regression and Selection Bias

The instrumental variables (IV) regression literature provided the justification for using the IV method in the present study. The objective of health outcomes research is to estimate the effects of medical interventions on patient health and well-being. The most appropriate study designs for achieving this result are experimental. However, these designs are often not employed for various reasons. As a result, interest is growing among health services researchers in the use of IV regression in non-experimental studies (Posner, Ash, Freund, Moskowitz, & Shwartz, 2002). An IV is an observable variable (or set of variables) that can be used as a proxy for random assignment of subjects to treatment and control groups in order to calculate unbiased (or consistent) program effect estimates (Harris & Remler, 1998).

IV regression is conducted in two stages. In the first stage, the instrument (or set of instruments) is used to predict program exposure status and in the second stage, the differences in the outcome variables are examined as a function of the differences in the predicted exposure status in order to evaluate the causal effect of participation on the outcome variable. Variables that are selected as instruments must not be related to the outcome variable (at least beyond its effect on participation) or be related to unobserved confounding variables that are not included in the model. This lack of association must be acceptable conceptually since it cannot be tested empirically (Posner et al., 2002).[6] The two IV stages are usually executed simultaneously using the two-stage least squares (2SLS) regression procedure. If the stages are performed separately, then incorrect estimates of the model sum of squares and parameter standard errors will be obtained (Linden & Adams, 2006).

The basis for IV regression is as follows. If X and Y are the observed treatment and outcome variables, then their relationship to a third variable Z (the instrument or set of instruments) should also be observed. Z is presumed to be associated with X but not Y except through its association with X. Under this condition, the Z-Y association can be depicted as the product of the Z-X and X-Y associations:

$$\text{Association}_{ZY} = \text{Association}_{ZX} * \text{Association}_{XY}$$

This equation allows for the solution of the X-Y association and is useful when the observed X-Y association is confounded by unmeasured covariates. IV estimation acts as a randomization device if properly executed, thus allowing researchers to estimate the effect of X on Y (Greenland, 2000; Leigh & Schembri, 2004).

[6] However, if researchers have more instrumental variables than are needed, they can test whether the additional instruments are uncorrelated with the disturbance term in the regression model (Wooldridge, 2002).

The importance of a valid IV estimator lies in the fact that it can be viewed as a ratio representing the joint projection of Y and X on Z, thus isolating a specific portion of the covariance between X and Y. If this covariation is to be used as the basis of a valid causal inference of the effect of X on Y, then it must not be attributable to any extraneous variables that cause both Z and Y. The regression coefficient for X must be assumed to be constant for all subjects in the target population in order to justify the causal effect estimate obtained through the covariation as the population-level causal effect of X on Y (Morgan & Winship, 2007).

While IV regression is widely used in program evaluations, its use is somewhat controversial with some critics arguing that the procedure is sensitive to violations of its assumptions and that it fails to produce results that approximate an experimental design (Harris & Remler, 1998; Shadish et al., 2002). Moreover, there is no guarantee that IV regression will produce estimates that are more robust than those produced through OLS regression. For instance, Posner et al. (2002) and Hadley, Polsky, Mandelblatt, Mitchell, Weeks, Wang, Hwang, and the OPTIONS Research Team (2003) examined the effectiveness of early stage breast cancer screening and three breast cancer treatment interventions for elderly women respectively. In these studies, the researchers compared OLS and IV regression models to determine which methods offered the best adjustments for selection bias.

Posner et al. found that both methods produced similar results, which they argued strengthen the credibility of the OLS model. (Researchers often select the OLS model if it produces results similar to the IV model because OLS is considered to be an efficient estimator.) They recommended that researchers use several statistical procedures to determine how selection bias affects their results before deciding upon which method offers the best estimates of program participation effects. Hadley et al. (2003) also found similar results and argued that researchers should consider using both methods to estimate a range of possible outcome effects.

However, other studies have found that IV regression produced better results than standard regression. For example, Fortney, Booth, Zhang, Humphrey, and Wiseman (1998) evaluated an Alcoholics Anonymous (AA) treatment program using both logistic and IV regression to test for the influence of selection bias. (Logistic regression is similar to OLS, but is used when the outcome variable is dichotomous.) Their analysis indicated that the IV model produced better estimates of program effects than the logistic regression model due to the influence of selection bias. Linden and Adams (2006) compared OLS and IV regression models in their diabetes DM analysis. They concluded that IV regression using zip code instruments could control for selection bias. However, they also indicated that researchers should use both OLS and IV regression models and select the one that produces the best estimates.

Research Questions

The present study was guided by the following three research questions:

1. Which statistical method provides the best unbiased estimates of high intensity DM program participation on the outcome variables?
2. Do the parameter estimates and confidence intervals for the predictor variables differ depending upon which statistical method is used?

3. What are the advantages and disadvantages of using OLS and IV regression to evaluate high intensity DM program effectiveness?

The study hypothesis was that the three-digit zip code instrumental variables procedure will provide an unbiased estimate of the effects of disease management program participation on a group of high intensity Virginia Medicaid diabetes patients enrolled in the Virginia *Healthy Returns*^{SM} Disease Management Program.

Methodology

As previously mentioned, the purpose of this study was to compare and contrast OLS and IV regression using three-digit zip code instruments, which are two methods that can be employed to estimate the effects of participation in the *Healthy Returns*^{SM} DM Program for a group of high intensity diabetes patients. Comparisons between the magnitudes, direction, and significance of the treatment effect estimates produced through both methods were made (Hadley et al., 2003). The Hausman specification test was performed to determine if the differences obtained between the OLS and IV regression models were significantly different and the relevancy of the zip code instruments was assessed by using both the Sargon chi-square test and the first stage model *F*-statistics from the IV regressions (Ender, n.d.; Hadley et al., 2003; Greene, 2003; Baum, 2006; Stock & Watson, 2007). Additional information on the study population, data source, and study variables is provided below.

Population

The population of interest for the present study consisted of all Virginia Medicaid diabetes recipients who were continuously enrolled in the Virginia Medicaid *Healthy*

Returns^{SM} DM Program during calendar (CY) year 2007. Individuals eligible for the program are classified as either high intensity or standard intensity participants.[7] Because the *Healthy Returns*^{SM} DM Program operated as an opt-out program during the study period, almost all Medicaid recipients with diabetes were automatically enrolled. Recipients who were identified as high intensity based on their probability of incurring high future medical costs were given the option of participating in the high intensity intervention. Those who agreed were enrolled in the high intensity program, while those who declined were enrolled in the standard intensity program as high intensity on demand patients. (High intensity participants who were enrolled in the standard intensity program were excluded from this study. Additional information on the study population is presented in Chapter 3.)

Standard intensity patients receive DM services such as an initial enrollment phone call, a welcome kit providing detailed information on their respective chronic condition, and quarterly educational newsletters. These recipients also have access to a 24 hour seven day-a-week call line that is staffed by licensed medical professionals. High intensity patients receive the standard services plus regularly scheduled phone calls from HMC nurses who have access to their prescribed plans of care or nationally recognized evidence based guidelines. The nurses use this information to assist the recipients in managing their illnesses. As part of this process, high intensity open recipients complete annual health

[7] Since the data for this study were obtained, the contractor that operates the *Healthy Returns*^{SM} Program for the Virginia Department of Medical Assistance Services changed its classification of DM participants. Participants are now classified as high, moderate, and low risk (HMC, 2008). However, this classification was not used in the present study because the data were collected under the previous administrative structure.

assessments that measure their mental and physical functioning abilities (Health Management Corporation [HMC], 2007).

Enrolling a majority of recipients into the DM program prevents the use of a control group composed of non-participants. Linden, Adams, and Roberts (2005) argue that this issue can be addressed by treating high intensity DM patients as the treatment group and standard intensity patients as the control group. The rationale for this comparison is that the high intensity patients actively receive the intervention, while the standard intensity patients do not (Linden et al., 2005). As a result, the treatment group for this study consisted of the high intensity diabetes patients, while the control group consisted of the standard intensity patients.

Data Source

The analysis data came from the Virginia Medicaid Management Information System (VaMMIS) and consisted of recipient and paid claims records. VaMMIS is the Virginia Department of Medical Assistance Services' (DMAS) administrative database that is used to perform both recipient and provider eligibility/enrollment as well as provider payment functions (i.e., providers are physicians, hospitals, clinics, etc. that participate in the Virginia Medicaid program). All Medicaid records within VaMMIS are maintained as SAS files; however, the data obtained for this study were analyzed using SAS, SPSS, and STATA statistical programs.

Because the analysis data contains protected health information (PHI) (i.e., recipient names, addresses, social security numbers, and Medicaid identification numbers), the researcher was required to enter into a formal written agreement with DMAS as stipulated under the federal Health Insurance Portability and Accountability Act of 1996 before receiving any Medicaid recipient claims data. Under the agreement, the researcher was required to appropriately safeguard all data containing PHI, report to DMAS any misuse of PHI data, provide a copy of the study to DMAS, and either return or destroy all analysis data upon study completion.

Variables

Three outcome variables were examined in the present study: total diabetes costs, total emergency department visits, and total hospital stays (Linden & Adams, 2006). Predictor variables included actual program participation, gender, age, and a propensity score. The rationale for selecting these predictor variables was twofold: 1) they were available, and 2) they (or similar variables) were used in other studies (Christakis et al., 2004; Linden & Adams, 2006, Afifi et al., 2007). Medicaid claims data were collected for the study subjects for the CY 2007 time period and quantitative variables were summed for all subjects (i.e., each subject had one total diabetes-related cost number).

Summary

The purpose of the present study was to compare ordinary least squares and instrumental variables regression using Linden and Adams' (2006) three-digit zip code instrument procedure in order to estimate the effects of high intensity participation in the Virginia Medicaid *Healthy Returns*^{SM} Disease Management Program. The purpose of the comparison was twofold: 1) to test the utility of the zip code instrumental variables procedure and 2) to determine which method offered the best solution to issues involving selection bias in observational research. Future research implications resulting from the analytical findings are discussed in Chapter 5.

Chapter 2

Literature Review

This chapter provides a context for understanding the importance of the present study, the statistical procedures that were used, and the findings and conclusions that resulted from the analysis. Studying methods for evaluating disease management programs is important because chronic disease represents a serious issue that is facing the U.S. health care system. With more than 90 million Americans suffering from chronic illnesses, providing care to these individuals accounts for approximately 75 percent of the nation's health care costs, which total more than $1 trillion annually (Centers for Disease Control and Prevention [CDC], 2005a). However, chronic disease patients often receive inadequate care due to a variety of factors including the prevalence of the diseases among low-income people and the substantial amount of resources that are required to manage the conditions (Lohr, Keeler, Calabro, & Brook, 1986; Bodenheimer, 2000; Center on an Aging Society, 2004).

Since the early 1990s, disease management (DM) has been promoted as a mechanism for improving quality of care for chronic disease patients, while reducing health care costs. In fact, the disease management industry has actively promoted its ability to achieve these ends (Bodenheimer, 2000). Even though millions of individuals currently receive services through DM programs, little evidence exists on the cost effectiveness of disease management or on its ability to improve patient health outcomes (Linden, 2006). These shortcomings may stem from the influence of two factors: 1) the availability of data for evaluating DM program effectiveness and 2) the use of weak observational designs that are subject to internal validity threats such as selection bias (Linden & Adams, 2006, Ofman et al., 2004).

The purpose of this study was to examine the utility of an econometric procedure proposed by Linden and Adams (2006) for evaluating disease management program effectiveness using data collected on a group of diabetes DM patients. The authors argue that the method is appropriate for DM evaluations because it can reduce the influence of selection bias. Additional information on disease management, patient self-management education and self-efficacy theory, diabetes-related disease management evaluations, and the statistical analyses examined in the present study are provided in the following sections.

The literature reviewed in this study was identified through several avenues. First, the instrumental variables (IV) regression references used by Linden and Adams (2006) were obtained. Second, searches for key phrases such as disease management, diabetes disease management, disease management evaluation, instrumental variables regression, instrumental variables regression and disease management evaluation, instrumental variables regression and selection bias, selection bias models, self-efficacy theory and disease management, propensity scores, and observational research were performed on the MEDLINE/PubMed, Psychinfo, Social Sciences Index, and the Cumulative Index to Nursing and Allied Health search engines available through the Virginia Commonwealth University library. Additional information on the study topic was identified in research design and regression/econometrics textbooks.

Overview of Disease Management

Disease management is a coordinated intervention and communication system for individuals with chronic conditions that can be managed effectively through patient self-management efforts (Linden & Adams, 2006). While DM programs were initially implemented by pharmaceutical companies, most DM services are currently provided by disease management companies that sell their services to health maintenance organizations (HMOs), large companies, and public entities such as Medicaid (Bodenheimer, 2000). This section provides information on disease management, state Medicaid agencies' involvement in disease management activities, and the Virginia *Healthy Returns*^{SM} Disease Management Program.

Disease Management

Disease management is a population-based approach to providing health care services to individuals suffering from chronic illnesses. Population-based means that DM applies to all eligible recipients enrolled in a health care program offering these services. Generally, the eligible beneficiaries are identified through a predictive modeling analysis of their medical claims history. The claims analysis is used to evaluate recipients' support needs and to stratify them into high and standard risk levels. DM has evolved from a system that typically focused on opt-in recruitment, where individuals with a chronic condition were invited to participate, to opt-out care where all individuals with a chronic condition are automatically enrolled unless they specifically request to be excluded (Foote, 2003).

Under the disease management paradigm, patients are viewed as individuals who receive consistent medical care from the same providers over the course of their disease rather than as individuals who only receive discrete or fragmentary medical care. Disease management is suitable for chronic illnesses that have large research bases. Due to the amount of information that exists on these diseases, it is relatively easy for health care providers to develop evidence-based treatment protocols and to identify and measure appropriate evaluation outcomes. For these reasons, disease management programs are often developed around conditions such as diabetes, heart disease, cancer, asthma, hypertension, AIDS, angina, and kidney disease (Hunter & Fairfield, 1997).

Disease management generally consists of six main components: 1) population identification processes, 2) evidence-based practice guidelines, 3) collaborative practice models, 4) patient self-management education, 5) process and outcome measurement, and 6) routine reporting and feedback between patients, providers, and health plans. Population identification involves identifying a group of patients with a specific disease and then enrolling those individuals into the DM program. Identification is often accomplished through predictive modeling that uses demographic, health care usage, and expenditure variables to identify patients who are most likely to benefit from program participation. Evidence-based practice guidelines consist of providing participating physicians with disease-specific treatment standards to ensure that they provide patients with consistent care based on the latest clinical evidence guidelines. Collaborative practice involves assembling a multidisciplinary health care team (physicians, nurses, pharmacists, dieticians, etc.) to provide comprehensive health care services to DM patients (Center on an Aging Society, 2004).

Patient education is based on the concept that educated patients receive better care because they are more knowledgeable about their conditions. Appointment reminders, 24-hour call centers, home visits, and counseling are examples of services used to educate patients about their illnesses. Process and outcome measurement must be established prior to DM program implementation. It involves measuring variables, such as health care expenditures, patient satisfaction, and health care service usage, to evaluate the impact of the DM program. Finally, routine reporting and feedback consists of periodic communication between patients, physicians, and other multidisciplinary health care team members to ensure that patients are properly managing their conditions and receiving appropriate levels of care (Center on an Aging Society, 2004).

Because data on diabetes patients were analyzed in the present study, an example of a hypothetical diabetes DM program is presented to illustrate how disease management works. Diabetes is a disease characterized by high levels of blood glucose (or sugar) that occur when individuals are unable to regulate their insulin production (CDC, 2005b). As a result, diabetics must monitor their blood glucose and may take insulin or other drugs to control their conditions. Failure to monitor blood sugar levels can have serious consequences for diabetics, including blindness, limb amputations, stroke, or kidney disease. Thus, patient self-management plays a vital role in controlling diabetes, which makes it an ideal illness for inclusion in a disease management program (Congressional Budget Office, 2004).

A diabetes DM program may work by targeting resources toward improving process outcomes for the enrolled population. For example, a diabetes program may focus on increasing the number of patients who receive regular blood pressure screenings, annual foot and eye exams, annual cholesterol tests, annual kidney function tests, and biannual lab tests for hemoglobin Alc, which is a blood sugar monitoring test. The DM program may attempt to motivate diabetes patients into complying with these periodic measures in order to achieve both short and long-term positive health outcomes. Examples of short-term outcomes may include reductions in enrollee hospital stays and emergency department visits, while long-term outcomes may include reduced rates of amputations, heart attacks, and kidney disease occurrences (Congressional Budget Office, 2004).

The arguments made by the disease management industry on the benefits of DM services may be justified to a degree because some research suggests that these programs can both improve patient quality of care and reduce health care costs. For instance, Ofman et al. (2004) report that many DM programs are associated with improvements in the quality of care received by chronically ill patients, while Gillespie and Rossiter (2003) report that DM programs can reduce costs for Medicaid recipients by almost 33 percent due to reductions in hospital emergency department and urgent care visits.

State Medicaid Programs

Due to disease management's potential for reducing health care costs and improving the quality of care for chronically ill patients, many state Medicaid agencies have become interested in these services (Gillespie & Rossiter, 2003). Disease management appears to be well suited for Medicaid programs because they provide health insurance coverage to individuals who are more likely to be physically disabled and less educated than the general population, which are conditions that may contribute to the development of chronic disease. In fact, more than 60 percent of all adult Medicaid recipients suffer from chronic ailments such as diabetes, hypertension, and asthma. Chronically ill Medicaid recipients require more care than their healthier counterparts (Williams, 2004). In fact, one study found that average annual health care costs for chronically ill Medicaid recipients were approximately $6,672 compared to $432 for recipients without these conditions (Williams, 2004).

According to some observers, increasing health care expenditures have probably been the driving force behind the development of many state Medicaid DM programs (Gillespie & Rossiter, 2003). For instance, national Medicaid expenditures increased from $205.7 billion to $275.5 billion (or by about 25 percent) between fiscal years 2000 and 2003 (Holahan & Ghosh, 2005). Many states anticipate that their Medicaid populations will increase due to current economic conditions, which will force Medicaid programs to compete with other state programs for increasingly limited funding (Gillespie & Rossiter, 2003). As a result, 42 states have either implemented or are planning to implement DM programs along with other initiatives in an effort to be more cost effective (Afifi et al., 2007).

In addition to being classified as either opt-in or opt-out, state Medicaid disease management programs can generally be grouped into one of three models: pay for performance, centers for excellence, or the health outcomes partnership. As described by Gillespie and Rossiter (2003), the pay for performance model involves the enlistment of nontraditional providers in the care of patients with diseases such as Alzheimer's, HIV/AIDS, schizophrenia, and chronic ear infections. The fees paid to nontraditional providers are based on improved patient outcomes or reduced health care costs.[8]

The centers for excellence model focuses on providing services to specific high-cost, high-volume Medicaid recipients through a network of hospitals, physicians, and other health care providers participating in the state Medicaid program. Examples of diseases covered under this model include cystic fibrosis, epilepsy, hemophilia, and sickle cell anemia. Medicaid programs using this model are responsible for determining the number of centers for excellence that will operate within their states. This model requires the centers to have official written documentation on the quality of care that will be provided to Medicaid recipients with specific diseases as well as the outcomes that will be measured. Once the documentation is developed, the Medicaid agency furnishes a single prospective payment to the center to cover the costs for all DM services that are provided to Medicaid recipients. The centers are also responsible for reporting health outcome improvements to the state Medicaid agencies (Gillespie & Rossiter, 2003).

Finally, the health outcomes partnership approach is a model typically used to provide DM services to fee-for-service (FFS) Medicaid recipients (individuals who are not enrolled in a managed care program) who have high-priority diseases such as diabetes, asthma, hypertension, congestive heart failure, and chronic obstructive pulmonary disease.

[8] Since 2004, pay for performance (P4P) has evolved in Great Britain and the United States into a model that promotes value based health care by paying providers financial incentives for achieving certain clinical quality, patient experience, and information technology outcomes (Doran, Fullwood, Gravelle, Reeves, Kontopantellis, Hiroeh, & Roland, 2006; Cutler, Palmieri, Khalsa, & Stebbins, 2007; O'Kane, 2007).

Medicaid programs that use this DM model usually provide claims-based feedback reports, treatment guidelines, and other support systems to help health care providers better serve the Medicaid recipients assigned to them (Gillespie & Rossiter, 2003). The Virginia *Healthy Returns*^{SM} Disease Management Program is an example of such a model because it operates as a partnership between the Department of Medical Assistance Services and a private contractor to provide disease management services to FFS Medicaid recipients.

Virginia Healthy Returns^{SM} *Disease Management Program*

The Department of Medical Assistance Services (DMAS) is the state agency responsible for administering both the Medicaid Program and the State Children's Health Insurance Program (SCHIP) within the Commonwealth of Virginia. Medicaid and SCHIP are public insurance programs that provide health care coverage to qualified low-income individuals. Both programs are financed using state and federal funds and are administered by the state in accordance with rules and regulations promulgated by the federal government (DMAS, 2005b).

Due to the current popularity surrounding DM programs, coupled with the expectation of reduced health care costs, the General Assembly directed DMAS to develop a DM program for FFS Medicaid recipients diagnosed with diabetes, asthma, coronary artery disease, or congestive heart failure (DMAS, 2005b). (Chronic obstructive pulmonary disease was included in 2007.) Although not explicitly stated, the intent of the General Assembly's directive was to improve health outcomes for Medicaid FFS recipients with certain chronic conditions while reducing state Medicaid costs. Examples of health outcomes include patient satisfaction with the treatment program, cholesterol testing rate, use of hospital emergency departments for non-emergency care, number of days hospitalized, diabetes hemoglobin A1c (blood sugar) testing rate, use of appropriate medications, and mortality.

To solicit proposals from vendors interested in administering a Medicaid DM program, DMAS prepared a request for proposals in May 2005 that outlined the program's requirements. The agency received four solicitations from interested vendors and spent several months reviewing them. DMAS eventually awarded the contract to a local disease management vendor on September 22, 2005, and the Virginia *Healthy Returns*^{SM} Program became operational on January 13, 2006. The purpose of the program is to improve health outcomes for Medicaid recipients with certain chronic conditions while saving the State money by reducing overall health care costs for program participants. It is designed to achieve a variety of objectives including an overall reduction in hospital admissions, improper use of hospital emergency departments, and medical expenditures for program participants, increased participant and provider education on managing chronic conditions, and enhanced participant and provider satisfaction with the program (DMAS, 2005a).

Under the current DM contract, the contractor is required to operate the *Healthy Returns*^{SM} DM Program as a voluntary opt-in program for Medicaid recipients who have diabetes, asthma, coronary artery disease, congestive heart failure, or chronic obstructive pulmonary disease. The contractor is required to identify recipients with one of the chronic conditions through an analysis of Medicaid claims data using a proprietary predictive modeling procedure. Once the recipients are identified, the contractor is then required to contact them to determine if they are interested in participating. Recipients who are interested are enrolled, while recipients who are not interested are contacted later to determine if they have become interested in participating (DMAS, 2005a). However, the contractor did not administer the *Healthy Returns*^{SM} Program as an opt-in program.[9] Instead, it administered the program as opt-out and automatically enrolled recipients once they were identified through the claims analysis.

Healthy Returns^{SM} offers Medicaid recipients three main interventions: care management, a 24-hour telephone call center, and use of evidence-based treatment protocols. The care management component consists of the following services: continuing health status assessments of all program participants, educating patients about self-management, monitoring patient compliance with self-management protocols, and providing participants with educational materials on their respective chronic conditions. Care management services are provided through either telephone calls or in-person visits at the participants' residences (DMAS, 2005a).

The call center component consists of a center with a toll-free telephone number that is staffed by licensed medical professionals on a 24-hour, seven day-a-week basis. The medical professionals are primarily available to provide participants with referral numbers needed to obtain services from specialty providers. They are also available to answer any basic health-related questions that the participants may have about their respective chronic conditions (DMAS, 2005a).

Finally, the evidence-based treatment protocol component involves providing the primary care providers of the participating Medicaid recipients with evidence-based

[9] The program has been administered as opt-in since March 2008 (HMC, 2008).

treatment guidelines that are based on the latest scientific literature and are approved by nationally recognized experts. It is anticipated that participating Medicaid recipients will experience improved health outcomes because providers will follow the guidelines when developing their treatment regimens (DMAS, 2005a).

The contractor classifies Medicaid recipients with one of the chronic conditions as either high intensity open, high intensity on demand, or standard intensity based on their predicted risk for incurring future health care costs. When patients are identified as high intensity, the contractor contacts them to determine if they want to receive the high intensity intervention. Those who agree are classified as high intensity open and receive periodic telephone calls, individualized care plans, 24-hour access to program nurses, and quarterly mailings of disease-specific information. High intensity recipients who decline are classified as on demand and receive periodic educational mailings and have 24-hour access to program nurses if they choose to use it (which is not reflected in the claims data). Individuals identified as standard intensity are automatically enrolled and also receive the periodic mailings and have 24-hour access to program nurses (HMC, 2007).

Between January and July 2008, 18,166 Virginia Medicaid recipients were enrolled in the *Healthy Returns*^{SM} program. The breakdown for program participation was as follows: 10,378 asthma participants, 3,823 diabetes participants, 1,615 coronary artery disease participants, 1,478 chronic obstructive pulmonary disease participants, and 872 heart failure participants (HMC, 2008). The number of participants enrolled in the program changes often due to the frequency with which Virginia Medicaid recipients enter and leave the Medicaid program as well as the frequency with which recipients are diagnosed with these conditions.

Patient Self-Management Education and Self-Efficacy Theory

Chronic diseases affect more than 90 million Americans and account for at least half of the nation's total health care costs. The prevalence of chronic illnesses in the US population is growing. In fact, it is estimated that these diseases will affect at least half of the US population by 2020 (Redman, 2005).[10] According to Nuovo (2007), the nation's health care system is ill prepared to deal with patients who suffer from these illnesses because it is heavily oriented toward treating individuals with urgent care needs. Such conditions mean that poorly treated chronic diseases usually lead to personal care limitations, poor quality of life, premature wage losses, and high mortality before the age of 65 for individuals who suffer from them (Marks, Allegrante, & Lorig, 2005).

Because many chronic disease patients receive inadequate care from their health care providers, the responsibility for managing these conditions falls primarily on the individuals who suffer from them. Even if the health care system was structured to provide better care, chronic disease patients would still be responsible for approximately 90 percent of the care that is needed to manage their conditions (Suter, Hennessey, Harrison, Fagan, Norman, & Suter, 2008). As a result, chronically ill patients are the ones who are ultimately responsible for managing their conditions, despite the advice and care that they may receive from health care providers. In fact, patients self-manage their conditions every day by deciding what foods they will eat, how much they will exercise, if they will use tobacco products, and even if they will consume prescribed medications. Unfortunately, some patients do not (or cannot) manage their illnesses as well as others (Bodenheimer et al., 2002).

In response to these realities, the Disease Management Association of America (n.d.) recommends that disease management (DM) programs provide patients with self-management education services in order to teach them how to effectively manage their conditions.[11] Self-management education services are intended to compliment traditional DM patient education services that focus on technical competence and disease-specific information by teaching patients problem solving skills (Bodenheimer et al., 2002). In particular, self-management education services may teach DM participants techniques for dealing with everyday problems such as depression, anger, and fatigue; exercise regimens to maintain strength and cardiovascular fitness; appropriate use of prescription medications; behavior modification procedures; and strategies for evaluating new treatments and decision making (Farrell, Wicks, & Martin, 2004).

DM self-management education emphasizes enabling patients to use the technical information that they learn about their diseases to solve problems that are relevant to the daily management of their conditions. The goal of self-management education is to assist patients with developing a greater sense of confidence in their ability to live with and

[10] The growth in chronic diseases is due in part to the fact that the baby boomer generation is reaching the age of increased chronic disease prevalence (Bodenheimer et al., 2002).

[11] While the Virginia *Healthy Returns*^{SM} Program includes a self-management education component, the information presented in this section is not descriptive of that component. Instead, this information is intended to illustrate how self-management education and self-efficacy theory are applied in DM programs in general.

manage their chronic diseases on a daily basis (Nuovo, 2007). Short action plans that are similar to New Year's Eve resolutions play a key role in DM self-management education. The plans are developed by the patients and propose specific behavior changes that they are confident they can achieve. Examples of such behavior changes may include walking a certain number of minutes each day or losing a certain amount of weight each week. By accomplishing the plans, patients begin to build confidence in their ability to successfully manage their diseases through appropriate behavioral modifications (Bodenheimer et al., 2002; Nuovo, 2007).

Because DM self-management education seeks to instill confidence in patients to make life-long behavioral changes, self-efficacy theory plays a central role in these interventions (Allen, Iezzoni, Huang, Huang, & Leveille, 2008). Self-efficacy, which is based on social learning theory, is viewed as fundamental to successful behavioral change. It refers to the level of confidence that individuals have in their ability to perform specific tasks needed to achieve desired goals or outcomes (Clark & Dodge, 1999; Marks et al., 2005).

Self-efficacy results from the interaction of four sources of information: performance attainment, vicarious experiences, verbal persuasion, and physiological states. Performance attainment exerts the strongest influence on self-efficacy because it concerns the successful completion of a task. Vicarious experiences exert the second strongest influence on self-efficacy and occur when patients gain confidence in their ability through the successes and failures of others who were confronted by similar challenges. Verbal persuasion is the third influence and manifests itself through reassuring words and praises that allow individuals to feel confident in their ability to succeed. Finally, physiological states represent the actual senses that patients feel when they experience success or failure (Wolf, 2006). When successfully implemented, self-efficacy can exert a powerful influence in how individuals with chronic illnesses respond to their attempts to achieve appropriate disease management behavior.

Chronically ill individuals develop self-efficacy when they successfully achieve personal goals such as performing a desired level of physical activity or reducing the amount of fast foods they consume. By successfully achieving personal goals, the patients become more confident in their ability to perform the same behaviors again. This in turn increases the likelihood that they will repeat those behaviors (Clark and Dodge, 1999). This is the end state that DM self-management education seeks to instill in patients through the use of action plans. The ultimate goal of the plans is not so much to ensure that patients undertake difficult major life-long behavioral changes, but rather to ensure that patients identify and undertake simple easy to complete tasks that will eventually lead to these major changes.

Promoting self-efficacy is thus a critical issue for DM self-management interventions. In fact, research suggests that favorable self-efficacy beliefs influence depression, disability, medication use, diet, weight loss, and self-care behaviors (i.e., receiving the recommended number of blood glucose tests or cholesterol tests in a year) among chronically ill individuals (Marks et al., 2005). Moreover, research also indicates that DM self-management programs promoting self-efficacy can reduce health care costs and improve outcomes for chronic disease populations (Farrell et al., 2004).

While self-efficacy provides an important theoretical framework for understanding how DM interventions work, readers should note that it was not the central theme in the present study, which focused on examining statistical methodologies. However, self-efficacy theory could be used to guide DM evaluations. For instance, Lorig et al. (2001) evaluated the effects of a DM intervention that promoted self-management education and self-efficacy on four outcomes: health behavior, self-efficacy, health status, and health utilization. The researchers found that the intervention was associated with statistically significant improvements in health status, health behavior, and self-efficacy as well as a statistically significant decrease in health utilizations. Sui et al. (2007) also evaluated the effects of a self-efficacy based DM program in a Chinese population on various outcomes including self-management behavior, self-efficacy, coping strategies, and health outcomes. They found that participation in a DM program that promoted self-management education was associated with statistically significant increases in self-efficacy, coping strategies, and health outcomes.

Diabetes-Related Disease Management Evaluations

While the Virginia *Healthy Returns*^{SM} Disease Management Program currently covers five chronic diseases, the present study only examined data on diabetic Medicaid recipients. Diabetes was selected because it is important for policymakers to understand how disease management effects participants who suffer from this disease because it is a serious condition that has reached epidemic proportions, affecting approximately 18 million Americans (Shojania, Ranji, McDonald, Grimshaw, Sundaram, Rushakoff, & Owens, 2006; Choe et al., 2005). Diabetes is the fifth leading cause of death in the country and contributes to morbidity by placing people at increased risks for heart disease, blindness, and other chronic conditions. Substantial costs are also associated with diabetes. For instance, it is estimated that diabetes cost the nation approximately $132 billion in health care expenditures and lost productivity in 2002 (American Diabetes Association, 2003). Diabetes places substantial clinical and economic burdens on American society (Knight, Badamgarav, Henning, Hasselblad, Gano, Ofman, & Weingarten, 2005), but studies suggest that self-management services such as intensive glucose control can substantially reduce diabetes complications (Rothman & Elasy, 2005). For these reasons, it is important to study methods that can be used to evaluate the effectiveness of diabetes-related DM interventions.

The literature review is not limited to evaluations of specific diabetes DM programs. Instead, it includes evaluations of both diabetes specific and related DM programs and services to develop an understanding of the research designs and statistical methods that have been used to evaluate this topic. A variety of designs and statistical procedures have been employed. However, only two studies were identified that used statistical methods similar to the ones that were used in this study. While the studies tended to find that diabetes DM produced positive outcomes for patients, they encountered various methodological shortcomings that limited their usefulness. Additional information on the reviewed diabetes DM studies is provided in the subsections below.

Experimental Designs

The experimental design with random assignment of subjects to treatment and control groups is considered the gold-standard in research and program evaluation studies.

Random assignment reduces the influence of selection bias and other internal validity threats by giving each subject in the population of interest an equal chance of being included in the study groups. However, implementing experimental designs can be difficult unless they are conducted in controlled settings (Linden et al., 2005).

Five diabetes-related DM evaluations using experimental designs were identified in the literature review. The statistical procedures used in these studies come from the generalized linear model (GLM), which provides the foundation for most of the statistical analyses performed by social science researchers. The GLM underlies statistical procedures such as the *t*-test, analysis of variance (ANOVA), analysis of covariance (ANCOVA), regression analysis, factor analysis, cluster analysis, and discriminant function analysis. The GLM is important because it allows researchers to summarize a variety of outcomes. However, model specification is a major limitation faced by researchers who use the GLM because they must specify statistical models that best summarize their data. If important variables are not included in the models, then the parameter coefficients will be biased, which will result in statistical equations that do not correctly describe the data (Trochim, 2005).

GLM procedures used in the reviewed studies included *t*-tests, ANOVAs, regression analyses (ordinary least squares, logistic, and Poisson), and generalized estimating equations (GEE). GEE is a specialized regression procedure that is used to analyze longitudinal and other correlated data, especially when the outcomes are binary or count variables (Hanley, Negassa, Edwards, & Forrester, 2003). Odegard, Goo, Hummel, Williams, and Gray (2005) and Meigs et al. (2003) used *t*-tests and GEE procedures to evaluate the effectiveness of a pharmacist intervention on diabetes self-management, and to test the effects of a web-based decision support tool to improve the management of diabetes patients using evidence-based treatment protocols, respectively. Outcome variables examined included hemoglobin A1c tests (used to monitor blood sugar levels), use of appropriate diabetes control medications, cholesterol tests, blood pressure, and eye and foot examinations.

Odegard et al. (2005) found that the pharmacist intervention did not significantly improve A1c test adherence for the treatment group and that no significant differences existed between the study groups on the use of appropriate medications. Meigs et al. (2003) concluded that the web-based decision support tool for diabetes management appeared to have the potential to improve evidence-based care of diabetes patients. However, both studies had limitations that may influence the usefulness of their results. For instance, Odegard et al. (2005) reported that their results may be influenced by regression to the mean because they specifically focused on diabetes patients with poor self-management skills. In addition, they randomized subjects within clinics as opposed to randomly selecting clinics for the study. As a result, providers who participated in the experiment could have provided services to both treatment and control subjects that may have caused treatment diffusion (i.e., treatment effects being spread to control group members) (McMillan & Schumacher, 2006).

The study by Meigs et al. (2003) also had several limitations. In particular, it was plagued by incomplete documentation that may have resulted in lower rates of patient compliance for some diabetes care activities used as outcome variables. Meigs et al.

(2003) also reported that some participating providers did not consistently use the web-based tool as intended when treating diabetes patients, which may have further skewed their results.

The GLM procedures used by Landon et al. (2007), Choe et al. (2005), and Sadur, Moline, Costa, Micalik, Mendlowitz, Roller, Watson, Swain, Selby, and Javorski (1999) included *t*-tests, and ordinary least squares (OLS), logistic, and Poisson regression analyses. Landen et al. (2007) used OLS and logistic regression to examine changes in disease-specific quality of care indicators for 9,658 randomly selected diabetes, asthma, and hypertension patients who received care at 64 (44 treatment and 20 control) community health centers. Diabetes-related indicators that served as outcome variables included A1c tests, cholesterol tests, and blood pressure. Choe et al. (2005) used OLS and logistic regression to examine treatment and control group differences in an evaluation of the effectiveness of pharmacist-provided DM services on blood sugar and preventive care measures for a group of randomly assigned diabetes patients. Outcome variables included A1c tests, eye examinations, and urine microalbumin tests. Finally, Sadur et al. (1999) used *t*-tests and Poisson regression to examine differences between the treatment and control groups on various outcome measures including A1c tests, self-reported measures of self care practices, satisfaction with general medical care, and health utilizations (e.g., number of visits to hospital emergency departments and visits to physician offices).

Landon et al. (2007) found that the diabetes and asthma patients who received care from the community health centers demonstrated greater improvements in the monitoring and treatment of their conditions, while hypertension patients did not. Choe et al. (2005)

found that proactive diabetes case management provided by a pharmacist substantially improved glycemic control for diabetes patients, and Sadur et al. (1999) determined that DM services improved glycemic control, self-efficacy, and patient satisfaction for diabetes patients, while reducing their health care utilization.

However, these studies also encountered limitations. Landon et al. (2007) reported that they had to rely on matching treatment and control subjects because they were unable to perform a pure randomized trial where subjects were randomly assigned to study groups. Thus, their analysis may not have accounted for all potential confounding variables, which could bias their regression models. They also reported that their results may have been overstated because some of the community health centers may not have fully implemented the intervention. Choe et al. (2005) reported that the generalizability of their study was hampered due to its small scope (it only involved 80 patients at one site), while Sadur et al. (1999) reported that they failed to obtain complete information on all study subjects, which could influence the accuracy of their findings.

Quasi-Experimental Designs

According to Trochim (2005), quasi-experimental designs are similar to experimental designs, but lack random assignment of subjects to treatment and control groups. These designs are more commonly used than experimental designs due to the difficulties that researchers encounter when implementing experimental studies (Trochim, 2005). Six quasi-experimental studies were identified in the literature review. The studies employed a variety of statistical procedures to examine the effects of DM interventions on various outcome variables. While the studies tended to find that diabetes DM produced positive results, limitations were present that may reduce their usefulness.

Two studies were identified that used GLM procedures similar to instrumental variables (IV) regression. In the first study, Wendel and Dumitras (2005) tested the feasibility of using a two-part selection bias regression model for DM program evaluations. The model was developed by James Heckman (an econometrician) in the 1970s to control for selection bias in observational studies. In the first step, probit regression is used to create a selection model that predicts the probability of program participation using observable patient characteristics. The residuals from this equation are then used to create the Inverse Mills Ratio (IMR), which represents the predicted probability of program enrollment. In the second step, linear or probit regression is used to produce unbiased estimates of the impact of program participation on the outcome variable, while controlling for the IMR and other observed confounders (Ettner, 2004; Sales, Plomondon, Magid, Spertus, & Rumsfeld, 2004). The regression steps are performed simultaneously. The Heckman method adjusts for potential selection bias by including the IMR as a variable in the outcome equation in order to model the correlation between the treatment variable and the error term (Ettner, 2004).

To perform the study, Wendel and Dumitras (2005) obtained administrative data from a managed care diabetes DM program. Because they were interested in assessing the change in health care costs generated by DM program participation, they used costs as the outcome variable. Their analysis found that DM program participation had a statistically significant effect on the outcome variable. Based on the analysis, the authors concluded that the Heckman method offered a viable means of evaluating a DM program using observational data. While Wendel and Dumitras (2005) did not report any study limitations, their analysis may be limited because the Heckman method is highly sensitive to model specification. As a result, they could have calculated biased (e.g., incorrect) estimates if important variables related to either recipients' participation decisions or the outcome variable were omitted (Sales et al., 2004).

In the second study, Afifi et al. (2007) used a variant of the two-part model to evaluate the effects of the Florida Medicaid disease management program on four utilization outcomes: number of annual inpatient hospital stays, emergency department visits, inpatient days, and outpatient visits. The first part of their modeling procedure used logistic regression to predict whether a recipient had any utilization, while the second part applied only to recipients who utilized one of the four health care services. OLS regression was used in the second stage to model the logarithm of the annualized utilization variables. According to Mullahy (1998), the two-part model is an appropriate procedure to use in health outcomes research settings (such as disease management) where information may only be available on subjects who utilized particular services.

As part of the analysis, Afifi et al. (2007) also calculated propensity scores for the treatment and control group subjects for use as a covariate in the regression analysis. The propensity score (which is estimated using logistic regression) is a method that corrects for selection bias. It differs from IV regression in that it can only control for bias resulting from observed variables, while IV corrects for bias due to unobserved variables. The propensity-score method allows researchers to match subjects on the observed covariates that may account for bias in the sample (Schneider, Carnoy, Kilpatrick, Schmidt, & Shavelson, 2007).

Based on their analysis, Afifi et al. (2007) found that the Florida DM program appeared effective at reducing hospital inpatient and emergency department visits for diabetes, congestive heart failure, and asthma patients, but not for hypertension patients. They concluded that some cost savings might be associated with the DM program because hospital stays and emergency department visits are two of the most expensive components of care. Their analysis further revealed that DM services do not appear to be effective until after recipients have been exposed to the treatment for some time. However, Afifi et al. (2007) reported that their analysis could have been hampered by selection bias because the propensity score procedure does not control for unobserved confounding variables. If any of the variables in their regression models were associated with unobserved confounders, then their parameter estimates could be biased.

Another Medicaid study identified during the literature review involves an evaluation of a Virginia Medicaid disease management program. Zhang, Wan, Rossiter, Murawski, and Patel (2008) employed GLM procedures to evaluate a DM program that was operated by DMAS prior to the implementation of the *Healthy Returns*^{SM} DM Program. This program was known as the Disease State Management (DSM) Program and it operated between 1999 and 2001. Chronic diseases covered under the DSM Program included diabetes, hypertension/congestive heart failure, depression, peptic ulcer disease, and asthma/chronic obstructive pulmonary disease. Zhang et al. (2008) used analysis of variance and covariance procedures to evaluate the effects of the program on several outcome variables including drug compliance, quality of life, hospitalizations, physician office visits, and emergency department visits. They also performed a cost savings analysis as part of their study. Their analysis indicated that the DSM program significantly improved participants' drug compliance and quality of life, while reducing hospitalizations, physician office visits, and emergency department visits. They also estimated that the program saved the Commonwealth of Virginia approximately $3 million. However, they reported that their analysis may be influenced by two limitations: selection bias because participants were allowed to self-select into the study groups and attrition because many of the control subjects left during the study.

The studies by Berg and Wadhwa (2007), Christakis et al. (2004), and Villagra and Ahmed (2004) also used GLM procedures to evaluate diabetes DM effects. For instance, Berg and Wadhwa (2007) evaluated a telephonic DM program for elderly diabetes patients using *t*-tests and propensity scores. Outcome variables consisted of utilization measures including number of hospitalizations, emergency department visits, and physician visits. In an unpublished study, Christakis et al. (2004) used OLS, logistic, and Poisson regression to evaluate Washington State's Medicaid DM program for patients with kidney disease, asthma, congestive heart failure, and diabetes. While outcome variables were examined for all four diseases, the diabetes specific outcomes were emergency department visits, number of hospitalizations, length of hospital stays, A1c tests, and eye tests.

Finally, Villagra and Ahmed (2004) used *t*-tests and an intention-to-treat analysis to evaluate the first year effects of a multistate diabetes DM program on several outcome variables including health care costs, emergency department visits, and number of days hospitalized. Because the DM programs were phased in over a three-year period, researchers used the sites where the diabetes DM programs operated as historical controls.

Berg and Wadhwa (2007) found significant differences between the treatment and control groups on health care service utilizations, prescription drugs, and diabetes testing procedures. They concluded that a commercially delivered diabetes DM program could significantly reduce hospitalizations, while increasing the use of diabetes related prescription drugs and clinical procedures. Their study may be limited because they did not have complete information on the control group members, which was needed in the analysis.

Christakis et al. (2004) also found that DM services were associated with improved care for Medicaid patients with kidney disease, asthma, and diabetes, but not for patients with hypertension. They reported that their findings might be hampered by selection bias because patients were not randomized into the DM program. Finally, Villagra and Ahmed (2004) determined that the overall costs of care were significantly lower for the DM subjects in their study. They also determined that quality of care (e.g., A1c testing, diabetes control drugs, and eye exams) was significantly better for the DM subjects. Villagra and Ahmed (2004) reported that their study was limited by its design, which may not have controlled for all biases and confounders that could influence the results, and that the data were susceptible to regression to the mean.

Non-Experimental Designs

Non-experimental designs are used when researchers are unable to manipulate variables, or are interested in either describing various phenomena or examining relationships. Examples include descriptive, relational (comparative and correlational), and causal-comparative designs (McMillan & Schumacher, 2006). Five non-experimental studies were identified for review. These studies also tended to find that diabetes DM interventions produced positive results for participants. The studies employed GLM procedures such as *t*-tests, ANOVAs, generalized estimating equations (GEE), mixed-methods regression models, and logistic regression models. In addition, one study performed a trend line analysis, while another study used a cost-comparison approach to evaluate a DM program. These studies are discussed in more detail below.

Morisky, Kominski, Afifi, and Kotlerman (2008) used both generalized estimating equations and linear mixed-method regression models to estimate the effects of the Florida Medicaid DM program on behavioral health and physiological limitations for participants with congestive heart failure, hypertension, diabetes, and asthma. Their outcome variables included mental and physical health assessment scores, medication compliance scores, blood glucose levels, and cholesterol scores. Coberly, McGinnis, Orr, Coberly, Hobgood, Hamar, Gandy, Pope, Hudson, Hara, Shurney, Clarke, Crawford, and Goldfarb (2007) performed a trend line analysis of a DM program to determine the relationship between DM telephonic contact and increased clinical testing rates. The outcome variables were A1c and cholesterol testing rates for patients during their first 12 months of participation in a diabetes DM program. Mangione et al. (2006) used mix-effects regression models to determine whether DM services provided by physician groups are associated with improved diabetes care processes and control of diabetes outcomes. Their main outcome variables included blood pressure, A1c tests, and cholesterol tests.

Krein and Klamerus (2000) used logistic regression to determine if patients who were enrolled in the Michigan Diabetes Outreach Network program received recommended diabetes-level care. Their main outcome variable was whether or not the patients received the recommended level of diabetes care. Finally, Fireman et al. (2004) analyzed financial data from a commercial DM program to determine if it had produced any cost savings. Variables examined in the study included number of emergency department visits, hospital admissions, hospital days, and clinic visits.

Four of the reviewed studies concluded that DM services were beneficial. Morisky et al. (2008) found that participation in the DM program was associated with increased health behaviors conducive to better health outcomes such as a reduced rate of smoking and adherence to medical regimes for hypertension and diabetes patients. Coberly et al. (2007) found a positive relationship between the frequency of telephone contact and increased Alc and cholesterol testing among diabetes DM patients. Mangione et al. (2006) determined that diabetes DM strategies were associated with better diabetes care, while Krein and Klamerus (2000) concluded that diabetes patients were more likely to receive Alc and eye exams after participating in the program. However, Fireman et al. (2004) reported that their analysis did not reveal any evidence of cost savings. Thus, they concluded that the rationale for DM programs should rest on effectiveness and value rather than the potential to reduce costs.

Despite these findings, however, the five studies also had limitations. For example, Morisky et al. (2008) used data that was subject to social desirability bias because it was collected through self-reported behavioral assessments. Coberley et al. (2007) reported that they used metrics that were less comprehensive than those recommended by national guidelines for evaluating health outcomes. Moreover, they relied upon administrative claims data as opposed to medical records (which are preferred) for measuring clinical testing outcomes. Thus, their findings may be distorted. Mangione et al. (2006) reported that their statistical analysis was not sensitive enough to detect modest associations, which may have prevented them from detecting treatment effects on all outcome variables.

Krein and Klamerus (2000) indicated that their study was unable to determine how much of the increase in the recommended level of diabetes care received by patients was attributable to the program and not to other factors. They further reported that their study was plagued by missing data. Finally, Fireman et al. (2004) reported that their findings may be biased because they did not have access to a control group composed of similar chronic disease patients which was needed in order to make meaningful comparisons. They also indicated that they lacked data on their subjects' functional status and work productivity, which would have allowed them to determine if these factors had improved over time while the subjects were enrolled in the DM program.

Conclusions

The primary purpose of this section was twofold: 1) to identify some of the designs, statistical procedures, and variables that researchers have used to evaluate DM programs, and 2) to determine if any researchers have employed IV or similar regression procedures to evaluate these programs. By addressing these purposes, additional justification for the analytical procedures that were employed in the present study can be obtained.

This review found that various experimental, quasi-experimental, and non-experimental designs were employed for the evaluations, and that most of the statistical procedures used came from the GLM. The review also revealed that some researchers have used the propensity score method to control for selection bias when evaluating DM programs. In addition, the review revealed that while many studies use clinical outcome measures, such as A1c and cholesterol testing rates, a number also use cost and utilization measures, such as hospital emergency department visits, as outcome variables. Finally, the review found that no evaluations were performed using IV regression, which is not surprising because this procedure has only obtained limited popularity among health services researchers (Newhouse, 2005). However, IV regression, which will be discussed further in the next section, is an appropriate procedure to use when analyzing data influenced by selection bias (Basu, Heckman, Navarro-Lozano, & Urzua, 2007).

While the primary justification for the analytical procedures that were used in this study come from Linden and Adams (2006), the reviewed articles provide additional justification for the study's design, statistical procedures, and outcome variables. In particular, the analysis was performed using a quasi-experimental design with treatment and control groups (e.g. Afifi et al., 2007), and program effects were estimated using the propensity score method and OLS and IV regression (e.g., Landon et al., 2007 and Afifi et al., 2007) to control for selection bias due to observed and unobserved confounders. The outcome variables were total diabetes costs, hospital emergency department visits, and hospitalizations for a 12-month time period (Berg & Wadhwa, 2007, Christakis, et al., 2004, and Villagra & Ahmed, 2004). The reviewed articles provide the justification for the study's design, statistical procedures, and outcome variables.

Statistical Models for Program Evaluation

The present study examined two methods that could be used to evaluate the effects of participation in a disease management program for Medicaid diabetes recipients. Because the analysis data came from administrative claims, this study was conducted as an observational (or quasi-experimental) study. According to Rosenbaum (1995), an observational study is an empirical investigation that seeks to derive cause-and-effect relationships under situations where controlled experimentation is not feasible. Observational studies focus on treatments, interventions, and public policies and their associated effects. Non-experimental studies that lack these characteristics are not considered observational (Rosenbaum, 1995).

In experiments, researchers randomly assign subjects to treatments. Subjects who receive the treatment form the experimental (i.e., program, intervention, or treatment) group, while subjects who do not receive the treatment form the control group. Randomization ensures that the treatment and control subjects are comparable. However, researchers conducting observational studies cannot control subject assignments. Because researchers lack this ability, systematic differences (i.e., selection bias) may exist among the study subjects (McMillan & Schumacher, 2006). Comparing outcomes across subjects in an observational study may subsequently confound the effects of the treatment with the effects of the subjects' preexisting differences. More specifically, the estimated treatment effect may be biased. Ordinary least squares (OLS) and instrumental variables (IV) regression are two generalized linear modeling procedures that can be used in observational studies to derive unbiased treatment effect estimates (Haro, Kontodimas, Negrin, Ratcliffe, Suarez, & Windmeijer, 2006). Additional information on these procedures is provided in the following subsections.

Ordinary Least Squares Regression

OLS regression is a widely recognized statistical methodology that involves relating a quantitative outcome variable to one or more quantitative and/or qualitative predictor variables. The end result is a mathematical model that predicts the outcome variable for the given set of predictors (Mendenhall & Sinich, 2003). Equation 1 depicts a bivariate (or simple) regression model.

[1] $Y_i = \alpha + \beta X_i + u_i$

In this equation, Y is the outcome variable (annual diabetes costs), X is the predictor variable (high intensity open DM program participation coded as 1 = participant and 0 = non-participant), *i* refers to the *i*th individual (any individual in the study), α is the Y intercept (or the value of Y when X = 0), β is the slope of the regression line (the change in Y when X increases by one, or in causal terms, the effect of X on Y), and u is the disturbance or error term. This term is assumed to have a mean of zero and to be randomly distributed across study subjects (Mohr, 1995). In a regression model, only one α is calculated; however, a β (parameter estimate or coefficient) is calculated for each predictor variable that is included.

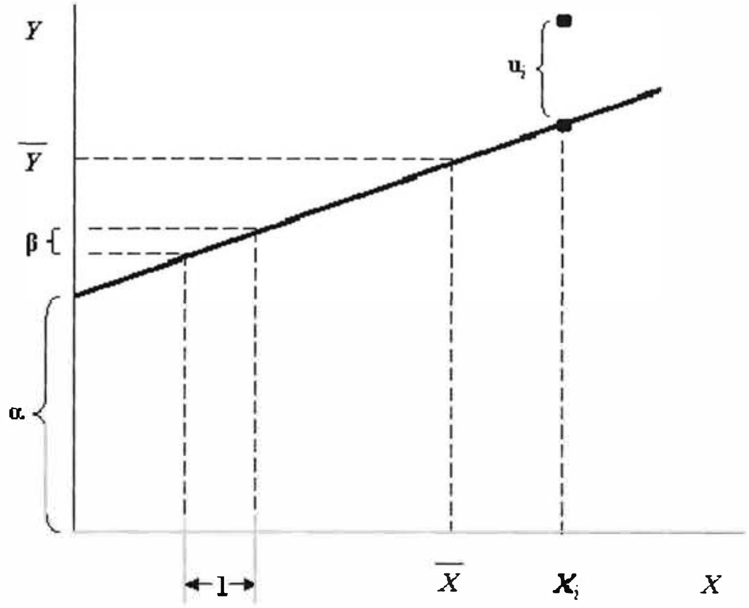
When a simple regression analysis is used to model the effects of a treatment (or program), the two regression parameters – α and β – have special meanings. For a treatment variable with only two categories, the regression line (β) will always pass through the mean of Y for each category of the predictor. If the treatment variable equals zero, then α represents the mean outcome value for the control group. If the treatment variable equals one, then $\alpha + \beta X_i$ represents the mean outcome variable for the treatment group. Because β denotes the causal effect of the treatment variable, X, on Y, β is the treatment effect (subject to selection and other internal validity threats). The discovery of β is the purpose of a summative program evaluation study. Moreover, because β is defined as the change in Y when it's associated X variable increases by one, β is the mean value of the outcome variable for the treatment group minus the mean value of the outcome variable for the control group (Mohr, 1995). An example of a bivariate regression is presented below in Figure 1.

The bivariate regression model can be extended by including additional predictors (or covariates). When additional variables are included, the procedure is known as multiple regression, which is illustrated in Equation 2.

[2] $$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_n X_{ni} + u_i$$

This equation shows that two or more predictors (up to the *n*th predictor) can be added to the model. In a multiple regression model, the βs are interpreted differently than in a bivariate model. For instance, β_1 represents the mean change in Y for every one-unit increase in its associated X variable, when holding fixed all other predictors in the model. The remaining βs and their respective variables are interpreted similarly (Mendenhall & Sincich, 2003). As another example, consider a model that includes two predictors: treatment program (coded as 1 = participant and 0 = non-participant) and age. Then β_1

Figure 1: Bivariate Regression



Source: Mohr (1995)

(and its respective treatment variable) answers the question, What is the difference in Y (e.g., diabetes costs) for subjects in the different study groups who are the same age? Specifically, it allows for the determination of the treatment effect given the age of the subjects (Mohr, 1995).

Statistical significance tests are performed on the parameter estimates (βs) in the regression model to test whether the estimates differ from zero. If the tests are significant, then it can be assumed that the estimates differ from zero, which indicates that the observed relationship between the predictor and outcome variable is probably not due to chance. A statistically significant test for the treatment variable's parameter estimate would indicate that the treatment had a significant effect on the outcome variable.

The primary function of including additional variables in multiple regression is to reduce the influence of imperfect randomization or even the lack of randomization. If subjects are not randomized, then the researcher must be concerned about important differences that may exist between the treatment and control group subjects, which could explain the observed results (Mohr, 1995). Addressing selection bias through OLS regression works if the researcher can account for all preexisting group differences that are related to the outcome variable. This is difficult to accomplish because program participants often differ in a variety of observable and unobservable ways from non-participants (Weiss, 1998). If the researcher is aware of preexisting differences (and can measure them), then they can be controlled by including them as covariates in the regression analysis. However, if the researcher is unaware of the preexisting differences (or is unable to measure them) then they will not be controlled, which may lead to confounding in the data (Wunsch, Linde-Zwirble, & Angus, 2006).

In regression, observable and unobservable factors that affect the outcome variable but are not included in the model represent the error term in the relationship (Wooldridge, 2006). In observational studies, the error term can include factors such as preexisting differences between the treatment and control groups. Selection bias occurs when these differences (if correlated with one or more of the predictors and the outcome) are not statistically controlled (Salkever, Slade, Karakus, Palmer, & Russo, 2004). In other words, selection bias occurs when at least one of the predictors is correlated with the error term (Ettner, 2004). One assumption of multiple regression is that the error term has an expected (or mean) value of zero given any of the values of the other predictors in the model. Failing to include important variables that are related to the predictors causes this assumption to fail, thus producing biased regression results (Wooldridge, 2006).

OLS regression can be used to estimate treatment effects in observational studies if overt selection bias is controlled (Haro, et. al., 2006). However, OLS regression cannot effectively address hidden selection bias. When hidden selection bias is a concern, researchers typically use other methods to estimate treatment effects such as IV regression (Winship & Morgan, 1999).

Instrumental Variables Regression and Selection Bias

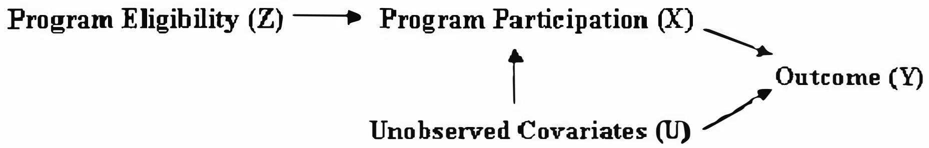
A structural equation modeling technique known as instrumental variables (IV) regression was developed by econometricians in the 1920s to address instances in which the OLS predictor(s) are correlated with the error term (Linden & Adams, 2006; Newhouse, 2005). IV regression was first used to estimate supply and demand curves, but has since been applied to issues involving measurement error and omitted variables bias (Schneider et al., 2007). In fact, IV regression is widely used to control for selection bias in observational studies (Basu et al., 2007). The popularity of IV regression in observational studies stems from its ability to estimate unbiased relationships between outcome variables and predictors by purging the predictors of the portion of their variance that is not independent of the error term.

In regression, a predictor variable that is correlated with the error term is known as an endogenous variable (Winship & Morgan, 1999). In a program evaluation context where hidden selection bias may be suspected (e.g., a situation where subjects self-selected into the program or were otherwise assigned nonrandomly), a good IV would be a variable that is associated with the treatment variable, but is uncorrelated with the omitted variables and has no association with the outcome, except through the treatment variable. Because the IV is related to the treatment variable, but is uncorrelated with other predictors of the outcome variable, the causal effect of the instrument on the outcome is therefore proportional to the causal effect of the treatment on the outcome (Schneider et al., 2007).

Based on this information, an IV must satisfy two assumptions: 1) it is independent of the error term in the regression model, and 2) it affects the problematic predictor variable (i.e., the treatment variable that is influenced by selection bias), but not the outcome. If the first assumption fails (that is, if the IV actually affects the outcome directly or if there are no observed variables that do not affect the outcome directly), then the results of the IV regression will be biased and the effect of the treatment will be unidentified. If the second assumption fails (that is, if the IV's variation does not produce much variation in the treatment variable), then the random error term may mask the effect of the treatment variable. If this occurs, then IV regression will produce results similar to OLS regression. Thus, these two assumptions must be satisfied in order for IV regression to produce useful results (Newhouse & McCellan, 1998). Essentially, an IV is a variable (or set of variables) (*Z*) that is correlated with the program variable (*X*), but not related to unobserved confounding variables of the outcome variable (*Y*).[12] Thus, the IV can only impact the outcome variable through the program intervention. This relationship is graphically depicted in Figure 2.

[12] More than one variable can be used as an instrumental variable (IV). When this occurs, the IV can be referred to as the set of instruments.

Figure 2: Instrumental Variable Estimation



Source: Newhouse and McClellan (1998); Linden and Adams (2006)

IV regression is performed using a structural (or simultaneous) equation modeling process known as two-stage least squares (2SLS) regression (Linden & Adams, 2006). In the first stage, the instrument (or set of instruments) and any covariates are used to predict the endogenous variable in a regression equation. In the second stage, the outcome variable is regressed on the fitted values from the first stage regression plus any covariates. If the IV is uncorrelated with the omitted variables, the predicted value of the outcome is also uncorrelated with the omitted variables. Thus, the bias in estimating the outcome variable from the exclusion of the variables that account for preexisting differences from the model is eliminated (Schneider et al., 2007).

2SLS regression is formally presented in equations 3 and 4.

[3] $X\text{–hat} = \alpha_0 + \alpha_1 Z_i + v_i$

[4] $Y = \beta_0 + \beta_1 X\text{–hat}_i + \varepsilon_i$

In equation 3, Z represents the instrument (or set of instruments) that is used to estimate X – hat, which is the predicted value of X. This value is then used in equation 4 instead of the actual X variable. It is assumed that Z is a significant predictor of who is likely to participant in the program (X – hat). If the actual treatment variable was used (indicating if the subjects actually participated), the result would be confounded due to selection bias. Using X – hat given Z allows for an unbiased estimate of the program's impact on an outcome because Z predicts X in equation 3, but remains independent of the X – Y relationship. In the second stage, IV regression essentially becomes OLS regression and its results are interpreted in a similar manner. It is important to note that IV regression must be performed using 2SLS. Using OLS twice to estimate each stage results in incorrect estimations of the model residual sum of squares and standard errors (Linden & Adams, 2006).

Generally, researchers estimate both OLS and IV regression models and compare them using a Hausman specification test to determine whether significant differences exist between the models (Ender, n.d., Greene, 2003, Hadley et al., 2003; Baum, 2006). If significant differences do exist, then the IV model is normally accepted because the OLS estimates are presumed to be biased. If the test is not significant, then the OLS results are accepted because this procedure produces more efficient parameter estimates than IV regression. This is due to the fact that IV estimates always have larger variances than the OLS estimates. In fact, sometimes researchers may even prefer the biased OLS estimates if they have smaller mean squared errors compared to the IV estimates (Winship & Morgan, 1999). In other words, IV regression does not always offer a good solution to selection bias issues in observational studies.

A number of articles using IV regression to control for selection bias were identified in the literature review. Six of these articles are reviewed in this subsection in order to provide insights into how these researchers applied IV regression. The researchers used a variety of instrumental variables to control for selection bias, some of which were based on geographic location. This observation suggests that some research may support Linden and Adams' (2006) decision to use three-digit zip codes, which represent large geographic areas, as instruments. In addition, some of the researchers indicated that IV regression is more suitable for addressing policy questions about the average treatment effects of health interventions than clinical questions about the possible effects that certain interventions may have on specific patients. Finally, some of the researchers found that IV regression offered a suitable solution to selection bias, while others did not. This may be due to the fact that finding variables that meet the two IV assumptions can be challenging.

The article by McCellan, McNeil, and Newhouse (1994) was probably the first health services study to use IV regression (McClellan & Newhouse, 2000). The authors used differential distances, a geographic location variable, as an instrument to account for selection bias in an observational study examining the effect of cardiac catheterization on mortality among Medicare patients with acute myocardial infarction (AMI). Specifically, they sought to determine if more intensive treatment of AMI in elderly patients reduced mortality. The researchers hypothesized that patient location independently affected choice of hospitals for AMI procedures. The instrument represented whether or not the patients' nearest hospital was a catheterization hospital. It was calculated by subtracting the distance between each patient's zip code and the zip code of the hospital where they received treatment. The outcome variables examined included mortality within one day, mortality within seven days, and mortality within 30 days. They found that treatment at high volume AMI hospitals yielded survival benefits for elderly patients (McClellan et al., 1994).

Landrum and Ayanian (2001), Stukel, Fisher, Wennberg, Alter, Gottlieb, and Vermeulen (2007), and Basu et al. (2007) also employed geographic location variables as instruments in IV regressions. Landrum and Ayanian (2001) used the density of cardiologists in a patient's county of residence as an instrument to estimate the effect of ambulatory specialty care on mortality following myocardial infarction, while Stukel et al. (2007) used regional cardiac catheterization rate as an instrument to estimate the effect of invasive cardiac management on acute myocardial infraction (AMI) survival. Basu et al. (2007) used a regional dummy variable to represent variations in physician practice patterns and a continuous variable that represented Medicare physician fee differentials at the three-digit zip code level to estimate the effect on five-year medical costs of breast-conserving surgery with radiation therapy compared to mastectomy in patients with breast cancer.

Based on their analysis, Landrum and Ayanian (2001) reported that simple and multiple IV regressions indicated that cardiology care was associated with 9.5 and 1.0 percent reductions in mortality rates, respectively. Stukel et al. (2007) reported that their IV analysis showed that cardiac catheterization was associated with a 16 percent reduction in mortality, while Basu et al. (2007) reported that breast-conserving surgery with radiation therapy was associated with an average cost of $41,493. Landrum and Ayanian (2001) and Stukel et al. (2007) both reported that IV regression appeared suitable for answering policy questions about the effects of health system factors on patient health outcomes rather than clinical questions concerning the effect of a particular medical procedure on specific patients. In addition, Basu et al. (2007) indicated that IV regression estimates were suitable for answering “economic” questions about the average or marginal treatment effects of certain interventions.

Long, Coughlin, and King (2005) also used IV regression to address a policy question while controlling for selection bias in a study that estimated the effects of Medicaid on access and use of health care services by low-income mothers relative to private insurance coverage or no insurance coverage. The authors replaced subjects’ actual insurance status with a predicted insurance status using four instruments: accessibility of private insurance, availability of public coverage, and family and community attitudes toward public assistance. They performed the analysis by estimating both OLS and IV models. The Hausman test was used to determine whether there was a significant difference between the models. They also found that IV regression produced better results than OLS regression due to selection bias. Thus, they concluded that the Medicaid program improved access to care for low-income mothers.

Hadley et. al. (2003) had a slightly different experience with OLS and IV regression. They used these procedures to estimate the outcomes of three treatments for early stage breast cancer in elderly women. Medicare fees, woman’s place of residence, and geographic areas were used as instruments. The Hausman test was used to determine if significant differences existed between the OLS and IV models. Hadley et al. (2003) found that the OLS regression results were not significantly different from the IV results. They found that the OLS estimates were preferable to the IV estimates because the latter were substantially larger and more unstable. Hadley et al. (2003) concluded that whether one accepts OLS or IV regression results should depend on three factors: 1) the extent to which observable information can be used as controls in OLS regression, 2) the results of statistical tests of the validity of the IV method, and 3) the similarity of the OLS and IV results.

Posner et al. (2002) used the propensity score method and logistic and IV regression to determine the effectiveness of screening in the identification of early stage breast cancer in elderly women. They also used a geographic location variable, the region in which the women lived, as the instrument. Posner et al. found that all three methods produced similar results, which they argued helped strengthen the credibility of the logistic regression model. They recommended that researchers performing observational studies consider the sources of bias that may affect their results and use the propensity score method to address overt selection bias and IV regression to control for hidden bias.

Finally, Malkin, Broder, and Keeler (2000) used IV regression to determine the effect of postpartum length of stay on new born readmissions. They used hour of delivery and method of delivery as instrumental variables and compared their analysis against standard statistical methods (which was probably OLS regression although they did not specifically state that). Their analysis found that IV regression might provide a better indication of the effects of length of stay on new born readmissions due to selection bias because standard statistical methods appeared to underestimate the effects. They concluded that lengthening postpartum hospital stays reduces readmissions.

The above information suggests that IV regression can offer a viable means of addressing both policy questions about the average effects of health care interventions and hidden selection bias when performing observational studies. The information further suggests that using geographic location variables as instruments can induce variation in an endogenous explanatory variable. However, finding suitable IVs that comply with the required assumptions may be challenging. In addition, the information suggests that OLS can sometimes provide better estimates of the effects of a particular treatment or program than IV regression. This may stem from the fact that finding a suitable instrument can be difficult.

Summary

This chapter provided a review of the literature on disease management, self-management education and self-efficacy theory, diabetes disease management evaluations, and OLS and IV regression. The intent was to provide a foundation for understanding the analytical design and methods that were used in the study. The review found that disease management has become very popular since the early 1990s, in part, due to efforts to both control costs and improve the quality of medical care that chronically ill individuals receive. Moreover, it found that disease management appears to be well suited for state Medicaid agencies because these programs serve a population that is more susceptible to chronic ailments than the general U.S. population.

The literature review further found that a variety of research designs, statistical methods, and outcome variables have been employed to evaluate the effectiveness of diabetes DM programs and services. While the primary justification for the study's analytical design comes from Linden and Adams (2006), the reviewed articles provide further justification for the study's design which was quasi-experimental and employed regression procedures to evaluate the effectiveness of diabetes DM services on three outcome variables. Of particular importance is the fact that no evaluations were identified that used IV regression, which was the focus of this study. IV regression is a widely used econometric procedure, but has received limited attention by the health services research community (Newhouse & McCellan, 1998). This may be due to the fact that IV regression has only recently been used by health services researchers. In fact, the first health services study that employed IV regression was only conducted in 1994 (McCellan & Newhouse, 2000). Finally, the literature review considered some of the technicalities involved with performing OLS and IV regression and how some health services researchers have employed these procedures. Of particular importance here is the fact that IV regression does not always provide a suitable solution to selection bias issues that are encountered in observational studies.

Definitions of Important Terms

Definitions for important terms used in the present study are provided below.

Administrative Claims Data: Medicaid claims data created for payment purposes rather than for research purposes. Claims data represent electronic versions of bills submitted by Medicaid providers (i.e., physicians, hospitals, pharmacies, etc.) for recipient office visits, hospital stays, pharmacy purchases, laboratory tests, or other encounters. Administrative claims data contain information on items such as: 1) the date and location of services, 2) type and cost of services, 3) procedures performed, 4) extent of services (i.e., hospital stays), and 5) recipient demographics (Piecoro, Wang, Dixon, & Crovo, 1999; Wyant & Parente, n.d.).

Diabetes: Diabetes is a chronic disease that occurs when an individual is unable to produce or use insulin, which is a hormone needed to convert sugar, starches, and other food products into energy (HMC, 2007).

Disease Management (DM): DM is a system of coordinated health care interventions and communications for populations with chronic disease (i.e., illnesses or conditions lasting more than three months in duration) that can be mostly controlled through patient self-care activities (Linden & Adams, 2006).

Disturbance (or Error) Term: The variable in an OLS regression model that contains unobserved factors affecting the outcome variable. The error term may also include measurement errors in the outcome or predictor variables (Wooldridge, 2006).

Endogenous Explanatory Variable: In OLS regression, an independent variable that is correlated with the disturbance term due to measurement error, omitted variables, or simultaneity (Wooldridge, 2006).

Estimate: The numerical value of an estimator derived from data on subjects in a specific sample (Stock & Watson, 2007).

Estimator: A procedure for using sample data to estimate the value of a population parameter. Ideally, researchers prefer estimators that get as close as possible to the unknown true value of the population parameter (Stock & Watson, 2007).

Exogenous Variable: Any variable in an OLS regression model that is uncorrelated with the disturbance term (Wooldridge, 2006).

High Intensity On Demand: Virginia Medicaid recipients who were contacted by HMC and declined to received the high intensity intervention, lost contact with HMC, or were never contacted by HMC. These individuals receive educational mailings (HMC, 2007).

High Intensity Open: Virginia Medicaid recipients who agree to receive the high intensity intervention. These individuals receive regular follow-up calls, individualized care plans, 24-hour access to program nurses, and quarterly condition-specific information (HMC, 2007).

Instrumental Variables (IV) Regression: An IV regression model is a linear equation that is used when one or more instrumental variables are available for the endogenous predictor (Wooldridge, 2006). IV regression is usually performed using two-stage least squares regression.

Instrumental Variable (IV): In a regression model containing an endogenous independent (or predictor) variable, the IV is a variable (or set of variables) that is not contained in the model, is uncorrelated with the model's disturbance term, and is partially correlated with the endogenous predictor (Wooldridge, 2006).

Observational Study: A study in which variables are observed instead of manipulated and subjects are not randomly assigned to treatment conditions (i.e., a quasi-experimental study) (Shadish et al., 2002).

Ordinary Least Squares (OLS) Regression: OLS regression is a statistical methodology that relates a quantitative outcome (or dependent) variable to one or more quantitative and/or qualitative independent (or predictor) variables. OLS estimates parameter coefficients for the predictors by minimizing the sum of squared residuals. The end result is a mathematical model that predicts the outcome variable for the given set of predictors (Mendenhall & Sinich, 2003; Wooldridge, 2006).

Propensity Score: The propensity score is the probability that an individual will be assigned to the treatment group instead of the control group based on a set of observed variables Z_i. The propensity score is formally defined as $P(Z_i) = \text{Prob}(T = 1|Zi)$ (Winship & Morgan, 1999).

Selection Bias: Nonrandom assignment to treatment conditions that result in subject characteristics between conditions that may be related to differences in outcome (Shadish et al., 2002).

Standard Intensity: Virginia Medicaid members who are considered at standard risk for future health care expenses and are able to manage their conditions with limited external support. They receive mail-in assessments, educational materials, quarterly disease-specific information, and 24-hour access to program nurses (HMC, 2007).

Two-Stage Least Squares Regression: An IV regression procedure, where the IV for an endogenous explanatory variable is obtained by regressing the endogenous variable on all exogenous variables (Wooldridge, 2006).

Chapter 3

Methodology

The purpose of this study was to compare and contrast ordinary least squares (OLS) regression and instrumental variables (IV) regression using a three-digit zip code instrument procedure developed by Linden and Adams (2006). OLS and IV regression are two statistical methods that can be used to estimate treatment effects in observational (or quasi-experimental) studies. OLS regression is used in observational research to assess the relationship between a treatment variable and an outcome, while adjusting for important explanatory variables to ensure comparability between the treatment and control groups (Newgard, Hedges, Arthur, & Mullins, 2004). However, OLS regression may fail to produce consistent effect estimates if certain important variables that are correlated with the treatment are excluded from the analysis (Foster & McLanahan, 1996; Guo, Barth, & Gibbons, 2006). This often occurs in observational studies because individuals who participate in treatments (or programs) usually differ systematically from those who do not.

IV regression, which approximates a pseudo-randomization, can ameliorate this by inducing variation in the treatment variable, but not in the outcome (Newhouse & McClellan, 1998; Scheider et al., 2007). To implement IV regression, researchers have to find one or more variables (called instruments) that produce this variation. Unfortunately, finding suitable variables that meet this criterion can be very challenging (McClellan & Newhouse, 2000). This study specifically examined the feasibility of using patient three-digit zip codes as instruments in an IV regression analysis to evaluate the effects of participation in the Virginia *Healthy Returns*^{SM} DM program.[13] Linden and Adams (2006) argue that patient three-digit zip codes can function as instruments by inducing variation in the treatment variable (i.e., the DM program participation variable). Thus, they reason that zip code instruments can offer a means of providing an unbiased estimate of DM causal effects on certain outcome variables.

This chapter reviews the methodology that was employed during the present study. Additional information on the study design, database, population, variables, and statistical procedures is provided in the sections below.

Study Design and Research Questions

Researchers seek to make causal inferences when evaluating social and medical interventions. Causality is strongest when researchers conduct controlled randomized experiments where subjects are randomly assigned to the study groups. Randomization balances the groups in terms of all relevant factors other than treatment exposure. It is because of randomization that causation can be inferred from experimental studies. However, experimental studies are usually difficult to implement for a variety of financial, ethical, and practical reasons. When experimental studies are not feasible, social scientists normally turn to observational research designs (Freedman, 2005).

Because Medicaid recipients self-select into the high intensity open treatment option of the Virginia *Healthy Returns*^{SM} Program, this study was conducted as an observational study (Freedman, 2006). A key characteristic of an observational study is

[13] The study did not attempt to determine the feasibility of IV regression in general. Instead, it sought to determine the appropriateness of a particular instrumental variables procedure that was proposed for evaluating DM programs.

that subjects are assigned nonrandomly to the treatment and control groups. The investigators simply observe what occurs during the study because they are unable to manipulate the treatment. For example, medical studies that examine issues, such as the effects of smoking, are observational because researchers cannot ethically randomize subjects to the treatment and control groups. A second characteristic of observational studies is that researchers try to estimate the effects of an intervention by comparing the treatment group to the control group on an outcome variable (Freedman, 2005). A third characteristic is that an observational study usually involves the analysis of data from a large administrative database to establish the consequences of social or medical interventions (Newhouse & McClellan, 1998).

Based on the above information, the present study employed an observational design because: 1) the high intensity open subjects self-selected into the treatment group, 2) the effects of the DM program were estimated by comparing the treatment and control groups on several outcome variables, and 3) the analysis data came from a large pre-existing administrative database. Because the study by Linden and Adams (2006) used a one year experience of a diabetes DM program, the study period for this analysis was from January 1, 2007 to December 31, 2007.[14] The objective of the study was to determine if an IV zip code procedure similar to the one developed by Linden and Adams using data from an Oregon managed care diabetes DM population could generalize to a Virginia Medicaid diabetes DM population. Three research questions were addressed in the study:

[14] The *Healthy Returns*^{SM} Disease Management program officially began in January 2006. Calendar Year (CY) 2007 was selected because it represents the program's most recent year of operation.

1. Which statistical method provides the best unbiased estimates of high intensity open DM program participation on the outcome variables?
2. Do the parameter estimates and confidence intervals for the outcome variables differ depending upon which statistical method is used?
3. What are the advantages and disadvantages of using OLS and IV regression to evaluate high intensity open DM program effectiveness?

Database

The study data came from the Virginia Medicaid Management Information System (VaMMIS), which is a large administrative database that the Virginia Department of Medical Assistance Services (DMAS) uses to process medical claims submitted by health care providers for providing services to Virginia Medicaid recipients. In January 2008, the contractor that administers the DM program for DMAS provided the researcher with a list of Medicaid recipient identification numbers for 2,741 diabetes recipients who were enrolled as either high intensity open, high intensity on demand, or standard intensity patients in the DM program during calendar year (CY) 2007.

The high intensity open participants served as the treatment group and the standard intensity participants served as the control group. The rationale for using the high intensity open recipients as the treatment group was that these individuals actively participated in the DM program, while the standard intensity recipients did not. According to Linden et al. (2005), this comparison is appropriate when evaluating population-based DM programs because most (i.e., 95 percent or more) of the eligible recipients will be enrolled, which limits the number of subjects available for comparison purposes.

All recipient-level claims and demographic data are stored within VaMMIS as SAS files. The researcher worked with staff in the DMAS information management division to develop SAS programs that used the recipient identification numbers to collect claims and demographic data on each subject. As part of this process, 339 subjects who were not continuously enrolled in both the Virginia Medicaid program during CY 2006 and 2007 and the *Healthy Returns*^{SM} DM program during CY 2007 were deleted from the analysis dataset. Recipients who participated in a DM pilot program that operated prior to January 2006 were also deleted. In addition, 775 high intensity on demand recipients were deleted from the dataset. The purpose of these deletions was to ensure that the treatment and control groups contained subjects who had equally experienced the study outcomes (i.e., hospitalizations, ED visits, and diabetes-related costs) (Linden et al., 2005) and to increase the chances of detecting treatment effects by reducing potential extraneous variability (i.e., background noise) due to the high intensity on demand recipients (Mendenhall & Sincich, 2003).[15] The data were then collapsed to form the study variables. Once the variables were developed, they were compiled into a dataset for use in the study's analysis stage. The computer programming needed to assemble the dataset occurred during February 2008.

Population and Sample

For this study, the population of interest was Virginia Medicaid diabetes recipients who were enrolled continuously in either the high intensity open or standard intensity

[15] High intensity on demand recipients can receive high intensity open services, which are not reflected in the claims data.

options of the *Healthy Returns*^{SM} program during CY 2007 (*N* = 1,627) and had not participated in the DM pilot program. (Continuous enrollment was defined as DM participation beginning on January 1, 2007 and ending on December 31, 2007.) The unit of analysis was the individual diabetes DM participant. The treatment group consisted of the 229 high intensity open participants, while the control group consisted of the 1,398 standard intensity participants.[16]

Study Variables

This section consists of two subsections. The first subsection provides information on the outcome variables, while the second presents information on the independent variables. Detailed information on both the zip code instruments and propensity score variables employed by Linden and Adams (2006) is provided in the second section.

Outcome Variables

Three outcome variables were examined in the present study: CY 2007 annual diabetes costs, hospital emergency department (ED) visits, and hospital days. The operational definitions of these variables are as follows:

- annual diabetes costs are costs related to any medical claim submitted to Virginia Medicaid during CY 2007 for one of the diabetes patients where

[16] According to Stevens (2002), studies involving treatment and control groups that consist of about 100 subjects each will achieve a power of at least 0.94. Thus, the probability of a type II error (or failing to find significant relationships that actually exist) would be 0.06 or less in these studies. Based on this information, power was not an issue in this study because the treatment and control groups consisted of 229 and 1,398 subjects respectively. These large samples made it highly probable that the statistical analyses identified significant relationships that actually exist in the study population.

diabetes was indicated as a diagnosis using international classification of diseases version 9 clinical modification (ICD-9-CM) codes 250.0 to 250.9,[17]

- hospital ED visits are the total number of ED visits made by recipients during CY 2007 for either emergency or non-emergency related conditions, and
- hospital days are the total number of days that recipients stayed in hospitals during CY 2007.

It should be noted that both the hospital ED and hospital days outcome variables may include data on both diabetes related and non-diabetes related conditions. It should also be noted that the hospital days variable may overstate the number of days that some recipients stayed in the hospital due to how the variable was calculated.[18] In addition, all outcome variables include a substantial number of zeros because many of the DM patients did not visit EDs, stay in hospitals, or incur diabetes-related medical costs during the study period. As a result, the zeros in these variables were not considered missing values because they represented legitimate observations.

The outcome variables were selected for three reasons: 1) Linden and Adams (2006) used them in their IV analysis, 2) several articles reviewed for this study used similar outcomes (Berg & Wadhwa, 2007, Christakis, et al., 2004, and Villagra & Ahmed, 2004), and 3) the Virginia *Healthy Returns*^{SM} DM Program is intended to reduce these

[17] According to the National Center for Health Statistics (2007) website, the ICD-9-CM is used to assign codes to diagnoses related to inpatient, outpatient, and physician office visits in the U.S. Searching for claims where diabetes is indicated as a diagnosis should eliminate claims that are not associated with diabetes, such as claims submitted for broken bones or sore throats.

[18] The variable was calculated using the "through" and "from" dates on hospital and physician inpatient/outpatient medical claims. Because there are no admittance or release dates on medical claims, this variable should probably be viewed as a proxy for the number of days that recipients were hospitalized during a calendar year.

outcomes for program participants. While justification exists for using these variables, several caveats are associated with their use. In particular, Linden (2006) reports that costs can be problematic due to fluctuations in provider reimbursements, insurance coverage, and technological innovations, which are beyond the control of most DM programs. Observed cost changes may thus not be entirely attributable to DM interventions. Linden (2006) further reports that ED visits and hospitalizations may be low in the target population prior to program implementation. If this occurs, then statistical analyses may not detect any significant program effects on these two outcomes. These caveats should be considered when interpreting the study's analytical results.

Independent Variables

Five independent (or predictor) variables were used in the study: actual program participation, predicted program participation, propensity scores, age, and gender. The program participation variable was a dichotomous variable coded as 1 = high intensity open participant and 0 = standard intensity participant. Following guidance from Linden and Adams (2006), predicted program participation represented the probability of a given individual enrolling in the Virginia *Healthy Returns*^{SM} DM Program. This variable was calculated using recipient zip codes. The actual program participation variable was used in the OLS regression models for comparison purposes. This variable was presumed to be influenced by selection bias due to preexisting differences between the treatment and control subjects (Wooldridge, 2006).[19]

[19] In structural equation modeling terms, predictor variables that are correlated with the error term are known as endogenous variables, while variables that are independent of the error term are referred to as exogenous (Freedman, 2006).

To perform IV regression, a researcher should select one or more instruments that are related to the endogenous predictor, but are not related to the outcome variable (Foster & McLanahan, 1996). For DM evaluations, this means that a variable must be identified that is predictive of a recipient's DM enrollment status, but is not associated with any of the unobserved covariates that influence the outcome variables. Linden and Adams (2006) argue that recipient three-digit zip codes make good instruments for two reasons: 1) a recipient who lives in a DM covered service area would be eligible for enrollment, but not necessarily enroll, and 2) living in a particular zip code may be independent of specific unobserved confounding variables.

Zip codes (i.e., zone improvement plans) are five-digit numbers that identify specific geographic mail delivery systems. Five-digit zip codes are assigned to every address in the United States. The first digit designates a general area of the country. For example "2" designates the District of Columbia, Maryland, North Carolina, South Carolina, Virginia, and West Virginia. The second and third digits refer to a US Postal Service sectional center facility (SCF). Each SCF functions as a distribution and processing center for approximately 40 to 150 surrounding post offices that are located within the facility's designated geographic area. The fourth and fifth digits designate one of the surrounding post offices. Zip codes are composed of clusters of addresses identified for mail delivery purposes. Urban and suburban zip codes are usually comprised of approximately located streets, while rural zip codes are comprised of selected roads or delivery routes (www.carrierroutes.com/ZIPCodes.html).

For this study, the instrumental variables were generated using the procedure developed by Linden and Adams (2006). The process began by identifying the number of unique five-digit zip codes for the study subjects. The zip codes were then collapsed based on the first three digits in order to produce new zip codes that represent larger contiguous geographical areas (e.g., each three-digit zip code represents a larger geographic area than each five-digit zip code). These zip codes were further examined to identify categories with less than nine frequency counts, which were collapsed into a category called "other" (Linden & Adams, 2006).

Linden and Adams (2006) argue that this zip code model mimics a "natural experiment" by assuming two types of zip codes: Zip Code-1, which represents high program participation, and Zip Code-2, which represents low program participation.[20] Linden and Adams argue that this "experiment" models how the differences in outcomes between these two zip codes are related to the differences in DM program participation rates. Because three-digit zip codes represent larger geographic areas than five-digit zip codes, they also argue that these larger areas may make participants and non-participants more similar on unmeasured confounders, which they indicate is true for measured demographic variables. They further argue that within a given three-digit zip code, the correlation between program participation and unmeasured confounders may be smaller than the correlation between zip codes. Thus, they reason that zip codes should have an

[20] Natural experiments refer to exogenous events that change subjects' environments (Wooldridge, 2006).

indirect effect on the outcome variables that is interceded only by DM program participation.[21,22]

Once the three-digit zip code categories were developed, logistic regression was used to estimate the probability of a given individual enrolling in the high intensity DM program using the participants' actual program participation status as the dependent variable (1 = high intensity open participant and 0 = standard intensity participant) and *k*-1 dummy variables to represent the three-digit zip code categories. (Categorical variables are always modeled as dummy variables with one less than the actual number of levels to prevent situations of perfect multicollinearity from occurring.) Additional covariates included in the logistic regression model were age, gender, and propensity scores, which were considered to be exogenous variables in this study. The estimated probability of program participation was used as the endogenous variable in the IV regression models and the zip code dummy variables were used as instruments (Linden & Adams, 2006; Linden, personal communication, June 16, 2008).

The feasibility of the logistic regression model was assessed using the likelihood ratio chi-square, *c*-statistic, Somer's D, and Gamma statistics. Descriptive statistics were also used to compare the characteristics of the treatment and control groups for the actual and predicted program participation variables to determine if significant differences existed

[21] Linden and Adams (2006) did not offer any empirical evidence to support their zip code model nor did they elaborate on how the differences in outcomes are related to differences in program participation.
[22] The US Census Bureau collects socio-economic data at the five- and three-digit zip code levels. While these data could potentially be analyzed to determine whether significant observable differences exist between the individuals who reside in the three-digit zip codes used in the present study, such an analysis was not performed because it was outside the study's scope. However, these data could be used in a future study examining the feasibility of three-digit zip code instruments.

between the groups (Peng, Lee, & Ingersoll, 2002; Mendenhall & Sincich, 2003; Linden & Adams, 2006). In addition, simple regressions were performed to determine if the predicted probability of program participation and the actual program participation variables were related to the zip code instruments (Wooldridge, 2006). As part of this process, a sensitivity analysis was performed. The analysis consisted of regressing the upper and lower 95% confidence limits for the predicted probability of program participation on the zip code instruments to determine if the relationship between the probability of program participation and zip codes was sensitive to the specific form of this variable.

Linden and Adams (2006) include a propensity score as an explanatory variable in their OLS and IV regression models.[23] The propensity score is the conditional probability of assignment to a particular treatment given a vector of observed variables (Rosenbaum & Rubin, 1983). The propensity score (PS), which is calculated using logistic regression, is defined as: $\text{PS} \equiv \Pr\{t = 1|x\}$, where t denotes whether a subject received the treatment and $\Pr\{t = 1|x\}$ is the probability of receiving the treatment given the observed covariates, x. Propensity scores allow researchers to estimate program causal effects by directly comparing treatment and control subjects on various outcomes. For this reason, researchers can relax the typical assumptions of regression (linearity, multicollinarity, parsimony, etc.) when calculating propensity scores (Love, 2004; Frank, Sykes, Anagnostopoulos, Cannata, Chard, Krause, & McCrory, 2008).

[23] Linden and Adams (2006) referred to the propensity score as a risk score in their article. Linden (personal communication, October 19, 2007) indicated that the risk score was actually a propensity score.

By collapsing the observed covariates into one score, the propensity score represents the probability that a particular subject would have received the treatment over another subject based on a larger collection of covariates. As such, the propensity score functions as a balancing score. In observational studies where subjects are nonrandomly assigned to the treatment and control groups, systematic differences are assumed to exist that can bias treatment effect estimates. Researchers use propensity scores to reduce overt bias by integrating this mechanism into the analysis in order to improve study group comparability. This process can be viewed as a means of reducing bias to obtain better program effect estimates through a quasi-randomization of the study groups (Newgard et al., 2004).

Propensity scores are generally employed using one or more of the following methods: 1) regression – where the propensity scores are added to the model as a covariate, 2) stratification – where the sample is divided into quintiles based on propensity scores to allow for the assessment of treatment effects within stratums, and 3) matching – where treatment and control subjects are matched based on similar propensity scores to allow for less biased estimates of program treatment effects (D'Agostino, 1998; Newgard et al., 2004). Weighting, which is rarely used in medical research, is another method of employing propensity scores (Austin, 2008; Frank et al., 2008). In this scenario, weights such as 1/PS and 1/(1 – PS) for the treatment and control subjects are created as a means for better estimating average program causal effects (Gelman & Hill, 2007). (The use of propensity score weighting is further explored in Appendix A.)

In their article, Linden and Adams (2006) did not specify which propensity score method they used; however, Linden (personal communication, October 19, 2007) reported that matching was employed. While matching is the preferred method of using propensity scores, it eliminates subjects from the analysis who cannot be matched, which can result in the loss of considerable data (Frank et al., 2008). It also requires the development of sophisticated matching algorithms and computer programs that are capable of matching pairs (or groups) of subjects and then either removing them from the pool of all available subjects or keeping them in for additional matching (Shadish et al., 2002; Morgan & Winship, 2007). For these reasons, the propensity score was simply used as a regression covariate in the study. The benefit of this approach is that the propensity score serves as an explanatory variable that controls for observable background characteristics in the regression analysis, while maintaining the integrity of the study sample (Wooldridge, 2002; Frank et al., 2008).

Linden and Adams (2006) estimated their propensity score regression model by using actual program participation status as the dependent variable and 150 clinical, financial, demographic, and utilization variables. Such an endeavor was not feasible for this study because the data came from a large preexisting database that contains thousands of raw data elements. Converting these elements into meaningful variables would require substantial computer programming efforts. Because variable selection for propensity score calculations is ad hoc (Brookhart, 2006; Sturmer, Joshi, Glynn, Avorn, Rothman, & Schneeweiss, 2006), the propensity scores were calculated using several variables selected because of their availability. These variables included CY 2006 diabetes-related costs, CY

2006 health care utilizations, age, gender, race, region of residence, country of origin, and citizenship status. The appropriateness of the propensity scores was assessed by examining the overlap in the distribution of scores between the treatment and control groups. A series of simple regressions were also performed to determine if the scores controlled for significant differences between the study groups on the covariates.

Finally, CY 2007 recipient age and gender were included as covariates in both the OLS and IV regression models. These variables were operationalized using the same definitions employed by Linden and Adams (2006) where age was measured in years and gender was coded as 1 = female and 0 = male.

Data Screening

Data screening consisted of calculating descriptive statistics for both the outcome and independent variables (Cohen, Cohen, West, & Aiken, 2003). Means, standard deviations, ranges, histograms, and box plots for the quantitative variables and frequency distributions for the qualitative variables were calculated as part of the initial data screening process. As previously mentioned, zeros in the outcome variables were not considered missing since some recipients did not incur costs or utilize inpatient services during CY 2007. Data screening was performed using SPSS version 15.0.

During data screening, particular attention was focused on identifying univariate and multivariate outliers because they can distort regression coefficients, standard errors, and model R^2 statistics (Tabachnick & Fidell, 2001; Cohen et al., 2003).[24] Outliers usually

[24] In regression, outliers are observations that deviate from other cases on the dependent variable, while leverage points are observations that deviate from others on the independent variables. Observations that deviate on both the dependent and independent variables are influential data points (Mays, personal

represent either contaminated data (i.e., measurement or coding errors) or accurate observations of rare cases (Cohen et al., 2003). For this study, univariate outliers were identified using box plots and standardized scores greater than 3.29 (Tabachnick & Fidell, 2001).[25] The data were screened for multivariate outliers by calculating Mahalanobis distance statistics for all study subjects in the dataset. Mahalanobis distance represents the distance that a subject is from the centroid of all subjects, where the centroid represents the intersection of the variables' averages. Mahalanobis distances were calculated by regressing each outcome variable on the predictors and then saving the distance statistics generated as part of the regressions. Mahalanobis distances were evaluated using the chi-square distribution ($p < 0.001$) with degrees of freedom equal to the number of predictors in the models. Cook's distances were also calculated as part of this process. Cook's distance is a leverage measure that assesses change in regression coefficients when observations that differ from others in the independent variables are deleted. Observations with distance scores greater than 1.0 were identified as outliers (Tabachnick & Fidell, 2001).

Based on these analyses, a number of outlier observations were identified. To determine if outliers should be deleted or retained, a series of regressions were performed with the outliers included and excluded and the variables containing them transformed using either the natural logarithm, reciprocal, or square root transformations to reduce their

communications, January 7, 2008). However, for simplicity the term "outlier" (i.e., outliers, leverage points, and influential data points) was defined in this study as a case with an extreme value on one variable (a univariate outlier) or a combination of extreme values on two or more variables (a multivariate outlier).

[25] Tabachnick and Fidell (2001) report in their textbook on multivariate statistics that among continuous variables, cases with very large standardized scores in excess of 3.29 are potential outliers.

impact on the regression models. Transformations can reduce the influence of outliers by changing the variables' distributions to be approximately normal (Tabachnick & Fidell, 2001). Diagnostic statistics were performed on the transformed variables and the variables with deleted values to determine if these options reduced the outliers' influence.

As part of this process, regression analyses were performed to allow the researcher to determine if the outlier adjustments produced any notable changes in the overall regression models. Specifically, the researcher sought to determine if the significance of the model *F* statistics or predictors changed or if the R^2 statistics fluctuated substantially. If no noticeable changes are observed, then little is probably gained by transforming the variables or deleting observations from the dataset. However, if changes are observed, then a decision can be made as to whether the transformed variables or the original variables should be used in the analysis.

Normality, linearity, and homoscedasticity were assessed to ensure that the data met these required OLS regression assumptions. Normality refers to the extent to which the variables have distributions that are approximately normal. Linearity is the degree to which a straight line relationship exists between two variables, and homoscedasticity is the extent to which the residual distribution has equal variances for all predicted values of the outcome variable (Cohen et al., 2003). Assessment of these assumptions was performed by calculating skewness and kurtosis statistics (they should be near zero) for each variable, and by regressing each outcome on the predictors to generate residual plots, where standardized residuals are plotted against the outcomes' predicted values.

Evidence for violations of the OLS regression assumptions exists if the skewness and kurtosis statistics are large, if noticeable patterns exist in the residual plots, or if the residuals are predominately located above or below the residual plots' mean zero lines (Stevens, 2002; Tabachnick & Fidell, 2001). The key to working with transformed variables is to identify the transformation that produces skewness and kurtosis values near zero, the cleanest residual scatter plot, and/or the least number of outliers. As a result, several transformations were applied before the most appropriate transformations were identified for use in the study (Tabachnick & Fidell, 2001).

Finally, the data were screened for bivariate and multivariate multicollinearity, which are serious conditions that exist when independent variables are highly correlated (i.e., contain redundant information). Multicollinearity can weaken statistical analyses by inflating error terms, which can cause parameter estimates to be non-significant (Tabachnick & Fidell, 2001). For this study, bivariate multicollinearity was defined as intercorrelations of at least 0.80 in a correlation matrix and multivariate multicollinearity was defined as variance inflation factors (VIF) that exceed 10 (Stevens, 2002). VIFs are calculated by regressing the outcomes on the predictor variables. Large VIFs indicate that there is a strong linear association between a particular predictor and the remaining predictors, suggesting the presence of multivariate multicollinearity. Screening for multivariate multicollinearity is important because it is possible for a predictor to only have weak or moderate bivariate correlations with other predictors, but then have a high multiple correlation when regressed with the other variables. If detected, multicollinearity can be addressed through variable deletion (Stevens, 2002).

As part of the multicollinearity screening process, particular attention was focused on determining if the predicted probability of program participation and the propensity score variables produced multicollinearity in the IV regression models since both variables represent the same concept (i.e., the probability of program participation). In addition to performing the correlation and VIF analyses, a quintile propensity score variable (levels 1 – 5) was calculated and included in the IV regressions in lieu of the quantitative propensity score variable (0.0 – 1.0) to determine if the models were sensitive to this variable's specific form.

Research Questions and Data Analysis

This section discusses the data analysis procedures that were used to answer the present study's three research questions. Additional information is provided in the subsections that follow.

Analysis of Research Questions One and Two

The first two research questions developed for the study were:

1. Which statistical method provides the best unbiased estimates of high intensity DM program participation on the outcome variables? and
2. Do the parameter estimates and confidence intervals for the predictor variables differ depending on which statistical method is used?

These questions were addressed by using STATA version 10.1 to estimate a series of OLS and IV regression models for each outcome variable. Statistically significant results were determined using an alpha level of 0.05.

Research question one was addressed by estimating several OLS and IV regression models. The OLS models were estimated by regressing the outcomes on the actual DM program participation, propensity score, age, and gender variables. Three blocks of OLS and IV regression models were developed for the study (i.e., a total of 18 regression models were developed). In the first block, the IV models were estimated following the method used by Linden and Adams (2006) where each outcome was regressed on the predicted probability of program participation (the endogenous variable), zip code instruments, age, gender, and propensity score variables. In the second block, the predicted probability of program participation was replaced with the actual program participation variable to determine if the IV regression models were sensitive to the specific form of the endogenous variable used in the regressions. In the third block, the propensity score variable was replaced by a quintile variable to determine if the regressions were sensitive to this variable's specific form.

The IV regressions were performed using two stage least squares regression (2SLS), which estimates two least squares equations simultaneously. In the first equation, both the zip code instruments and exogenous variables (which were age, gender, and propensity scores) were used to estimate X-hat (the predicted value of program participation). This variable was then inserted into the second equation instead of the actual endogenous variable and the outcomes were regressed on the predicted value of program participation and the exogenous variables. 2SLS must be used for the IV regression procedure because running OLS regression twice to perform the IV procedure results in incorrect estimations of the residual sum of squares across all observations and their standard errors (Linden & Adams, 2006; Gelman & Hill, 2007).

The feasibility of the IV regression models was assessed using a variety of statistical tests. For instance, a series of simple regressions were performed to determine if the instruments were related to both the program participation and predicted program participation variables. The residuals from each IV regression model were also regressed on the outcome variables to test the assumption that the IV estimates were not related to the outcome variables. The Sargon chi-square test was used to test the assumption that at least some of the zip code instruments were uncorrelated with the disturbance terms in the IV models. The Hausman specification test was used to determine if significant differences existed between the OLS and IV regression coefficients. Finally, the relevancy of the instruments was assessed by examining the first-stage *F*-statistics from the IV regressions (Hadley et al., 2003; Baum, 2006; Linden & Adams, 2006; Wooldridge, 2006; Stock & Watson, 2007).[26]

Research question two was addressed by comparing the OLS and IV coefficients and confidence intervals for the variables in the three regression blocks. Particular attention was directed toward discussing why IV regression produces large coefficient variances and how these variances can lead to model instability. As part of the discussion, the calculations used to produce the OLS and IV estimates were briefly reviewed.

[26] Linden and Adams (2006) appeared to use the Durbin-Wu-Hausman *F*-test to determine if significant differences existed between their OLS and IV regression models. The Hausman specification test can be performed either as a chi-square or *F*-test. The Durbin-Wu-Hausman *F*-test was not used in this study because the "hausman" subcommand in the "ivregress 2sls" command in STATA version 10.1 reports a chi-square test (Ender, n.d.; Hadley et al., 2003; Greene, 2003).

Information on the relevancy of the zip code instruments (or the extent to which the instruments explain variation in the program participation variable) and the limitations of using instruments from natural experiments to estimate average treatment effects were also included.

Analysis of Research Question Three

The final research question developed for the present study was:

3. What are the advantages and disadvantages of using each statistical method to evaluate high intensity DM program effectiveness?

This question was addressed through the researcher's assessment of the advantages and disadvantages of using both methods for DM program evaluations. Particular emphasis was directed toward discussing the feasibility and limitations of using each method.

Institutional Review Board

Even though this study involved the analysis of preexisting administrative data, the researcher still had to link subjects to their confidential health records in order to create the study variables. The study thus involved human subject research, which required it to fall under the purview of Virginia Commonwealth University's Institutional Review Board (IRB). However, because the researcher protected the identifies of all subjects by removing their Medicaid recipient identification numbers from the analysis dataset, the university's IRB determined that the study qualified for an exemption under 45 Code of Federal Regulations (CFR) 46.101(b)(4), which states that research activities are exempt as long as subject identities are protected (U.S. Department of Health and Human Services, 2007; Ann Nichols-Casebolt, personal communication, April 9, 2008).

Summary

This chapter reviewed the methodology that was employed in the present study. Information was presented on the study design, database, population, variables, and statistical procedures. The study was conducted as an observational study because data from a large preexisting database was used to determine whether an instrumental variables regression procedure proposed by Linden and Adams (2006) could generalize to a Virginia Medicaid population. The planned analysis attempted to follow their procedure, albeit with some modifications.

Chapter 4

Results

The results of the study are presented in this chapter, which begins with a discussion of the propensity score and instrumental variables calculation procedures. The data screening procedures employed to assess the appropriateness of the variables for the ordinary least squares (OLS) and instrumental variables (IV) regression models are next reviewed followed by a discussion of the analytical results from the OLS and IV regression models. The chapter concludes with a summary of important study findings.

Propensity Score Variable Calculation

According to some observers, selecting covariates from which to estimate propensity scores is a critical step that should be based on a-priori theoretical grounds and previous research (Yanovitzky, Zanutto, & Hornik, 2005; Guo et al., 2005). However, Linden et al. (2005) maintain that disease management (DM) evaluators should use any variables that are available because they often have access to limited amounts of data. For this reason, the propensity score variable was calculated using the variables presented in Table 1.[27,28] As can be seen from this information, most of the study subjects were female (1,152), white (1,377), and resided in the western part of the State of Virginia (1,336). This information also shows that considerable variability exists in the 2006 age (*M* = 45.16, *SD* = 15.51), hospital days (*M* = 1.07, *SD* = 11.39), emergency department (ED)

[27] An argument could be made that the variables used in the propensity score model were selected based on prior research because Linden et al. (2005) used age, sex, geographic location, number of hospital admissions, number of emergency department visits, and total costs to calculate propensity scores for a group of congestive heart failure DM participants.

Table 1
Descriptive Statistics for Observed Covariates Used in the Propensity Score Regression Model (N = 1,627)

Variable	Frequency (%)	Mean (SD)
Gender/Male	475 (29.2%)	
Gender/Female	1,152 (70.8%)	
Race/White	1,377 (84.6%)	
Race/Black	208 (12.8%)	
Race/Other	42 (2.6%)	
Region/Coastal	70 (4.3%)	
Region/Northern	72 (4.4%)	
Region/Central	149 (9.2%)	
Region/Western	1,336 (82.1%)	
Country of Origin/US	1,568 (96.4%)	
Country of Origin/Other	59 (3.6%)	
Primary Language/English	1,612 (99.1%)	
Primary Language/Other	15 (0.9%)	
US Citizen/Yes	1,581 (97.2%)	
US Citizen/No	46 (2.8%)	
Age (2006)		45.16 (15.51)
Hospitalizations (2006)		1.07 (11.39)
ED Visits (2006)		1.62 (3.28)

(continued)

[28] At a minimum, Linden, Adams, and Roberts (n.d.) argue that participant age and sex should be included in the propensity score calculation.

Diabetes Related Costs (2006)	$1,963.09 ($7,084.46)

visits (*M* = 1.62, *SD* = 3.28), and diabetes-related cost (*M* = $1,963.09, *SD* = $7,084.46) variables. The standard deviations for the hospital days, ED visits, and cost variables are much larger than the means, which indicates that the variables are positively skewed due to the presence of a large number of zeros. This finding is not too surprising because large numbers of zeros are frequently found in many variables that are examined in health services research (Mullahy, 1998).

Ideally, actual pre-program (i.e., CY 2005) cost, utilization, and clinical variables should be used to calculate propensity scores (Linden et al., 2005)[29], but the researcher did not have access to this information. In addition, while both quadratic and interaction terms can be included in propensity score regression models, only three interaction terms for the age, utilization, and cost variables were included in the propensity score model developed for the present study. These terms were added in order to maximize model fit and because previous research suggests that health care costs and utilizations may be influenced by patient age (Lynn & Adamson, 2003; Jones & Richmond, 2006). Interaction terms were used sparingly in the regression model because the inappropriate use of these terms may alter the estimated propensity scores by inflating coefficient variances (Baser, 2006).

[29] Debate exists among researchers as to which variables to include in propensity score models. Some argue that only variables that predict treatment assignment should be included, while others argue that all variables that are potentially related to the outcome should be included. Still others argue that propensity scores should only be calculated using variables that are related to both the treatment and outcome variables (Austin, Grootendorst, & Anderson, et al., 2007).

The propensity scores were estimated using the logistic regression option in SPSS version 15.0. The program participation variable (coded as 1 = high intensity open participant and 0 = standard intensity participant) was used as the dependent variable in the regression model. The results of the regression for the propensity to be a high intensity open participant are presented in Table 2. Only the 2006 ED visits, cost, and age x ED visits variables were significantly related to the participation variable; however, all

Table 2
Logistic Regression for Propensity of Participating in the High Intensity Open DM Intervention (N = 1,627)

Independent Variable	Estimate	Std. Error	Wald $\overline{\chi}^2$	p-value
Constant	-3.492	1.168	8.934	0.003
Female	0.133	0.179	0.548	0.459
Race/White	-0.254	0.741	0.117	0.732
Race/Black	0.198	0.743	0.071	0.790
Race/Other	Reference			
Region/Coastal	0.580	0.371	2.438	0.118
Region/Northern	0.411	0.451	0.831	0.362
Region/Central	0.265	0.292	0.824	0.364
Region/Western	Reference			
Country of Origin US	0.281	1.056	0.071	0.790
English Language	-0.166	0.772	0.046	0.829
US Citizen	1.043	1.324	0.620	0.431
Age (2006)	0.001	0.007	0.008	0.928

(continued)

ED Visits (2006)	-0.370	0.124	8.926	0.003
Hospital Days (2006)	0.191	0.140	1.858	0.173
Cost (2006)	0.000	0.000	6.680	0.010
Age x ED Visits	0.009	0.003	11.303	0.001
Age x Hospital Days	-0.004	0.003	2.799	0.094
Age x Costs	0.000	0.000	0.733	0.392

13 variables were retained because propensity scores should include all variables that may play a role in the treatment assignment process, even if they are not statistically significant (Shadish et al., 2002).[30]

The results of the overall propensity score regression model appeared adequate. The regression model correctly classified 79.0 percent of the subjects using 14 percent as the classification cutoff score (14.0 percent of the subjects were in the treatment group). The likelihood ratio chi-square statistic of the propensity score model was significant (χ^2 = 218.50, df = 16, p = 0.000), indicating that the model provided a good fit to the data. More importantly, the propensity score model *c*-statistic was 0.783, indicating that the model discriminated effectively between the individuals in the treatment and control groups based on the observed covariates. In other words, the model assigned a high probability of

[30] It is acceptable to use nonparsimonious models to estimate propensity scores. In fact, some researchers use hundreds of variables to estimate propensity scores (Stukel et al., 2007).

treatment assignment to the treatment subjects in 78.3% of all possible subject pairs (Hosmer & Lemeshow, 2000; Peng et al., 2002).[31,32]

After performing the regression, the adequacy of the actual propensity scores was assessed by examining the overlap in the distribution of scores between the treatment and control groups. This was accomplished by reviewing the ranges of the estimated propensity scores for the treatment and control subjects as well as two histograms that depicted the distribution of scores for both groups (Love, 2003; Linden et al., 2005). Overlap between the two study groups is necessary for comparability between the treatment and control subjects. A lack of overlap implies that there are combinations of covariate values that are found in only one of the two study groups, suggesting that the treatment and control subjects cannot be meaningfully compared (Landrum & Ayanian, 2001).

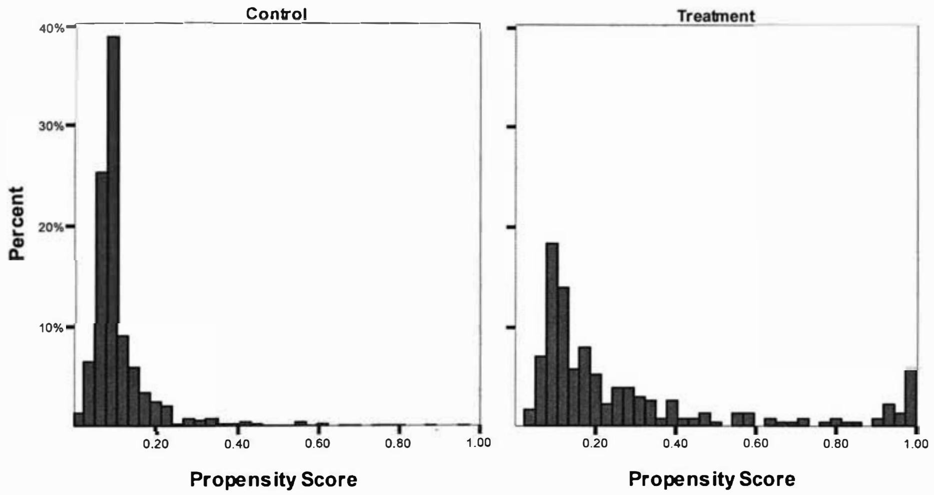
Based on this examination, substantial overlap appeared to exist between the treatment and control groups. The estimated propensity scores for the treatment subjects ranged between 0.03 and 1.00 (*M* = 0.30, *SD* = 0.28), while the estimated propensity scores for the control

[31] While propensity score methods have been in existence for more than two decades, debate still exists over how best to calculate and evaluate propensity score regression models (Weitzen, Lapane, Toledano, Hume, & Mor, 2004; Austin et al., 2007; Baser, 2006; Hill, 2008; and Caliendo & Kopeinig, 2008). For instance, some researchers use goodness of fit measures such as the Hosmer-Lemeshow chi-square test and *c*-statistics in addition to propensity score overlap and covariate balance to assess propensity score models (Baser, 2006), while others use *c*-statistics and likelihood ratio tests to assess the models (Normand, Sykora, Mamdani, Rochon, & Anderson, 2005; Jones & Richmond, 2006). Still, others argue that goodness of fit measures are of little value in assessing propensity score models (Love, 2004). Due to this debate, the researcher assessed the appropriateness of the propensity score model using the likelihood ratio test, *c*-statistic, propensity score overlap, and covariate balance. While not reported in the text, the researcher did calculate the Hosmer-Lemeshow test statistic to assess the model's classification power. The statistic was significant (χ^2 = 17.22, df = 8, p = 0.000) indicating that the model did not have good classification power. However, Allison (1999) reports that this statistic is "ad hoc" and may not be very powerful.

subjects ranged between 0.00 and 0.97 ($M = 0.12$, $SD = 0.10$). This overlap is illustrated in Figure 3, which shows that a majority (86.7%) of the scores fell in the 0.00 – 0.20 categories. This finding suggests that the treatment group mirrors the larger population of diabetes DM recipients from which they were selected (Linden et al., 2005). Figure 3 also shows that a slightly larger percentage of treatment subjects had scores greater than 0.40. Overall, these findings indicate that the treatment and control groups had reasonable overlap at all but the highest values of the propensity scores (Newgard et al., 2004).

Finally, simple regression analyses were performed to determine if the propensity scores produced balance between the two study groups on the covariates (Table 3). To

Figure 3
Propensity Score Distributions



[32] The *c*-statistic ranges between 0.5 and 1.0, with a value of 0.5 indicating that the model is no better than assigning subjects randomly to the treatment and control groups and a value of 1.0 indicating that the model assigns high probabilities to all subjects in the treatment group (Peng et al., 2002).

Table 3

Simple Regression Analysis to Assess Covariate Balance (N = 1,627)

Variables	Treatment (*n* = 229)*	Control (*n* = 1,398)*	Unadjusted p-value**	Adjusted p-value**
Female	166 (72.5%)	986 (70.5%)	0.546	0.771
Nonwhite Race	52 (22.7%)	198 (14.2%)	0.001	0.441
Western Region	170 (74.2%)	1,166 (83.4%)	0.001	0.485
Country of Origin US	225 (98.3%)	1,343 (96.1%)	0.110	0.507
English Language	226 (98.7%)	1,386 (99.1%)	0.511	0.964
US Citizen	226 (98.7%)	1,355 (96.9%)	0.147	0.507
Age	46.6 (14.25)	44.93 (15.70)	0.131	0.577
Hospital Days	4.90 (29.75)	0.44 (1.91)	0.000	0.644
ED Visits	2.36 (3.65)	1.50 (3.20)	0.001	0.933
Diabetes Related Costs	$7,473.33 ($16,962.42)	$1,060.48 ($2,380.31)	0.000	0.191

*Frequencies and (percentages) are presented for the categorical variables, and means and (standard deviations) are presented for the quantitative variables.

**Adjusted p-values were calculated with the propensity score included in a logistic regression analysis modeling the relationship between each covariate and the program participation variable.

perform these analyses, the treatment variable was regressed separately on each covariate with and without the propensity scores. To simplify the analyses, the race and geographic region variables were transformed into two new dichotomous variables: nonwhite race (1 = yes, 0 = no) and western region (1 = yes, 0 = no). After examining the *p*-values unadjusted for the propensity scores, it was determined that significantly different covariate distributions existed between the study groups on the nonwhite race, western region, and 2006 age, utilization, and cost variables. Using the covariate distributions as a proxy for comparability between study groups indicated that the two groups were not comparable on these variables. However, after adjusting for the propensity scores, no significant differences remained in the covariate distributions of the study groups (Newgard et al., 2004). These results suggest that the propensity scores could control for some potential overt confounding variables and were appropriate to use in the OLS and IV regression models.

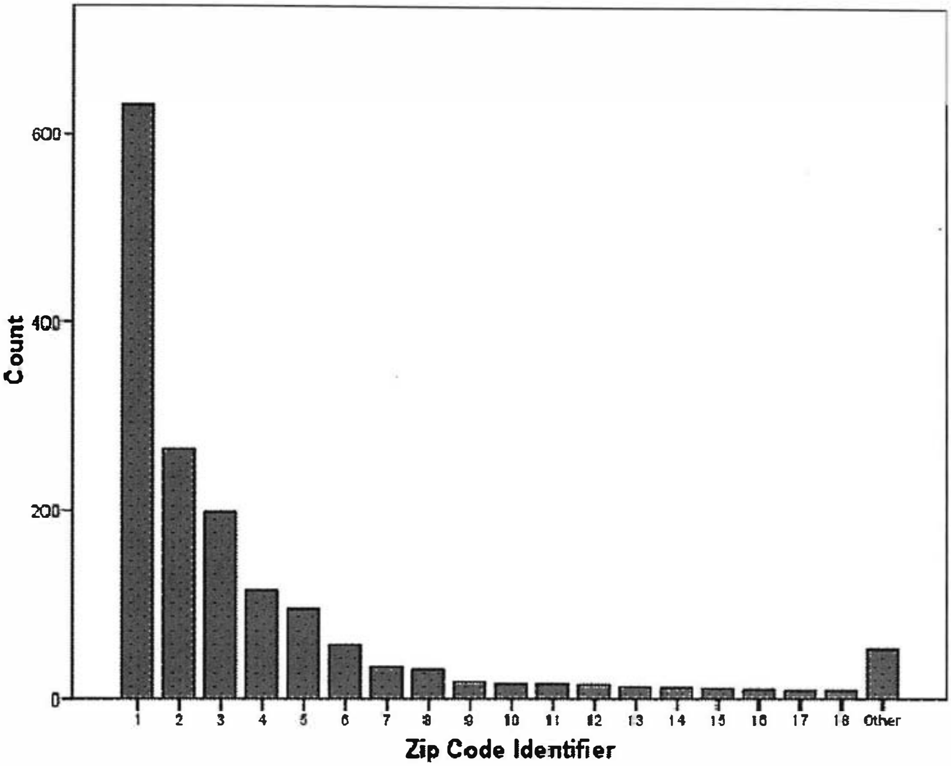
Instrumental Variables Calculation

The instrumental variables calculation procedure began by identifying the unique number of five-digit zip codes in the sample (Linden & Adams, 2006). A total of 337 unique five-digit zip codes were identified. By collapsing these based on the first three zip code digits, 32 new zip code categories were developed. An examination of these zip code categories revealed that 14 categories contained nine or less participants. (A total of 54 participants were contained in these nine categories.) Following guidance by Linden and Adams (2006), these participants were grouped into a zip code category called "other".[33] A new categorical variable was then created (called zip-code identifier) that contained 19 levels corresponding to the 19 three-digit zip code categories. The frequencies of the zip code categories used as instrumental variables in the study are depicted in Figure 4. These

[33] The researcher originally intended to group individuals in zip code categories with nine or less participants into geographically similar zip code categories with more than nine participants. However, upon examining a three digit zip code map of Virginia, it was not readily apparent which zip code categories these individuals should be grouped in because the categories cover large geographic areas bordering multiple three-digit zip code categories.

Figure 4

Frequency Distribution for the Three-Digit Zip Code Identifier Variable



results are comparable to the results reported by Linden and Adams (2006). The zip code categories were used as dummy variables in the first stage of the two stage least squares (2SLS) IV regression models.

A logistic regression model was next developed to estimate the probability of a given individual participating in the high intensity open option of the Virginia *Healthy Returns*^{SM} DM program using age (years), gender (1 = female and 0 = male), propensity scores (0.0 - 1.0), and the zip code categories (Linden & Adams, 2006). While not specifically stated, Linden and Adams (2006) apparently performed this calculation to determine if the IV assumption that Z is associated with X was satisfied, and to calculate the probability of program participation that they used as the endogenous variable in the two-stage least squares (2SLS) IV regression models.[34] In other words, they did not use actual program participation as the endogenous variable in the IV regression models.

The reasoning of Linden and Adams (2006) for doing this is unclear, however, because 2SLS automatically performs this calculation through two simultaneous least squares regressions (Greene, 2003). Essentially what the two stages do simultaneously is this: the first stage uses a set of instruments (and the other exogenous variables in the model) to generate least squares predictions of the endogenous variable, X, and then uses these predictions in lieu of the actual X to explain variation in the outcome variable in the second stage.[35]

Because the present study sought to test the feasibility of Linden and Adams' IV procedure, logistic regression was used to estimate the probability of program participation. The predicted values of program participation from the logistic regression were then used to determine if Z is associated with X, and a 2SLS regression model was developed for each outcome variable using this value as the endogenous variable in the

[34] Specifically, they reported (p.151) that "[i]ndependent variables included those exogenous variables from the first stage regression: age, gender, risk score, as well as the 'plug-in' or predicted value of program participation." Linden reported that the predicted probability of program participation was used as the endogenous variable instead of actual program participation (Linden, personal communication, June 16, 2008).

[35] The first stage attempts to identify variation in the treatment variable that is uncorrelated with the disturbance term. The second stage uses OLS regression to regress the outcome on just that portion of the treatment variable identified as being uncorrelated with the disturbance term in the first stage. This occurs because the predicted value of the treatment variable from the first stage is a linear combination of the instruments and the other exogenous variables. Because the instruments are uncorrelated with the disturbance term, the treatment variables' predicted value is uncorrelated with the disturbance term (Judd & Kenny, 1981).

first stage and its least squares predictions as the plug-in values in the second stage. (These regressions are referred to as the block one models in the OLS and IV regression section in Chapter 4.)

However, using the estimated probability of program participation as the endogenous variable in 2SLS appears unnecessary because this variable is calculated in the first stage. Based on the information presented in their article, Linden and Adams (2006) calculated the probability of program participation using gender, age, propensity scores, and zip code categories in a logistic regression model. They then used this value as the endogenous variable instead of actual program participation in the first stage where they appear to re-estimate this probability again by regressing it on the same set of variables used to calculate it in the logistic regression. (In other words, the estimated probability of program participation served as the dependent variable in the first stage of 2SLS instead of the treatment variable.) This procedure does not appear to comply with the 2SLS guidance provided by Greene (2003), Wooldridge (2002 and 2006), Gelman and Hill (2007), and Stock and Watson (2007). As a result, the researcher recalculated the three 2SLS models using actual program participation as the endogenous variable in the first stage and its least squares predictions as the plug-in value in the second stage to determine if the models were sensitive to the specific form of the program participation variable used in the analysis. (These regressions are referred to as the block two models in the OLS and IV regression section of Chapter 4.)

The logistic regression model developed to predict the probability of program participation contained 18 dummy variables that represented the 19 three digit zip code categories as well as 2006 age (years), gender (1 = female and 0 = male), and the propensity score (0.0 – 1.0). This model was estimated based on the actual high intensity open program participation variable with 1 indicating program participation (or treatment) and 0 indicating nonparticipation (or control). The model was used to estimate each subjects' probability of high intensity open DM participation, which was used as an endogenous variable in the 2SLS regressions.

Diagnostic statistics indicated that the logistic regression model was sound. In particular, the likelihood ratio chi-square statistic was significant ($\chi 2 = 232.77$, df = 21, $p = 0.000$) indicating that the model adequately fit the data, and the *c*-statistic was 0.786 indicating that the model discriminated effectively between the treatment and control subjects. The model also correctly classified 75.2% of the subjects, using 14% as the cutoff score. Moreover, the Somer's D and Gamma statistics for the model were 0.57 and 0.58, which indicated that it had good predictive ability. For comparison purposes, Linden and Adams (2006) reported that their model was significant ($p < 0.0001$) and had very good predictive ability based on the Somer's D and Gamma measures, which were 0.61 and 0.65.

Table 4 is similar to a table that Linden and Adams (2006) presented in their article comparing the characteristics of the treatment and control groups using actual and predicted program participation status. They reported that their predicted enrollment model "compared favorably" to the actual participation data and that there were no significant differences in the age and gender distributions for the predicted enrollment

Table 4
A Comparison of Group Characteristics Based on Actual and Predicted High Intensity Program Participation

	Group	N (%)	Age (SD)	Female Count (%)
Actual				
	Treatment	229 (14.1)	46.6 (14.25)	166 (72.5)
	Control	1,398 (85.9)	44.93 (15.70)	986 (70.5)
Predicted				
	Treatment	465 (28.6)	48.47 (14.23)	362 (77.8%)
	Control	1,162 (71.4)	43.84 (15.81)	790 (68.0%)

groups. They also noted that their predicted model slightly over-estimated program participation. Their actual study groups consisted of 1,952 (77%) program participants and 582 (23%) control participants. However, their predicted study groups contained 2,029 (81%) program participants and 505 (19%) control participants. Based on this information, they concluded that their results satisfied the IV regression assumption that Z is associated with X.

The results of the enrollment prediction model developed for the present study do not appear to compare favorably to the actual enrollment data because the model substantially over predicted high intensity program participation by 236 subjects. Significant differences also existed between the age and gender variables for the predicted enrollment groups. Based on the reasoning used by Linden and Adams (2006), these results appear to suggest that the zip code instruments are not related to the predicted probability of program enrollment. However, this assumption can be tested by regressing X on Z. If the null hypothesis is rejected at the 0.05 level, then statistically significant evidence exists that Z and X are related (Wooldridge, 2006). Linden and Adams (2006) did not report employing this procedure in their article.

To perform this test in the present study, the researcher regressed the probability of program enrollment on a scaled quantitative version of the zip code identifier variable where the levels represented the number of subjects in each three-digit category (i.e., level 1 = 631 subjects and level 19 = 10 subjects). The results of this regression suggest that Z is related to X ($t = 9.41, p = 0.000$).[36] To determine if the results were sensitive to the estimated probability of program enrollment, the researcher regressed the program participation probability's upper and lower 95% confidence limits (CL) on the scaled zip code variable. The results indicated that the initial regression model was not sensitive to the estimated probability of program enrollment (95% lower CL: $t = 2.51$ $p = 0.021$; 95% upper CL: $t = 34.36$, $p = 0.000$). In addition, the researcher regressed the actual program participation variable on the scaled zip code variable using logistic regression. The results further indicated that Z and X are significantly related (Wald $\chi^2 = 15.91, p = 0.000$). While Table 3 does not appear to provide evidence that Z is related to X, the results of the above regressions do provide evidence that Z and X are related.

[36] Because some observers may argue that Z is actually a categorical variable, the researcher regressed predicted program participation on 18 zip code dummy variables. This regression also revealed that Z and X are related ($F = 21.80, p = 0.000$). The researcher also regressed actual program participation on 18 dummy zip code variables. This regression further indicated that X and Z are related ($F = 3.56, p = 0.000$). However, the R^2 for this regression was 0.04, indicating that Z explain little variation in actual program participation. The implications of this finding is explored further later in this chapter.

Data Screening

Descriptive statistics for the study's outcome and independent variables are presented in Table 5. The information shows that considerable variability exists in the 2007 age, utilization, and cost variables. (For consistency, the 2007 age variable was used in the OLS and IV regression models because 2007 cost and utilizations were used as the outcomes.) The information also shows that the propensity score and predicted program

Table 5
Descriptive Statistics for the Study Variables (N = 1,627)

Variables	Mean (SD)	Frequency (%)	Range
Outcome Variables			
Hospital Days (2007)	0.99 (3.80)		0 - 64
ED Visits (2007)	1.61 (3.18)		0 - 45
Diabetes-Related Costs (2007)	$2,491.49 ($7,366.45)		$0 - $144,705.10
Independent Variables			
Age (2007)	46.16 (15.15)		2 - 90
Propensity Score	0.14 (0.15)		0.00 – 1.00
Predicted Enrollment Probably Score	0.14 (0.14)		0.00 – 0.96
Program Participation (1 =yes)		229 (14%)	
Female (1 = yes)		1,152 (70.8%)	

enrollment probability variables are very similar. These results are not surprising, however, because the propensity score and probability of program participation variables were calculated using essentially the same set of covariates. Due to the similarity between the propensity score and program participation probability variables, including both in the OLS and IV regression analyses may produce multicollinearity. Linden and Adams (2006) did not report descriptive statistics for their study variables, so it is impossible to compare their propensity score and program participation probability variables to the ones developed for the present study. However, they did report checking their regression models for multicollinearity, but detected none. The multicollinearity issue is examined later in this section.

Boxplots and histograms for the quantitative variables as well as frequency distributions for the qualitative variables were generated during the data screening process. The boxplots revealed that univariate outliers were present in all of the quantitative variables, and the histograms revealed that these variables were all positively skewed, with the exception of the age variable that was approximately normally distributed. Skewness and kurtosis statistics further indicated that the distributions for the utilization, cost, and propensity score and program participation probability variables deviated substantially from normality.[37] This finding may not be too problematic, however, because regression is robust to violations of normality (Mendenhall & Sincich, 2003). The frequency

[37] The skewness and kurtosis statistics for the variables were as follows: Age (-0.60, -0.91), ED Visits (4.50, 34.74), Hospital Days (7.55, 83.99), Cost (8.72, 119.38), Propensity Scores (3.79, 15.82), and Program Participation Probability (3.53, 13.79). Ideally, these statistics should be close to zero (Tabachnick & Fidell, 2001).

distributions for the gender and program participation variables did not reveal any aberrant observations.

To examine the variables further for univariate outliers, standardized scores were calculated for all six quantitative variables. Using 3.29 standard deviations as the threshold for identifying univariate outliers, none were detected in the age variable, but numerous outliers were found in the utilization, cost, and propensity score and program participation probability variables (Tabachnick & Fidell, 2001).[38] Mahalanobis and Cook's distances were next calculated to screen for multivariate outliers. Mahalanobis distances were evaluated using the chi-square distribution with degrees of freedom equal to the number of predictors in the model ($\chi^2 = 20.52$, df = 5, $p < 0.001$).[39,40] Cases with Mahalanobis distances greater than 20.52 and Cook's distances greater than 1.0 were identified as outliers (Tabachnick & Fidell, 2001). Using these measures, 54 multivariate outliers were identified.

Outlier observations either represent contaminated data or accurate observations of rare cases. Outliers due to contaminated data are usually either corrected or deleted. Because the data came from the Virginia Medicaid Management Information System (VaMMIS), the researcher assumed that the observations represented legitimate cases.

[38] The number of univariate outliers were as follows: ED Visits (28), Hospital Days (31), Cost (23), Propensity Score (44), and Program Participation Probability (50).

[39] For the purposes of data screening, five predictors were used in the OLS regressions: program participation, age, gender, propensity score, and the probability of program participation.

[40] Tabachnick and Fidell (2001) report that the Mahalanobis distance statistic can be used to identify multivariate outliers. In most large datasets, cases group around the point created by the intersection of the means of all variables in the set. The Mahalanobis distance represents the distance that the cases are from the intersection. Cases that are outside the group are multivariate outliers. Tabachnick and Fidell recommend using a chi-square distribution and a very conservative probability estimate ($p \leq 0.001$) to identifying multivariate outliers based on their Mahalanobis distances.

(However, this assumption could not be verified.)[41] Consequently, the researcher decided against simply deleting the outliers and instead focused on examining how their presence influenced the analytical results before determining their ultimate fate. This was accomplished by regressing the outcomes on the independent variables with the outliers included and excluded, examining residual plots, and applying transformations to the variables to determine if they reduced the number of outliers or produced any noticeable improvements in the fit of the regression models.

Each outcome was regressed on the independent variables with the propensity score and program participation probability outliers included and excluded. Based on a review of the model F statistics and p-values, R^2 statistics, coefficient p-values, and residual plots little appeared to be gained by excluding the outliers from the analysis. Each outcome variable with the outliers included and excluded was next regressed on the independent variables and the above statistical measures were reexamined. These results also indicated that little would be gained by excluding the outliers from the regression analyses. Finally, the regressions were performed with the multivariate outliers included and excluded, which further indicated that eliminating the outliers would produce no substantive improvements in the regression models. Based on this information, a decision was made to retain the outliers in the dataset.

The natural logarithmic (i.e., log) and square root transformations were then applied to the propensity score and program participation probability variables to

[41] Claims data can contain inaccuracies (Linden et al., 2005). However, the researcher was unable to compare the actual claims that were submitted by Medicaid providers to DMAS to verify the accuracy of the data used in this study.

determine if these transformations produced any improvements in the fit of the regression models. Such transformations are usually recommended for skewed variables (Tabachnick & Fidell, 2001). While the transformations improved the skewness and kurtosis measures of the variables, the regression analyses indicated that they produced no substantive improvements to the fit of the models.[42,43] Consequently, the untransformed propensity score and probability of program participation variables were used in the OLS and IV regression models.

The log transformation was next applied to the cost variable, while the square root transformation was applied to the ED visits and hospital days variables.[44] Log transformations are generally applied to dollar outcomes, while square root transformations are applied to count outcomes (Cohen et al., 2003; Mendenhall & Sincich, 2003).[45] These transformations improved the skewness and kurtosis statistics for the outcomes.[46] An examination of the residual plots for the transformed outcomes indicated that the transformations appeared to satisfy the assumptions of residual normality, linearity, and homoscedasticity. The log transformation also eliminated all univariate outliers in the cost outcome variable, while the square root transformations only reduced the number of

[42] The log transformation actually increased the kurtosis for the probability of program participation variable from 13.79 to 52.82.

[43] A decision was made against employing additional transformations due to the complexities associated with interpreting regression coefficients for transformed variables (Gelman & Hill, 2007).

[44] The natural log transformation does not calculate log transformed values for observations that are zero. The log transformation did not calculate logs for 252 subjects who did not incur diabetes-related health care costs during CY 2007. Zeros were inserted into the log transformed cost variable to replace these missing values.

[45] The researcher did not compare regression models between the original outcomes and their transformations because models are not directly comparable (Cohen et al., 2003).

outliers in the count outcomes. Residual histograms and p-p plots further revealed that the log cost and square root ED visits variables were approximately normally distributed, while the square root hospital days variable still remained skewed, although to a lesser extent than the untransformed hospital days variable.

Because the square root transformation neither resulted in the elimination of all outliers in the count variables nor produced a residual distribution that was approximately normal for the hospital days variable, log and reciprocal transformations were applied to these variables (Tabachnick & Fidell, 2001). However, the transformations were rejected based on regression analyses, which indicated that they did not result in any substantive improvements over the analyses performed using the square root transformation. While the square root transformation may not be optimal for the hospital days variable, this transformation was still applied to both count outcomes because: 1) it is viewed as an acceptable transformation for count outcomes when using OLS regression (Cohen et al., 2003; Mendenhall & Sincich, 2003), 2) it reduced variable skewness and kurtosis, and 3) it improved residual normality, linearity, and homoscedasticity.

Finally, the data were screened for multicollinearity by generating a correlation matrix and calculating variance inflation factors (VIFs). Bivariate multicollinearity was defined as intercorrelations of at least 0.80, while multivariate multicollinearity was defined as variance inflation factors (VIF) that exceed 10 (Stevens, 2002). All study variables were included in the correlation matrix. A review of the intercorrelations

[46] The skewness and kurtosis statistics for the log cost variable were –0.91 and 0.12 respectively, while these measures were 1.26 and 1.53 for the ED Visits and 3.20 and 11.71 for the Hospital Days square root variables.

revealed that a strong highly significant ($\rho = 0.94$, $p = 0.000$) correlation existed between the propensity score and program participation probability variables. While only the propensity score variable will be included in the OLS models, including both variables in the second stage of the IV regression models may result in unreliable regression coefficients, standard errors, confidence intervals, and *p*-values for the predictor variables. VIFs were next calculated by regressing each transformed outcome on the independent variables. None of the VIFs were greater than 10. However, the VIFs for the propensity score and program participation probability variables were approximately 8.0 in all regression models. While these VIFs did not meet the multivariate multicollinearity "rule of thumb" adopted for this study, the high values are problematic and may produce multicollinearity in the 2SLS models (Cohen et al., 2003).[47]

Because the present study sought to assess the feasibility of Linden and Adams (2006) IV regression procedure, the IV regression models were generated using the propensity score as an independent variable and the probability of program participation as the endogenous variable. However, to test the sensitivity of the IV regression models to the inclusion of these highly correlated variables, the propensity score variable was stratified into five quintiles, with each quintile containing an approximately equal number of subjects based on their propensity scores.[48] The OLS and IV regression models were

[47] When the probability of program participation was removed and the regressions redone, the VIFs for the actual program participation variable were 1.22.

[48] For each quintile, the number of subjects and the average propensity scores were as follows: quintile 1 (325 subjects, 0.081 average propensity score), quintile 2 (326 subjects, 0.083 average propensity score), quintile 3 (325 subjects, 0.103 average propensity score), quintile 4 (326 subjects, 0.110 average propensity score), and quintile 5 (325 subjects, average propensity score 0.328).

then redone using this stratified variable in place of the quantitative propensity scores to determine if it influenced the regression models. (These regressions are referred to as the block three models in the present study.) Including propensity scores as a quintile covariate in multiple regression is an acceptable means of employing this procedure (D'Agostino, 1998). In fact, Afifi et al. (2007) included a propensity score quintile covariate in the two-part regression models they used to evaluate the Florida Medicaid DM program.[49]

Ordinary Least Squares and Instrumental Variables Regression Models

The results of the ordinary least squares (OLS) and instrumental variables (IV) regression models developed for the present study are presented in this section. To address the study's research questions, three blocks of OLS and IV regression models were generated. Each block contained six regression models. The OLS and IV regressions were performed using STATA version 10.1.

In the first block, three IV regression models were developed using the procedure employed by Linden and Adams (2006). Two-stage least squares (2SLS) regression was used to generate the IV models. In the first stage, the probability of program participation (the endogenous variable) was regressed on age, gender, and the propensity score variables (the exogenous variables), plus the 18 dummy zip code instruments. In the second stage, the outcome variables were regressed on the predicted program probability values from the

[49] Afifi et al. (2007) actually used quintile indicators as covariates in their regression models. However, indicator quintiles were not used as covariates in this study for two reasons: 1) D'Agostino (1998) reports that it is acceptable to include "the propensity score quintile itself" (p. 2276) as a covariate in regression and 2) using quintile indicators in the block three models would have complicated efforts to assess the sensitivity of the regression models in all three blocks to the specific form of the propensity score variable.

first stage plus the age, gender, and propensity score variables. In the OLS models, the outcomes were simply regressed on the actual program participation, age, gender, and propensity score variables.

To test the assumption that the IV estimates were not significantly related to the outcomes, residuals from each IV model were regressed on the zip code instruments (Linden & Adams, 2006). Nonsignficant model *F* statistics suggest that no direct relationships exist between the instruments and the outcomes. The Sargon chi-square test was used to test the assumption that at least some of the instruments were uncorrelated with the disturbance terms in the IV models.[50] While not definitive, a nonsignificant *p*-value for this test suggests that the instruments and the disturbance term are uncorrelated (Gujarati, 2003; Baum, 2006). Finally, the Hausman specification chi-square test was used to determine if significant differences exist between the OLS and IV coefficients (Ender, n.d.; Greene, 2003; Hadley et al., 2003; Baum, 2006).[51] If no significant differences exist, then the OLS model becomes the default because its coefficients are more efficient (Linden & Adams, 2006).

In the second block, the same procedures were performed as above except that the actual high intensity program participation variable was used in the 2SLS regressions to determine if the models were sensitive to the particular form of the endogenous variable.

[50] According to Linden and Adams (2006), the IV assumption that Z is not associated with the unobserved error cannot be "definitively" assessed. However, Wooldridge (2006) argues that if researchers have more than one instrumental variable, they "can effectively test whether some of them are uncorrelated with the structural error" (p. 533).

[51] Wooldridge (2002) indicates that the Hausman test is a test of the possible endogeneity of the problematic independent variable (i.e., the treatment variable when subjects are assigned nonrandomly). If the problematic variable is uncorrelated with the disturbance term, then the OLS and IV models should only differ by sampling error.

In the third block, the same procedures were performed as in the second block, but the quantitative propensity score variable was replaced with a quintile variable to test the sensitivity of the models to this variable's specific form.

The results of the three regression model blocks are presented in the following subsections.

Block One OLS and IV Regression Models

The results for the block one OLS and IV regression models are presented in Table 6. To test the assumption that the IV estimates were not directly related to the outcomes, residuals from each IV model were regressed on the 18 dummy zip code instruments (Linden & Adams, 2006). The p-values for the model F-test suggested that no direct relationships exist between the instruments and the residuals for the cost model ($F = 1.28$, $p = 0.183$) and the hospital days model ($F = 0.80$, $p = 0.707$). However, a direct

Table 6
Block One: OLS and IV Regression Models Using Linden and Adams' Procedure

	OLS Regression			*IV Regression*		
Variable	*Coefficient (SE)*	*95% CI*	*p-value*	*Coefficient (SE)*	*95% CI*	*p-value*
Outcome: Diabetes-Related Costs (Log)						
Intercept	3.11 (0.22)	2.68 – 3.54	0.000	3.11 (0.22)	2.67 – 3.55	0.000
Program	1.53 (0.20)	1.14 – 1.93	0.000	0.84 (1.37)	-1.84 – 3.52	0.538
Female	-0.06 (0.14)	-0.34 – 0.21	0.648	-0.06 (0.14)	-0.34 – 0.22	0.655
Age	0.04 (0.00)	0.03 – 0.05	0.000	0.04 (0.00)	0.03 – 0.05	0.000
Propensity	3.71 (0.46)	2.79 – 4.62	0.000	4.36 (1.38)	1.67 – 7.06	0.002
Adjusted R^2	0.16			0.13		
Sargon $\chi 2$						0.109
Hausman $\chi 2$						0.604

(continued)

Outcome: Emergency Department Visits (Square Root)						
Intercept	1.05 (0.08)	0.89 – 1.22	0.000	1.05 (0.08)	0.89 – 1.22	0.000
Program	0.20 (0.08)	0.04 – 0.35	0.011	0.04 (0.52)	-0.97 – 1.05	0.940
Female	0.16 (0.05)	0.05 – 0.26	0.003	0.16 (0.05)	0.05 – 0.26	0.003
Age	-0.01 (0.00)	-0.01 – -0.01	0.000	-0.01 (0.00)	-0.015 – -0.008	0.000
Propensity	0.78 (0.18)	0.43 – 1.13	0.000	0.93 (0.52)	-0.09 – 1.95	0.074
Adjusted R^2	0.05			0.05		
Sargon $\chi 2$						0.000
Hausman $\chi 2$						0.759
Outcome: Hospital Days (Square Root)						
Intercept	-0.00 (0.08)	-0.15 – 0.15	0.993	-0.00 (0.08)	-0.15 – 0.15	0.989
Program	0.42 (0.07)	0.28 – 0.56	0.000	0.51 (0.47)	-0.42 – 1.44	0.283
Female	-0.01 (0.05)	-0.11 – 0.08	0.771	-0.01 (0.05)	-0.11 – 0.08	0.772
Age	0.00 (0.00)	-0.00 – 0.01	0.060	0.00 (0.00)	-0.00 – 0.01	0.062
Propensity	1.15 (0.16)	0.84 – 1.47	0.000	1.07 (0.48)	0.13 – 2.00	0.025
Adjusted R^2	0.08			0.07		
Sargon $\chi 2$						0.574
Hausman $\chi 2$						0.847

relationship appears to exist between the instruments and the residuals for the ED visits model ($F = 2.36$, $p = 0.001$). This finding suggests that three-digit zip codes are not good instruments for the endogenous variable in the ED visits model. The Sargon chi-square test indicated that the zip code instruments were not associated with the disturbance terms in the cost ($p = 0.109$) and hospital days ($p = 0.574$) IV models. However, the test found that at least some of the instruments were associated with the disturbance term in the ED visits IV model ($p = 0.000$). This finding also provides evidence that three-digit zip codes are probably not good instruments in the ED visits model.

The Hausman specification test indicated that there were no significant differences between the OLS and IV models. Based on this information, the OLS models should be used to interpret the effects of the high intensity DM program on the outcomes for the 1,627 diabetes study subjects because the parameter estimates for the OLS and IV models were not significantly different. The implication is that the probability of program participation variable (the endogenous variable in the first stage of the 2SLS equations) may not actually be endogenous (i.e., correlated with the models' disturbance terms) (Wooldridge, 2002). Nevertheless, care should be exercised when using the OLS models to interpret the causal effects of the high intensity DM program because they may not actually control for all overt and hidden biases. Failing to control for these biases can result in incorrect regression coefficients, standard errors, confidence intervals, and *p*-values for the model covariates.

The information in Table 6 also shows that the probability of program participation was not significant in the IV models. This finding is surprising since the program variable is highly significant in the OLS models. It may be due to: 1) how the IV program variable was calculated (i.e., it was estimated in a separate logistic regression model and then used as the treatment variable in the first stage of 2SLS where it was estimated again using the same variables that predicted it in the logistic regression model), and 2) the high correlation between the program probability and propensity score variables may have resulted in larger variances for the IV estimators, which can translate into larger confidence intervals and imprecise *p*-values (Wooldridge, 2006).

To investigate the second stage IV regression models further, variance inflation factors (VIFs) were calculated for all independent variables. The VIFs for the propensity score and probability of program participation variables were 10.3 in all IV regression models indicating that multicollinearity was probably present. Multicollinearity is a serious problem in OLS regression because it can produce large standard errors for the coefficients. However, it is even more serious in 2SLS regressions because coefficient variances will be larger due to: 1) the predicted value of the endogenous variable having less variation than the actual endogenous variable and 2) the correlation between the predicted value of the endogenous variable and the other exogenous variables being much stronger than the correlation between the actual endogenous variable and the other variables (Wooldridge, 2006). Because multicollinearity appears to be present in the block one IV models, the overall results may be incorrect.

Because the diabetes-related cost variable was log transformed, the OLS regression coefficients for the cost model are interpreted as percent changes (Wooldridge, 2006).[52] The program variable in the model represents the average effect of the high intensity program on the outcome when the other covariates are held fixed. The results of this model indicate that high intensity DM program participation is associated with a 153% increase in diabetes-related health care costs on average when controlling for age, gender, and propensity scores. The 95% confidence interval for the coefficient estimate indicates that the true program effect is probably between 114% and 193%. While the estimated program effect may seem large, it should be noted that average diabetes-related costs for the treatment and control groups were $7,990.39 and $1,590.74, respectively (t = -5.97, p =0.000).[53]

The program estimate in the IV model is much smaller, suggesting that high intensity program participation is only associated with an 84% increase in diabetes-related health care costs on average. The 95% confidence interval for the IV program estimate (-1.84 to 3.52) is also much larger than the estimate's OLS confidence interval. This is due to the fact that the IV standard error of the estimate is larger (1.37) than the OLS standard error (0.20).[54] In IV regression, coefficient confidence intervals and standard errors are usually larger than in OLS regression because the variance of the IV estimator is calculated by incorporating $R^2{}_{\mathrm{X,Z}}$ (i.e., a measure of the strength of the linear relationship between X and Z) into the denominator of the variance formula (Wooldridge, 2006).[55] However, the overall usefulness of this information is questionable because the program participation variable is not significant (p = 0.538) and the IV cost model was rejected in favor of the OLS model.

An examination of the coefficient estimates and confidence intervals for the other independent variables reveals that the age and propensity score variables are significant in

[52] According to Wooldridge (2006), when the dependent variable is log transformed, the interpretation of the regression coefficient is as follows: % change in y = (100*coefficient)change in x.

[53] This is equal to a 402% change in diabetes-related health care costs between the treatment and control groups.

[54] Wooldridge (2006) indicates that this is the "price paid" to get a consistent estimator of the outcome if the treatment variable is endogenous.

both the OLS and IV regression models. In particular, the OLS and IV age estimate (0.04) and confidence intervals (0.03 to 0.05) were the same; however, the IV coefficient and confidence intervals for the propensity score were larger than its OLS estimate (4.36) and confidence intervals (1.66 to 7.06). The fact that the propensity score variable was significant in both the OLS and IV models is not surprising because this variable represents a "composite confounder" designed to reduce overt bias by controlling for 13 potential confounders (Newgard et al., 2004).

The coefficients in the ED visits and hospital days OLS and IV models are interpreted differently because the square root transformation was applied to these outcomes. In the OLS ED visits model, high intensity DM program participation is associated with a 0.20 ($p = 0.011$) increase on average in the square root of ED visits, while holding fixed the age, gender, and propensity score variables.[56] The 95% confidence interval for the program variable is 0.04 to 0.35. The IV estimate is much smaller, indicating that high intensity program participation is only associated with an average increase of 0.04 in the square root of ED visits when controlling for the effects of the other covariates. The estimate's IV confidence interval (-0.97 to 1.05) is also larger than its OLS confidence interval. However, these findings may also not be very informative because the p-value of the IV program estimate is 0.940 and the IV model was rejected in favor of the OLS model.

[55] In IV regression, the asymptotic standard error of β_1 is $\sigma_2/SST_x*R^2_{x,z}$ while the asymptotic standard error of β_1 in OLS regression is σ_2/SST_x. Because R^2 is always less than one, the IV variance is usually larger than the OLS variance. In fact, the smaller the $R^2_{x,z}$, the larger the IV variance (Wooldridge, 2006).

[56] Square root transformation estimates can be converted back to original units by squaring them. For instance, high intensity DM program participation is associated with a 0.04 (or 0.20^2) increase in ED visits.

The female, age, and propensity score variables were significant in the OLS ED visits model, while only the female and age variables were significant in the IV model. The variables' coefficient estimates, standard errors, confidence intervals, and *p*-values were identical in both models. This finding is surprising because IV regression estimates are usually larger than the OLS estimates (Winship & Morgan, 1999).

Finally, the OLS hospital days model indicates that the high intensity DM program is associated with an average increase of 0.42 ($p = 0.000$) in the square root of hospital days when controlling for the other covariates. The 95% confidence interval for this variable is 0.28 to 0.56. The IV program effect estimate is larger than the OLS estimate. The IV model estimates the average effect of the high intensity DM program to be 0.51 with a 95% confidence interval of -0.42 to 1.44. The wide confidence interval is due in part to the coefficient's large standard error (0.47). However, the usefulness of the IV estimate is also questionable because the variable's *p*-value was 0.283 and the IV model was rejected in favor of the OLS model.

In addition, only the propensity score covariate was significant in both the OLS and IV hospital days models. While the propensity score's OLS coefficient is larger (1.15) than its IV coefficient (1.07), its IV standard error is larger (0.48) than its OLS standard error (0.16). The variable's IV confidence interval is thus wider (0.13 to 2.00) than its OLS confidence interval (0.84 to 1.47).

The adjusted R^2 statistics for the six OLS and IV regression models are as follows: diabetes-related costs (0.16 and 0.13), ED visits (0.05 and 0.05) and hospital days (0.08 and 0.07). While some observers may argue that the regression models are not practically significant because the covariates did not account for much of the variability in the outcomes, it is important to remember that the adjusted R^2 statistic will always be smaller than the R^2 statistic because it penalizes for the inclusion of irrelevant variables in the model. Thus, it is possible to have practically significant regression models that account for a large portion of the variation in the outcome with small adjusted R^2 statistics. However, Linden and Adams (2006) report that small R^2 statistics (in the 0.03 to 0.06 range) are generally reported in the health services research literature for these outcomes.[57] In any case, it should be noted that R^2 statistics will usually be smaller for IV models than for OLS models. IV regression is intended to provide better estimates of the effect of X on Y when X is related to the disturbance term. Goodness-of-fit measures are not a factor in this process (Wooldridge, 2006).

For comparison purposes, Linden and Adams (2006) rejected their diabetes-related cost and ED visits IV models, but accepted their hospital days IV model based on the Hausman test results. They also found significant DM treatment effects in their OLS and IV estimates for the cost and hospital days models, but did not find significant effects in the ED visits model. In addition, they reported low adjusted R^2 statistics for their OLS and IV cost (0.095 and 0.094), ED visits (0.006 and 0.006) and hospital days (0.034 and 0.029) models.

[57] According to Wooldridge (2006), R^2 statistics will be larger for OLS models than for IV models because OLS "minimizes the sum of squared residuals" (p. 525). Wooldridge further reports that R^2 statistics are not very useful measures in IV regression.

Block Two OLS and IV Regression Models

The results of the block two OLS and IV regression models are presented in Table 7. These models were developed in the same manner as the block one models, but the actual program participation variable was used as the endogenous variable in the 2SLS regressions. Model development for the second block began by testing to determine if the assumption that the instruments are not directly related to the outcomes was met. This was accomplished by regressing the residuals from each IV model on the zip code instruments. The model *F*-tests indicated that no direct relationships existed between the instruments and the residuals for the cost model ($F = 1.38, p = 0.130$) and the hospital days model ($F = 0.86, p = 0.626$). However, the *F*-test for the ED visits model indicated that a direct relationship existed between the instruments and the residuals ($F = 2.49, p = 0.001$). These results further suggest that three-digit zip codes may not be good instruments in the ED visits model.

The Sargon test indicated that the instruments were not related to the disturbance terms in the cost ($p = 0.092$) and hospital days ($p = 0.552$) models. However, the test indicated that the instruments were related to the disturbance term in the ED visits model ($p = 0.000$). This finding provides additional evidence that three-digit zip codes may not be appropriate instruments for the treatment variable in the ED visits model.

The Hausman test revealed that no significant differences existed between the OLS and IV models when using actual program participation as the endogenous variable. Based on this information, the OLS models could be used to interpret the effects of the high

Table 7

Second Block: OLS and IV Regression Models Using Actual Program Participation as the Endogenous Variable

	OLS Regression			*IV Regression*		
Variable	*Coefficient (SE)*	*95% CI*	*p-value*	*Coefficient (SE)*	*95% CI*	*p-value*
Outcome: Diabetes-Related Costs (Log)						
Intercept	3.11 (0.22)	2.68 – 3.54	0.000	3.11 (0.22)	2.68 – 3.54	0.000
Program	1.53 (0.20)	1.14 – 1.93	0.000	0.84 (1.35)	-1.80 – 3.48	0.532
Female	-0.06 (0.14)	-0.34 – 0.21	0.648	-0.06 (0.14)	-0.34 – 0.21	0.650
Age	0.04 (0.00)	0.03 – 0.05	0.000	0.04 (0.00)	0.03 – 0.05	0.000
Propensity	3.71 (0.46)	2.79 – 4.62	0.000	4.36 (1.36)	1.71 – 7.02	0.001
Adjusted R^2	0.16			0.16		
Sargon $\chi 2$						0.092
Hausman $\chi 2$						0.604
Outcome: Emergency Department Visits (Square Root)						
Intercept	1.05 (0.08)	0.89 – 1.22	0.000	1.05 (0.08)	0.89 – 1.22	0.000
Program	0.20 (0.08)	0.04 – 0.35	0.011	0.04 (0.52)	-0.97 – 1.05	0.940
Female	0.16 (0.05)	0.05 – 0.26	0.003	0.16 (0.05)	0.05 – 0.26	0.003
Age	-0.01 (0.00)	-0.015 - -0.008	0.000	-0.01 (0.00)	-0.015 - -0.008	0.000
Propensity	0.78 (0.18)	0.43 – 1.13	0.000	0.93 (0.52)	-0.09 – 1.95	0.074
Adjusted R^2	0.05			0.05		
Sargon $\chi 2$						0.000
Hausman $\chi 2$						0.759
Outcome: Hospital Days (Square Root)						
Intercept	-0.00 (0.08)	-0.15 – 0.15	0.993	-0.00 (0.08)	-0.15 – 0.15	0.989
Program	0.42 (0.07)	0.28 – 0.56	0.000	0.51 (0.47)	-0.41 – 1.43	0.278
Female	-0.01 (0.05)	-0.11 – 0.08	0.771	-0.01 (0.05)	-0.11 – 0.08	0.770

(continued)

Age	0.00 (0.00)	-0.00 – 0.01	0.060	0.00 (0.00)	-0.00 – 0.01	0.060
Propensity	1.15 (0.16)	0.84 – 1.47	0.000	1.07 (0.47)	0.14 – 2.00	0.024
Adjusted R^2	0.08			0.08		
Sargon $\chi 2$						0.552
Hausman $\chi 2$						0.847

intensity DM program on the outcome variables. This information suggests that the IV regression models developed for the present study were not sensitive to the particular form of the endogenous variable included in the analysis.

The variables' coefficient estimates, standard errors, confidence intervals, and *p*-values in the block two models were very similar (and in some cases identical) to those in the block one models. In particular, no insignificant ($\alpha = 0.05$) variables in the block one IV models were significant in the second block IV models. The 2SLS IV regression models were screened for multicollinearity by calculating variance inflation factors (VIFs). The VIFs for the predicted value of program participation from the first stage regression and the propensity score variables were 10.3 in all regressions. Based on this information, substantial multicollinearity was present in the models due to these variables. Thus, the usefulness of the second block IV regression models appears questionable.

Block Three OLS and IV Regression Models

Table 8 contains the results of the block three OLS and IV regression models. In these models, actual program participation was used as the endogenous variable; however, the quintile propensity score was used in place of the continuous propensity score variable. To test the assumption that the instruments were not directly related to the outcomes,

Table 8

Block Three: OLS and IV Regression Models Using Actual Program Participation as the Endogenous Variable and the Propensity Score Quintile Variable

	OLS Regression			*IV Regression*		
Variable	*Coefficient (SE)*	*95% CI*	*p-value*	*Coefficient (SE)*	*95% CI*	*p-value*
Outcome: Diabetes-Related Costs (Log)						
Intercept	2.88 (0.24)	2.40 – 3.35	0.000	3.45 (0.33)	2.80 – 4.10	0.000
Program	1.87 (0.20)	1.48 – 2.25	0.000	8.15 (1.33)	5.55 – 10.75	0.000
Female	-0.06 (0.14)	-0.34 – 0.22	0.658	-0.06 (0.18)	-0.41 – 0.30	0.751
Age	0.04 (0.00)	0.03 – 0.05	0.000	0.04 (0.01)	0.03 – 0.05	0.000
Propensity	0.25 (0.05)	0.15 – 0.34	0.000	-0.27 (0.12)	-0.52 - -0.03	0.028
Adjusted R^2	0.15			0.08		
Sargon $\chi 2$						0.222
Hausman $\chi 2$						0.000
Outcome: Emergency Department Visits (Square Root)						
Intercept	1.01 (0.09)	0.83 – 1.19	0.000	1.06 (0.10)	0.86 – 1.26	0.000
Program	0.27 (0.07)	0.12 – 0.42	0.000	0.85 (0.40)	0.06 – 1.64	0.035
Female	0.16 (0.05)	0.05 – 0.27	0.003	0.16 (0.06)	0.05 – 0.27	0.004
Age	-0.01 (0.00)	-0.015 - -0.009	0.000	-0.01 (0.00)	-0.015 - -.008	0.000
Propensity	0.05 (0.02)	0.01 – 0.09	0.008	0.00 (0.04)	-0.07 – 0.08	0.964
Adjusted R^2	0.05			0.04		
Sargon $\chi 2$						0.000
Hausman $\chi 2$						0.136
Outcome: Hospital Days (Square Root)						
Intercept	-0.05 (0.09)	-0.22 – 0.11	0.540	0.08 (0.10)	-0.12 – 0.29	0.415
Program	0.54 (0.07)	0.40 – 0.67	0.000	2.04 (0.41)	1.23 – 2.85	0.000
Female	-0.01 (0.05)	-0.11 – 0.08	0.795	-0.01 (0.06)	-0.12 –0.10	0.837

(continued)

Age	0.00 (0.00)	-0.000 – 0.006	0.073	0.00 (0.00)	-0.000 – 0.006	0.062
Propensity	0.07 (0.02)	0.03 – 0.10	0.000	-0.06 (0.04)	-0.13 – 0.02	0.138
Adjusted R^2	0.06			0.04		
Sargon $\chi 2$						0.706
Hausman $\chi 2$						0.000

residuals from each IV model were regressed on the instruments. The model *F*-tests indicated that no direct relationships existed between the instruments and the residuals for the cost ($F = 1.12, p = 0.324$) and hospital days ($F = 0.74, p = 0.768$) models. However, the *F*-test for the ED visits model indicated that a direct relationship existed between the instruments and the outcome ($F = 2.56, p = 0.000$). This finding provides additional evidence that three-digit zip codes are not good instruments to use in the ED visits model.

As with the previous blocks, the Sargon test indicated that the instruments were not related to the disturbance terms in the cost ($p = 0.222$) and hospital days ($p = 0.706$) models, but that at least some of the instruments were related to the disturbance term in the ED visits model ($p = 0.000$). However, the Hausman test results in the block three models were different from the test results in the previous blocks. In particular, the Hausman test indicated that significant differences existed between the OLS and IV models for the cost ($p = 0.000$) and hospital days ($p = 0.000$) outcomes. This information may indicate that the OLS cost and hospital days models were inconsistent due to the endogeneity of the program participation variable.

To determine if the block three models were influenced by multicollinearity, VIFs were calculated for the 2SLS IV regression models. All VIFs were well below the 10.0

threshold suggesting that multicollinearity was not present. These findings indicate that the IV models developed for this study were sensitive to the quantitative form of the propensity score variable. This is probably due to the fact that multicollinearity apparently results when propensity scores and the estimated program probability variable are included as covariates in the second stage of the 2SLS IV regression models.

Unlike the previous two model blocks, the coefficients for the program participation variables were significant in all block three IV regression models. This result may be due to multicollinearity that was present in the IV regression models of the two previous blocks. The information in the block three cost models indicates that program participation in the OLS model was associated with a 187% increase on average in diabetes-related health care costs, while program participation was associated with an average increase of 815% in health care costs in the IV model. While the IV program estimate is substantially larger, the OLS estimate is closer to the estimate reported in the first two blocks. The confidence intervals for the block three OLS estimate includes the point estimate from the previous blocks; however, the IV confidence intervals are extremely large and do not include the IV estimates from those blocks.[58] The results for the program variable in the block three ED visits and hospital days models are similar to those in the cost model.

For the other covariates, use of the quintile propensity score did produce some changes in the coefficient standard errors and confidence intervals. However, it did not

[58] While the confidence interval for the ED visits IV model includes the OLS point estimate, the confidence intervals for the cost and hospital days IV models do not contain their respective OLS point estimates.

produce any changes in the statistical significance of the coefficients except for the propensity quintile variable in the hospital days model, which was nonsignificant (p = 0.138). In the previous blocks, the propensity score was significant (p = 0.025 and 0.024 for the block one and two models, respectively). The other covariates that were significant in the first two blocks remained significant in the third block.

The information from the block three models suggests that the high intensity DM program had a significant positive effect on diabetes-related costs, ED visits, and hospital days. The direction and magnitude of the program coefficients is surprising because they are all positive and considerably larger than the OLS coefficients. It would seem reasonable to expect the DM program to have a negative effect on the outcomes because the program is intended to achieve that result, and for the IV estimates to be smaller than the OLS estimates (and also for the IV estimates' confidence intervals to include the OLS estimates) if they are actually influenced by omitted variable bias. This finding may be problematic because the fact that the IV estimates in the block three models are considerably larger than the OLS estimates is not entirely consistent with omitted variable bias in OLS regression (Wooldridge, 2006).

Several reasons may exist that account for these results. First, the positive program effect may be a result of the high intensity DM participants having more advanced illnesses than the standard intensity recipients and thus requiring more intensive health care services. Second, the positive effect may be due to the fact that the individuals in the high intensity program were selected for this intervention because they have higher health care costs and utilizations. The regression models thus reflect the fact that the high intensity program by design is associated with increased health care costs and utilizations. Third, enough time has not elapsed for the high intensity program to reduce health care costs and utilizations. DM is intended to be a long-term process and it may take several years before the high intensity program produces a negative effect in the study outcomes (Afifi et al., 2007).

Fourth, the instrumental variables in the IV regressions are "weak," meaning that the partial correlation between the instruments and the endogenous variable is low. If this occurs, then the IV estimator will have large asymptotic bias, which may result in inconsistent regression estimates, standard errors, and confidence intervals, thus making model inferences very unreliable (Staiger & Stock, 1997; Wooldridge, 2006; Gelman & Hill, 2007). One method to test for weak instruments is to calculate the model *F*-statistic from the first stage IV regression. This statistic tests the hypothesis that the instruments do not enter into the first stage regression (Staiger & Stock, 1997). The rule of thumb for determining whether the instruments are weak is an *F*-statistic that is less than 10, which implies that the model lacks substantive predictive power (Stock, n.d.).

To test for weak instruments, first stage *F*-statistics were calculated for each block three IV regression model. The calculation yielded an *F*-statistic of 3.28 ($p = 0.000$) for each model.[59] Based on this information, it appears that the zip code instruments are weak

[59] *F*-statistics were also calculated for the block one and two first stage IV regression models. The *F*-statistic for the first block models was 326.83 ($p = 0.000$), while the *F*-statistic for the second block models was 2.02 ($p = 0.006$). The fact that the *F*-statistic for the block one models was so large is probably not that informative because the R^2 statistic for the regression of the predicted values from the first stage on the exogenous variables (excluding the instruments) was 0.90, which suggests a high level of multicollinearity (Wooldridge, 2006). A similar analysis performed for the block two models also yielded an R^2 statistic of 0.90. The multicollinearity is due to the fact that the predicted values are a linear function of the exogenous variables including the instruments.

in the present study and probably contribute to the IV estimates' large coefficients, standard errors, and confidence intervals.[60] As a result, the block three IV models are probably biased and should not be used to interpret the effects of the DM program even though the Hausman test indicated that the IV cost and hospital days models should be preferred over the OLS models. It should be noted that due to the large variances that IV estimates tend to have, some researchers prefer to interpret treatment effects using OLS estimates even if they are biased (Winship & Morgan, 1999). Thus, the researcher would probably have recommended interpreting the effects of the high intensity DM program using the OLS models developed in this study even if the instruments were not weak (assuming of course that the IV coefficients, standard errors, and confidence intervals were as large as the ones presented in Table 6).

Summary

The objective of this study was to test the feasibility of an instrumental variables (IV) regression procedure developed by Linden and Adams (2006) for evaluating disease management programs. Their procedure involved using three-digit recipient zip codes as instruments in two-stage least squares (2SLS) IV regression. As part of their procedure, they used propensity scores as a covariate to control for overt biases and the predicted

[60] Some observers may argue that this information contradicts the discussion on the association of X and Z presented on pages 74 and 75. That discussion differs from the above discussion because it focused on whether a relationship existed between the instruments and the endogenous variable. The above discussion is focused on the extent to which at least one of the instruments is useful for predicting program participation, given the exogenous variables. Moreover, Morgan and Winship (2007) report that researchers do not agree on how large the association between Z and X must be before IV regression can proceed. Generally, researchers agree that Z is too weak if the regression of X on Z is not statistically significant at a small level. However, if this regression is performed using a large dataset, a weak Z can still produce significant results.

value of program participation, from a separate logistic regression model, as the endogenous variable in their IV models.

The study found that propensity scores that appear to control for potential overt biases in ordinary least squares (OLS) regression can be calculated using claims data from the Virginia Medicaid Management Information System. To test the feasibility of Linden and Adams' (2006) IV regression procedure, three blocks of OLS and IV regression models were developed. The appropriateness of the IV models was determined, in part by comparing them to OLS models, which are generally viewed as being more efficient estimators.

The first block IV regression models used the predicted value of program participation from a separate logistic regression as the endogenous variable and the propensity score as an exogenous covariate. While the Sargon test revealed that the zip code instruments were probably not good instruments in the ED visits IV model because they were associated with the disturbance term, the Hausman test indicated that using three-digit zip codes as instruments did not appear to produce IV models that were significantly different than their OLS counterparts. The analysis performed for the present study suggested that zip codes may not be good instruments to use in IV models that are intended to estimate DM treatment effects. However, the results from the first block models may be questionable because the models were biased due to multicollinearity, which resulted from including both the predicted value of program participation and propensity scores as independent variables in the second stage of 2SLS IV regression.

In the second block models, actual program participation was used as the endogenous variable to test the sensitivity of the IV models to this variable's specific form. The analytical results were very similar to those reported in the first block models suggesting that the IV models were not sensitive to the endogenous variable because no significant differences were found between the OLS and IV models. However, these results are also questionable due to multicollinearity that resulted from including both the predicted value of program participation (calculated by regressing actual program participation on the instruments and exogenous variables in the first stage of 2SLS regression) and the propensity scores as covariates in the IV models.

Finally in the third block models, actual program participation was used as the endogenous variable; however, a quintile propensity score was used in place of the quantitative propensity score to test the sensitivity of the models to this variable's specific form. The analysis revealed that the models were sensitive to this variable because significant differences were found between the OLS and IV regression cost and hospital days models. These results suggest that when multicollinearity is eliminated, using three-digit zip codes as instruments resulted in IV regression models that were preferable to the OLS models presumably due to the endogeneity of the treatment variable. However, these IV models should probably still be rejected because an examination of the first stage IV *F*-statistics revealed that the instruments were weak, which probably produced biased coefficients, standard errors, and confidence intervals for the model covariates.

Chapter 5

Discussion and Conclusions

The present study examined the feasibility of an instrumental variables (IV) regression procedure developed by Linden and Adams (2006) for use in disease management (DM) program evaluations. In disease management, individuals are often assigned nonrandomly to the treatment and control groups (i.e., they self-select into the programs). Nonrandom assignment can complicate efforts to evaluate DM programs because the treatment (or participation) variable may be correlated with unobserved confounding variables. Unless evaluators can control for these variables by including them in regression models, DM program effect estimates may be biased. Linden and Adams (2006) argue that IV regression using patient three-digit zip codes as instruments can be used to derive unbiased DM treatment effect estimates by eliminating the correlation between the treatment and unobserved confounding variables.

IV regression can be difficult to employ, however, because researchers must find one or more instruments that are: 1) directly correlated with the treatment variable, but not the outcome variable, and 2) uncorrelated with the unobserved variables that influence treatment assignment (Wooldridge, 2006). Linden and Adams (2006) hypothesize that participant three-digit zip codes meet these two assumptions and are thus valid instruments for DM program participation. To test their hypothesis, they used IV regression to evaluate the effects of participation in an Oregon managed care diabetes DM program on three outcome variables: total diabetes costs, emergency department (ED) visits, and hospital days. They assessed the appropriateness of their IV regression models by comparing them against comparable ordinary least squares (OLS) regression models. Based on the analysis, Linden and Adams (2006) found that IV regression using zip code instruments only appeared to produce an unbiased treatment effect estimate on the hospital days outcome. While they noted that their procedure was successfully employed in one model, they reported that it might not generalize to other DM programs.

The present study sought to test the generalizability of Linden and Adams' (2006) IV regression procedure by using it to evaluate the effects of a high intensity Virginia Medicaid diabetes DM program. The study was guided by three research questions:

1. Which statistical method provides the best unbiased estimates of high intensity DM program participation on the outcome variables?
2. Do the parameter estimates and confidence intervals for the predictor variables differ depending upon which statistical method is used?
3. What are the advantages and disadvantages of using OLS and IV regression to evaluate high intensity DM program effectiveness?

The study hypothesis was that instrumental variables regression using three-digit zip code instruments will provide an unbiased estimate of the effects of disease management participation on a group of high intensity Virginia Medicaid diabetes patients enrolled in the *Healthy Returns*^{SM} DM Program. The remainder of this chapter focuses on answering the research questions and discussing the study limitations, implications of the study findings for the medical and health policy communities, future research possibilities, and important study conclusions.

Research Question Analysis

Analytical findings for the three research questions are presented in this section, which begins with a discussion of the best statistical method to use for evaluating the effects of high intensity participation in the Virginia *Healthy Returns*^{SM} DM program. An examination of the ordinary least squares (OLS) and instrumental variables (IV) regression coefficients and confidence intervals is then presented followed by a discussion of the advantages and disadvantages of using OLS and IV regression in DM evaluations.

Best Statistical Method for Evaluating the High Intensity DM Program

The first research question focused on determining whether OLS or IV regression using three-digit zip code instruments produced the best unbiased estimates of high intensity DM program participation on the study outcomes. To address this research question, three blocks of OLS and IV regression models were generated. The first block was performed using the procedure employed by Linden and Adams (2006) where the estimated probability of program participation from a separate logistic regression model was used as the endogenous variable and age, gender, and a quantitative propensity score variable were used as covariates. The second block was performed using actual program participation as the endogenous variable and age, gender, and a quantitative propensity score variable as covariates. Finally, the third block was performed using actual program participation as the endogenous variable and age, gender, and a quintile propensity score variable as covariates.

To test the IV regression assumption that Z is related to X, a series of simple regressions were performed to determine if the instruments were related to both the predicted value of program participation and the actual program participation variables. Residuals from each IV regression model were also regressed on the outcome variables to test the assumption that the IV estimates were not directly related to the outcomes. The Sargon chi-square test was used to test the assumption that at least some of the zip code instruments were uncorrelated with the disturbance terms (i.e., instrument exogeniety) in the IV models. The Hausman specification chi-square test was used to determine if significant differences existed between the OLS and IV regression coefficients (i.e., treatment endogeniety). In addition, the IV regression models were checked for multicollinearity by calculating variance inflation factors (VIFs). Finally, the relevancy of the zip code instruments was tested by examining the first stage *F*-statistics from the IV regressions to determine if the instruments were weak (i.e., instrument relevancy).[61]

The following observations are made based on the above analyses:

1) The assumption that Z is related to X appeared to be met based on the results of the simple regressions that assessed this relationship. These results are comparable to results reported by Linden and Adams (2006) and suggest that at least some of the three-digit zip codes may be appropriate instruments for the high intensity DM program participation variable.
2) The assumption that no direct relationships exist between the instruments and the outcomes appeared to be fulfilled for the cost and hospital days IV

[61] The usefulness of instrumental variables regression depends on whether the instruments are valid. Invalid instruments can result in meaningless IV regression models (Stock & Watson, 2007). To ensure that the instruments are valid, Wooldridge (2006) and Stock and Watson (2007) recommend testing for the endogeniety of the problematic explanatory variable, the exogeniety of the instruments, and the relevancy of the instruments.

regression models, but not for the ED visits IV models. This suggests that three-digit zip codes may be appropriate instruments in the cost and hospital days models, but not in the ED visits models. Linden and Adams (2006) reported that no direct relationships existed between the zip code instruments and outcomes in their IV regression models.[62]

3) The assumption of instrument exogeniety was demonstrated in the cost and hospital days IV regression models, but not in the ED visits models. This information further suggests that three-digit zip codes may be appropriate instruments for use in the cost and hospital days models, but not in the ED visits models. Linden and Adams (2006) did not report checking for instrument exogeniety in their study.

4) The assumption of treatment endogeniety did not appear to be met in the first two regression blocks, but did appear to be met in the block three IV cost and hospital days models. This information indicates that the IV regression models did not contribute to the prediction of the outcome variables in the first two blocks, but may have contributed to the prediction of costs and hospital days in the third block. Linden and Adams (2006) reported that treatment endogeniety was only met for their hospital days IV model.

[62] Linden and Adams (2006) assessed this assumption by regressing the residuals from each IV model on the outcomes. However, Morgan and Winship (2007) report that this is actually "a strong and untestable assumption" (p. 196). They argue that Z and Y will always be related after conditioning on X because Z is a "collider" that is caused by both Z and U (or unobserved variables). Based on this information, the assumption that no direct relationships exist between the instruments and the outcomes may not have actually been achieved in the present study.

5) Multicollinearity was detected in the first two IV regression blocks, but not in the third IV regression block. Thus, the usefulness of the overall results from the first two blocks may be questionable. Linden and Adams (2006) reported that multicollinearity was not present in their regression models. Multicollinearity apparently occurred in the present study when the predicted value of program participation was combined with the quantitative propensity scores in the second stage of the IV regressions because these two variables essentially represent the same concept. As a result, the quantitative propensity score variable was replaced with a quintile variable in the block three regressions to eliminate multicollinearity from the models.
6) While statistical validity tests tended to favor the block three IV cost and hospital days regressions, the assumption of instrument relevancy was not met based on an examination of the first stage model *F*-statistics. First stage *F*-statistics provide a measure of the information content contained in the zip code instruments (or the extent to which the instruments explain variation in the treatment variable), given the other exogenous variables included in the IV model. Because weak instruments can bias IV regression estimates, the results of the block three models (as well as the other two blocks) may be unreliable (Stock & Watson, 2007). This finding could explain why the block three IV estimates appeared unstable compared to their OLS counterparts (Stock, n.d.).

Despite the fact that some of the above observations tended to favor the IV regressions (i.e., evidence was obtained suggesting that Z is associated with X, Z is not associated with Y, and X is not associated with U), the IV models are probably unstable due to multicollinearity and weak instrument bias.[63] Thus, OLS regression using a propensity score covariate appears preferable in this study for estimating the "unbiased" effects of high intensity participation in the Virginia *Healthy Returns*^{SM} DM Program. It should be noted that this finding does not suggest that IV regression using three-digit zip code instruments is inappropriate for estimating program effects in other DM evaluations because it only applies to the present study. Different analytical results might be obtained if this IV regression procedure is applied to other samples.

While the study hypothesis was not supported, an important analytical finding involves the use of propensity scores as a control variable. The present study demonstrated that propensity scores that control for some observable differences (i.e., overt biases) between the treatment and control groups can be calculated using data from the Virginia Medicaid Management Information System. When combined with OLS regression, propensity scores may offer a parsimonious means of deriving less biased estimates of the effects of the high intensity DM program on certain outcomes by controlling for many observable variables that may account for nonrandom assignment.[64] In fact, propensity scores combined with OLS regression are popular estimators in program evaluation because the scores can act as a control function by containing information relevant for estimating treatment effects (Wooldridge, 2002).

[63] "U" refers to unobserved variables that are related to X and Y.

[64] There are restrictions on the number of covariates that can be included in regression models. A general rule is that there should be one predictor for every 20 subjects. Including all available covariates (kitchen-sink regression) may produce instability in the models, incorrect results, and decreased statistical precision (Newgard et al., 2004).

However, there are some limitations involved with using propensity scores as a regression covariate. First, propensity scores can only control for observed covariates that are available in a dataset. They will not control for unobserved confounders except to the extent that they are correlated with the measured variables used to estimate the propensity scores (Stukel et al., 2007). Researchers using propensity scores as a regression covariate should therefore note that treatment effect estimates may still be biased due to unobserved confounders.

Second, using propensity scores as a covariate may increase bias because regression adjustment imposes a linearity constraint that may be unrealistic when modeling treatment effects on some outcomes. While nonlinear terms can be added to regression models, it may be difficult for researchers to know how nonlinear propensity scores should be approximated. Failure to use the correct nonlinear approximations for variables may bias treatment effect estimates due to regression model misspecification (Winship & Morgan, 1999; Shadish et al., 2002).

Third, the ability of propensity scores to control for observed differences in regression models may depend on their specific form. In particular, quantitative propensity scores may offer a better means of controlling for preexisting differences than quintile propensity scores because they can account for more variability in the observed covariates than quintile scores. Information presented in Table 3 in Chapter 4 demonstrated that quantitative propensity scores could control for significant differences between study groups on the nonwhite race, western region, hospital days, ED visits, and diabetes-related cost variables. However, when the same regressions were performed using the quintile propensity scores, the regressions revealed that the quintile scores failed to control for significant differences between the study groups on the hospital days and cost variables (Appendix B). While the quintile propensity score variable may have eliminated multicollinearity from the block three IV models, the ability of this particular variable to control for observable differences may be limited.

Despite the above limitations, however, propensity scores can reduce bias in estimating treatment effects in nonrandomized observational studies. When combined with regression, this method may allow for stronger causal inferences in DM evaluations by reducing bias due to nonrandom assignment. This in turn may help to prevent inappropriate conclusions due to analyses performed on observational data that fail to correct for selection bias (Newgard et al., 2004).

OLS and IV Regression Coefficients and Confidence Intervals

The second research question focused on determining if the OLS and IV regression estimates differed. In general, OLS and IV estimates will differ because IV regression produces a larger asymptotic variance, which can increase the magnitude of the coefficients, standard errors, and confidence intervals (Winship & Morgan, 1999; Hadley & Cunningham, 2004; Wooldridge, 2006). When supported by the results of statistical tests that assess the validity of the IV models, differences between the OLS and IV coefficients may suggest that the OLS estimates are inconsistent due to omitted variable bias and should be rejected in favor of the IV estimates (Hadley et al., 2003; Linden & Adams, 2006; Wooldridge, 2006).[65]

To address the second research question, the researcher compared the OLS and IV coefficients and confidence intervals for the variables in the three regression blocks. The comparisons revealed that the block one, two, and three OLS and IV coefficients, standard errors, and confidence intervals for the age and gender variables remained relatively unchanged, while the program and propensity score OLS and IV estimates differed. This finding could suggest that the age and gender estimates were not correlated with the error term because their OLS and IV estimates were comparable in all regression blocks, while the program and propensity score variables were correlated with the error term in the OLS regressions (at least in the block three IV cost and hospital days models) (Lindrooth, Hoerger, & Norton, 2000; Long et al., 2005; Linden & Adams, 2006).[66,67] However, examining the differences between the OLS and IV estimates for the independent variables may not be very meaningful in the present study due to the multicollinearity and weak instrument bias that were encountered in the regression models.

[65] Hadely et al. (2003) further argue that researchers should determine the appropriateness of IV estimates based on the availability of important variables that could be used as covariates in OLS regression and the extent to which the OLS and IV estimates are similar.

[66] This finding seems feasible because age and gender were exogenous variables in the IV regressions. It is also not very interesting because treatment effect estimates, rather than covariate estimates, are usually the focus of program evaluation studies (Mohr, 1995). In addition, the observation is not entirely unexpected because similar results were observed in other studies reviewed by the researcher (Hadley & Cunningham, 2004; Linden & Adams, 2006).

[67] The propensity score variables developed for the present study may actually be endogenous because it is doubtful that they contain information on all observable variables that account for nonrandom assignment into the high intensity and standard intensity programs. Important excluded variables would thus be subsumed under the error term.

While IV regression can be used to overcome selection bias due to omitted variables, it does not always offer a good solution to this issue because IV coefficient estimates can be unstable due to large variances (Winship & Morgan, 1999). For this reason, Hadley et al. (2003) recommend that researchers exercise some discretion in accepting IV estimates based on statistical validity tests due to possible coefficient instability. They further recommend that researchers who use IV regression to estimate treatment effects in observational studies consider reporting both OLS and IV coefficient estimates to establish a range of possible treatment effects on the outcomes.

IV estimates are usually larger than OLS estimates because IV regression only uses a portion of the covariation in the treatment and outcome variables in the calculation process. Using only a portion of the covariation represents a direct loss of statistical power that can induce the IV estimators to exhibit increased sampling variance compared to that exhibited by similar OLS estimators. Thus, an unbiased IV estimator can actually be unstable compared to a biased and inconsistent OLS estimator. This issue can be especially troubling if the instruments are weak (Morgan & Winship, 2007).

To better understand how IV regression can increase coefficient magnitudes and how weak instruments can bias IV estimates by inflating coefficient variances, the calculations used to produce OLS and IV estimators are briefly reviewed. For simplicity, the calculations are limited to bivariate regressions (and one instrument for the IV estimator) where the outcome (Y) is regressed on the treatment variable (X) (the logic behind the calculations also applies to multiple regression). The OLS estimator for β_1 is calculated as:

$$[5] \qquad \text{ß-hat}_{OLS} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{S_{XY}}{S_X^2}$$

This equation simply means that the OLS estimator equals the sample covariance between the treatment (X) and outcome (Y) variables divided by the sample variance of the treatment variable. The OLS estimator can be derived as long as the denominator is greater than zero (i.e., meaning that there is variation in X). Equation 5 shows that the OLS estimator has more statistical power, and is therefore more efficient, than the IV estimator because it uses all of the covariation in the treatment and outcome variables in the calculation process (Wooldridge, 2006). As a result, OLS estimates will tend to be smaller and more efficient than IV estimates.

However, the IV estimator for β_1 is calculated as:

$$[6] \qquad \text{ß-hat}_{IV} = \frac{\sum_{i=1}^{n}(Z_i - \overline{Z})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(Z_i - \overline{Z})(X_i - \overline{X})} = \frac{S_{ZY}}{S_{ZX}}$$

This equation demonstrates that the IV estimator will often be larger than the OLS estimator because it is calculated as the covariance of the instrument (Z) and the outcome (Y) divided by the covariance of the instrument (Z) and the treatment (X) variables. The instrument and treatment variables must be correlated in order for this equation to be solved. Equation 6 indicates that IV regression only uses a portion of the covariation in the treatment and outcome variables in the calculation process because the numerator equals the effect of the instrumental variable on the outcome, while the denominator represents the effect of the instrumental variable on the treatment variable. If the instrument is dichotomous (as in the present study), then the numerator equals the mean outcome difference between Z = 0 and Z = 1 and the denominator equals the mean effect of the treatment (Wooldridge, 2006; Martens, Pestman, de Boer, Belitser, & Klungel, 2006).[68]

As previously mentioned, the IV coefficients, standard errors, and confidence intervals can be very unstable if Z is weak. A weak instrument will bias the IV estimator because the denominator in equation 6 will be small, which makes the estimator sensitive to small changes.[69] Because the covariance in the denominator behaves as a multiplier, a small correlation will produce a large standard error and confidence interval, which can make the coefficient estimate less reliable (Martens et al., 2006). Because the relationship between the instrument and treatment variable is independent of sample size, a weak instrument can produce an IV coefficient that contains no genuine information about the causal effect of the treatment on the outcome (Morgan & Winship, 2007). The results of the first stage IV *F*-statistics and second stage IV coefficients, standard errors, and confidence intervals for the three regression blocks suggest that this process may have occurred in the present study.[70]

Before concluding this discussion, it is worth mentioning one additional issue that may have contributed to the weak instrument bias encountered in the present study. According to Linden and Adams (2006), the zip code instrument they proposed using is

[68] The source for the equations is Wooldridge (2006).

[69] In fact, Morgan & Winship (2007) report that weak instrument bias can "explode" the IV estimator due to the small denominator.

[70] This comment refers to the role that sample size plays in OLS regression. A larger sample can produce a better OLS estimator (Stock & Watson, 2007).

based, in part, on a natural experiment that assumes the existence of a high enrollment rate zip code and a low enrollment rate zip code. The natural experiment that these geographic boundaries model is how differences in zip codes are related to differences in DM participation rates. Linden and Adams (2006) assume that because three-digit zip codes mimic a natural experiment, geographic proximity will make DM participants and non-participants more similar on unmeasured confounders. They are not alone in their use of geographic boundaries as instrumental variables. In fact, the use of naturally occurring instruments has gained considerable popularity among many researchers who view them as "gifts of nature" (Morgan & Winship, 2007).

However, not all researchers are convinced that naturally occurring boundaries make good instruments. In particular, some argue that the variation supposedly induced in the endogenous predictor by naturally occurring instruments is often either poorly explained and/or not supported by relevant theories. Moreover, random variation created by natural experiments does not necessarily ensure that the instrument has no direct effect on the outcomes. As a result, instruments from natural experiments can be "black boxes" that reduce their ability to produce unbiased treatment effect estimates needed for public policy development and guidance (Morgan & Winship, 2007). These concerns may explain why three-digit zip codes did not ultimately appear to be appropriate instruments for the high intensity DM participation variable used in the present study.

Advantages and Disadvantages of Using OLS and IV Regression in DM Evaluations

The third research question concerned the advantages and disadvantages of using OLS and IV regression to estimate treatment effects in DM evaluations. This research question was addressed through an assessment of what the researcher learned about these methods during the present study. Because both methods use least squares to minimize the sum of squared residuals, they have similar advantages and disadvantages. Nevertheless, researchers planning to estimate the effects of DM interventions when selection bias is present should still consider the advantages and disadvantages of both methods as well as their own strengths and capabilities (i.e., knowledge of regression methods, computer capabilities, access to data, etc.) before deciding upon which method (if either) to use.

An advantage of using OLS regression in DM evaluation is that OLS is a widely recognized data analysis technique that is used by researchers in a variety of social science disciplines to evaluate alternative explanations for study outcomes (Morgan & Winship, 2007). Because OLS regression allows researchers to control for (or fix) many correlated factors that may simultaneously affect the outcome variable, they can use it to infer causality when analyzing observational data (Wooldridge, 2006). In other words, OLS regression can allow researchers using nonexperimental data to do what natural scientists using experimental data can do: hold other factors fixed (Wooldridge, 2006). OLS regression has thus become the common language that many researchers "speak" when discussing the effects of social interventions on outcome variables (Stock & Watson, 2007). Using OLS regression can therefore allow DM evaluators to communicate their findings to a wide audience of social scientists.

Another advantage of using OLS regression in DM evaluation is its computational ease. All major statistical software programs include regression procedures that allow for estimating linear and nonlinear relationships between both continuous and categorical variables in cross sectional and longitudinal datasets (Morgan & Winship, 2007). Because these software programs can operate on personal computers, it should be very convenient for many evaluators to use OLS regression to estimate DM program participation effects (Kennedy, 2003). However, the evaluators must understand how to use and interpret OLS regression in order for the results to be meaningful. Failure to understand the mechanics of OLS regression can result in some very impressive looking findings that are essentially meaningless.

Other advantages of using OLS regression in DM evaluations include its theoretical properties that make it an unbiased and consistent estimator (Stock & Watson, 2007). For instance, OLS is considered to be an unbiased estimator under the Gauss-Markov theorem when certain statistical assumptions are met such as linearity, random sampling, no perfect collinearity, zero conditional mean, and homoscedasticity. Moreover, when these assumptions are met, OLS is viewed as the best linear unbiased estimator (BLUE) because it has the smallest variance conditional on the values of the independent variables (Wooldridge, 2006). Because it often produces the smallest variance compared to other estimators, many social scientists view OLS regression as an efficient estimator. When these characteristics are subsequently met, OLS can be an appropriate estimator to use in DM evaluations.

However, using OLS regression may have certain disadvantages when the statistical assumptions are not met. For instance, the validity of regression inferences depends partly on meeting the assumption of linearity between the outcome and independent variables. This assumption can be violated when the outcome is a dichotomous variable where y = 1 denotes one possible outcome and y = 0 denotes the other possible outcome (i.e., patient received adequate care = 1, patient did not receive adequate care = 0). While OLS regression can be used to model binary outcomes, it can result in some problematic results (i.e., probability estimates that are above 1 or below 0, nonnormal errors, and unequal variances).[71] As a result, a probit or logistic estimator may be more appropriate to use when the study outcomes are dichotomous because these estimators force the predicted values of the outcomes to be between 0 and 1 (Cohen et al., 2003; Mendenhall & Sincich, 2003; Wooldridge, 2006; Stock & Watson, 2007).

Regression inferences can also be invalid if the homoskedasticity assumption is violated which can occur when the study outcome is a count variable, such as the number of participant hospitalizations or ED visits in a calendar year. While there are certain variance-stabilizing transformations that can be applied to satisfy the homoskedasticity assumption for OLS regression use, other estimators may be more appropriate. In particular, the Poisson regression estimator or the negative binominal regression estimator may be preferable when the outcome is a count variable. While OLS regression can be used to model DM effects on count outcomes, failure to use a more appropriate estimator may result in biased results (Mendenhall & Sincich, 2003; Wooldridge, 2006; Stock & Watson, 2007).

In addition, the validity of regression inferences can be questionable if the zero conditional mean assumption is violated, which is a key assumption that must be met in

[71] An OLS regression model is referred to as a linear probability model when the dependent variable is binary because y can only assume two values 0 and 1. Because y-hat represents the predicted probability of success in a probability model, the probability estimates must be between 0 and 1 (Stock & Watson, 2007).

order for OLS regression to provide unbiased results. The zero conditional mean assumption assumes that all important variables that are related to the outcome and the DM treatment variable are included in the model (Wooldridge, 2006). Failure to meet this assumption will result in omitted variable bias. For example, by failing to include factors that determine high intensity DM participation such as personal motivation, the OLS estimator of the slope in the regression of diabetes-related costs on high intensity program participation (while controlling for age, gender, and propensity score) could be biased. In other words, the mean of the sampling distribution of the OLS estimator may not equal the true effect on diabetes-related costs of a unit change in high intensity participation (i.e., moving from participant to nonparticipant) (Stock & Watson, 2007). Thus, the estimated coefficient for the treatment variable cannot be interpreted as the effect of the high intensity program because it also captures part of the effect of the omitted variables (Ettner, 2004).

Because most DM programs allow subjects to self-select into the treatments, it is very likely that OLS program effect estimates will be biased (or incorrect) due to omitted variables bias (Linden & Adams, 2006). Therefore, the primary advantage of using instrumental variables regression in DM evaluations is that it can offer a solution to omitted variables bias. IV regression can produce an unbiased effect estimate when the treatment variable (X) is correlated with the disturbance term (U) by dividing X into two variation parts. The first variation part is correlated with U (the unknown determinates of program participation), while the second variation part is uncorrelated with U. By using one or more instrumental variables that isolate the second part, IV regression focuses on variation in X that is uncorrelated with U and disregards variation in X that is correlated with U. IV regression can therefore derive unbiased treatment effect estimates by isolating variation in X that is uncorrelated with U (Stock & Watson, 2007).

Other advantages of IV regression include ease of interpretation due to its similarity with OLS regression, ease of calculation due to its inclusion in most statistical software programs, and its popularity among certain groups of social scientists such as economists (Linden & Adams, 2006; Wooldridge, 2006).[72]

The disadvantages of using IV regression in DM evaluation include the fact that researchers must find one or more variables (Z) that are correlated with the program participation variable (X), but not with the study outcomes (Y) or the unobserved variables (U) that influence both X and Y. Moreover, the instrumental variables must be strongly correlated with X in order for the regression coefficients to be reliable. Failure to use instruments that explain most of the variation in X may seriously bias the IV estimator by producing large standard errors and confidence intervals that do not contain the true value of the coefficients (Stock & Watson, 2007). Finding suitable variables that meet these often contradictory requirements can be very difficult in DM evaluation because evaluators will typically only have access to administrative claims and enrollment data, which can limit the number of suitable variables that can be used as instruments (Wooldridge, 2002; Linden & Adams, 2006).

[72] According to Wooldridge (2006), IV regression is second only in popularity to OLS regression in applied econometrics.

Other disadvantages of using IV regression in DM evaluation include the fact that the method still needs to meet certain theoretical OLS assumptions, such as no multicollinearity. Meeting this assumption can be especially difficult in two-stage least squares (2SLS) IV regression because the predicted value of X from the first stage is a linear combination of both the instruments and the exogenous variables that function as covariates in the second stage. However, failure to meet this assumption can bias IV regression results due to large coefficient standard errors and confidence intervals (Wooldridge, 2006).

Finally, another disadvantage of using IV regression in DM evaluation results from the fact that IV estimates are only based on a portion of the covariation in the causal and outcome variables. Using only a portion of the information in the data can result in a direct loss of statistical power thus resulting in IV estimators that exhibit more sampling variance than other estimators. As a result, a consistent and asymptotically unbiased IV estimator may be outperformed by a biased and inconsistent OLS estimator (Morgan & Winship, 2007).

Implication of Study Findings for the Medical and Policy Communities

The objective of the present study was to assess the feasibility of an instrumental variables (IV) regression procedure for disease management (DM) evaluations rather than to address clinically oriented issues concerning the quality and outcomes of DM services for specific patients or policy oriented issues about the overall effects of high intensity diabetes DM services on health care costs and utilizations. The study therefore focused on issues that may be of interest to quantitatively oriented health services researchers rather than to clinicians (i.e., practicing physicians and other health care providers) or health policy makers. While it may appear that both clinical and policy oriented inferences could be made based on the study results, caution should be exercised if attempting to use them for such purposes. For instance, some observers may conclude that the results suggest that the high intensity DM participants did not receive appropriate care or even that high intensity DM services are not effective because program participation was associated with significant increases in diabetes-related costs, ED visits, and hospital days. However, DM is a long-term process (Afifi et al., 2007) and the study results only represent subjects' one year experience in a high intensity diabetes DM program. As a result, they do not reflect changes that may occur in outcomes over time due to the program. Panel (or longitudinal) data studies are probably more useful for program evaluations that seek to assess the overall effects of interventions because they can account for changes that occur overtime for the same group of subjects while holding constant unobserved factors that may affect study outcomes (Wooldridge, 2006).

In any case, drawing inferences about the effects of the Virginia *Healthy Returns*^{SM} DM Program based on the results of this study are probably not appropriate because it has only been in operation for two years. Therefore, policy makers should probably consider the results of a panel data study before making any summative decisions about the *Healthy Returns*^{SM} program or high intensity DM services.

The above discussion does not suggest that examining patient quality of care and other clinical and policy oriented issues related to DM services is not important. In fact, studying these issues is very important because chronic diseases, such as diabetes, asthma, hypertension, and coronary heart disease, affect approximately one fourth of all Americans and account for almost 50 percent of all health care costs (Morisky et al., 2008). Thus, efforts to improve the quality of care that chronically ill patients receive have been given high priority and research needs to be directed toward determining if these interventions result in both decreased utilization of high cost health care services and improved patient health behaviors (Ofman et al., 2004).[73]

Addressing clinically oriented issues that may interest health care providers about the effects of DM services on specific patients is probably beyond the immediate scope of observational studies that employ IV regression methods to estimate treatment effects. Instead, clinical trials (i.e., experimental studies) are more appropriate for answering questions that are of immediate interest to physicians and other health care providers. In well-designed clinical trials, all subjects are randomized to the study groups in order to allow researchers to derive estimates of the average treatment effects for the entire population of subjects eligible for the trials. Assuming that a particular patient is a random draw from one of those populations, a physician could use the effect estimate from the trial to determine what the expected outcome of the treatment would be for his or her patient. In other words, clinical trials allow physicians to use the expected outcome of the treatment to make decisions about specific patients (Newhouse & McClellan, 1998).

However, IV methods are more appropriate for addressing policy issues that involve incremental decisions about health care services (i.e., whether additional chronic

[73] In fact, research has demonstrated that disease management can result in improved health status, improved health behaviors, and decreased health service utilizations for chronically ill patients (Lorig et al., 1999; Afifi et al., 2007; Morisky et al., 2008).

diseases should be included in a DM program or even if the DM program should continue to operate) than decisions about specific patient treatments. In particular, effect estimates that are derived using IV regression do not usually represent the average effect of an intervention in the entire population because they only apply to the subpopulation of patients who actually participated in the treatment. Specifically, IV estimates do not represent the average effect on a random patient from the entire population, but rather the average effect on an individual from the subpopulation that participated in the treatment. This marginal population is often not identifiable by physicians because it may not be immediately obvious to them that a particular patient is from that subpopulation. As a result, IV estimates may be only indirectly applicable to physicians (Newhouse and McClellan, 1998).

This does not mean, however, that physicians cannot use IV estimates for clinical decisions. For instance, physicians may find IV estimates to be more relevant than clinical trial estimates if their patients differ from the populations included in the trials. In that case, the IV estimates may provide better insights into the likely effects of a particular intervention than clinical trial results (Newhouse & McClellan, 1998).

Study Limitations

This study had several limitations. First, it was limited to diabetes recipients who were continuously enrolled in the DM program during CY 2007. The study results are thus not applicable to DM participants who have one of the other conditions covered under the Virginia *Healthy Returns*^{SM} DM program. The results may also not be applicable to diabetes program participants who enrolled after December 31, 2007 because they may be different from the study participants. The study is further limited because it only included fee-for-service Virginia Medicaid recipients. As a result, the findings are not generalizable to Virginia Medicaid recipients who receive diabetes DM services through one of the State's managed care organizations.

Second, the study findings may be limited due to how the propensity score variable was calculated. Propensity scores can control for overt bias if they are estimated using important predictors of group membership. The propensity score method assumes that all confounding variables that predict treatment assignment and are correlated with the outcome are included in the logistic regression model (Shadish et al., 2002). The predictors used to estimate the propensity scores in this study were selected for accessibility rather than for the role they played in identifying DM recipients for the high intensity intervention. It is very likely that important observable variables exist in the Virginia Medicaid Management Information System that should have been included in the propensity score model. However, the researcher was unable to include these variables for logistical reasons (i.e., computer programming capabilities and time and effort needed to develop the variables). As a result, the propensity scores used in the study are probably biased due to omitted variables, which could influence the results of the OLS and IV regression models developed for this study (D'Agostino & Kwan, 1995; Yanovitzky et al., 2005; Baser, 2006).[74]

[74] The *Healthy Returns*^{SM} DM contractor identifies patients as either high intensity or standard intensity based on a predictive modeling analysis of Medicaid claims data. Because the model is proprietary, the researcher is not aware of which specific variables are used to identify DM patients.

Third, the present study is limited due to the specific instrumental variables procedure that was employed. The study only considered the feasibility of using patient three-digit zip codes as instruments in a cross sectional analysis. The analytical results suggested that three-digit zip codes were not appropriate instruments to use in such an analysis. However, the IV estimation procedure can be used in time series and panel data regression methods (Wooldridge, 2006). If three-digit zip codes are used as instruments in an IV time series or panel data analysis, different results may be obtained. In addition, the study is limited because only one instrument was considered. Other variables may exist in the Virginia Medicaid Management Information System that could prove to be more appropriate instruments than zip codes. However, identifying these instruments was outside the scope of the present study.

Fourth, the study is limited because it focused on one statistical method for adjusting for hidden selection bias in DM program evaluations, which did not prove beneficial. Other methods exist for estimating program effects in observational studies that may be more appropriate to use for evaluating the *Healthy Returns*^{SM} DM program. Examples include the Heckman two-step method where a multiple regression model is estimated for an outcome variable concurrently with a selection model that compares program participants to nonparticipants on selected variables or fixed effects panel regression models that adjust for fixed unobserved characteristics that may be associated with selection into the treatment group (Schneider et al., 2007). These methods have been used in similar studies, but were not considered in the present study.

Future Research Possibilities

Several future research possibilities were identified during the present study. While the study results suggested that participant three-digit zip codes were not useful instruments in IV regression, the results are nevertheless limited to this particular study. Different results may be obtained if the procedure is used on different samples or in different ways. As a result, additional research could be performed using the IV regression procedure in a cross sectional analysis of a different sample of Virginia *Healthy Returns*^{SM} DM diabetes participants or on samples composed of participants with one of the other chronic conditions that are covered under the program, such as congestive heart failure, asthma, or coronary artery disease. Moreover, the IV regression procedure could be employed in a cross sectional analysis of one or more samples of Virginia Medicaid managed care DM participants.

If the results from these analyses support the results of the present study, then enough evidence may exist to conclude that three-digit zip codes are not appropriate instruments to use in DM evaluations. Researchers interested in using IV regression to estimate the effects of DM program participation should therefore focus attention on identifying other variables that can serve as instruments.

Another potential research area involves using patient three-digit zip codes as instruments in an IV time series (i.e., data collected over time on one or more variables) or panel data (i.e., data constructed from repeated cross sections over time on a set of subjects) data evaluation of a DM program. In fact, evaluating a DM program using time series or panel data may be more insightful than using cross sectional data because it will allow evaluators to assess how the average effects of DM participation on certain outcomes change over time. Moreover, these methods are suitable for DM evaluations of Medicaid programs (and maybe other DM programs) because recipient level claims data are typically compiled on both a monthly and yearly basis. IV regression using three-digit zip code instruments could therefore be used to estimate the effects of DM participation over time (Wooldridge, 2006).[75]

Another area of research could involve using a generalized method of moments (GMM) IV estimator instead of a two-stage least squares (2SLS) IV estimator to test the feasibility of using three-digit zip codes as instruments for DM participation in models where the outcomes are count variables. While two-stage least squares regression is probably the most commonly used IV method for both continuous and non-continuous outcomes, other methods exist that may be more feasible for count outcomes such the GMM IV method. The GMM IV method may be a better estimator than 2SLS regression because it can specifically account for a count outcome's Poisson distribution. By accounting for this particular distribution, results that are different from those derived in the present study may be obtained because the regression estimates would probably assume more plausible values (Johnston, Gustafson, Levy, & Grootendorst, 2008).[76]

Another area for future research involves recalculating the propensity scores developed in this study by including all observable variables that are used to assign

[75] In this case, the analysis could probably focus on estimating the effect of another month of DM participation on outcomes such as health care costs, hospital days, or emergency department visits (Wooldridge, 2006).

Virginia Medicaid recipients to either the high intensity or standard intensity programs in the logistic regression model that estimates the propensity scores. The propensity score developed for the present study is most likely biased because it does not include all variables that are used in the program assignment process. The ability of the propensity score variable to control for overt selection bias could therefore be improved by reestimating it using these variables and then reperforming the OLS regressions to derive less biased effects of high intensity DM participation.

In addition, a potential research area may involve testing the feasibility of the self-care education that participants receive when they are enrolled in the *Healthy Returns*^{SM} Program and other DM programs. Self-care management emphasizes the central role that patients play in managing their chronic illnesses. Appropriate self-care of chronic conditions is an important element of DM because it can result in improved health status and reduced health care utilizations for participants (Lorig et al., 2001). Because self-efficacy plays a role in the ability of individuals to manage their chronic conditions over the long-term, such a study could be guided by self-efficacy theory. Self-efficacy is important in DM because it involves the confidence that individuals have to perform behaviors that are needed to successfully manage their chronic conditions. Variables that could be used in such a study include health education components, health behaviors and status, and health care utilizations and costs (Lorig et al., 2001; Marks et al., 2005).

[76] In other words, the count outcomes would not have to be transformed to meet the heteroskedasticity assumption and the regression coefficients, standard errors and confidence intervals between the IV and OLS models may be more comparable.

Finally, another potential research area may involve examining the specific demographic characteristics of Virginia's three-digit zip code areas in order to better understand what these areas represent and how they may produce variation in the treatment variable that is not associated with variation due to the unobserved determinates of program participation. While the usefulness of this particular analysis may seem questionable due to the results of the present study, other research may reveal that three-digit zip codes are feasible instruments to use in IV regression. If this occurs, then the characteristics of the three-digit zip codes will need to be studied in order for researchers to better understand how the zip codes induce variation in the DM participation variable. As part of this analysis, the researchers could also determine which three-digit zip codes are weak so they can discard them in order to use the most relevant zip codes as instruments in their IV regression models (Stock & Watson, 2007). Failure to fully understand and explain this variation may lessen the ability of the instruments to provide informative information about the effects of the Virginia *Healthy Returns*^{SM} DM Program.

Conclusions

Three important conclusions emerged from the present study. First, IV regression using participant three-digit zip codes as instruments did not prove to be an effective means of estimating the effects of high intensity participation in the *Healthy Returns*^{SM} DM Program. Second, propensity scores developed using administrative claims and enrollment data from the Virginia Medicaid Management Information System can be estimated that control for preexisting observable differences between the study groups. Propensity scores that control for overt bias can therefore be included in OLS regression models to derive less biased estimates of high intensity program participation. Third, several advantages and disadvantages exist when using either OLS or IV regression. Researchers who plan to evaluate DM programs using either of these regression methods should consider their own personal strengths as well as the advantages and disadvantages of the methods before committing to either approach. Failure to do this may result in the researchers finding themselves in situations where they are unable to effectively use either method to evaluate DM programs.

List of References

List of References

Achen C.H. (1986). *The Statistical Analysis of Quasi-Experiments.* Berkeley, CA: University of California Press.

Afifi, A.A., Morisky, D.E., Kominski, G.F., & Kotlerman, J.B. (2007). Impact of disease management on health care utilization: evidence from the "Florida: a healthy state Medicaid program." *Preventive Medicine*, 44, 547 –533.

Allen, M., Iezzoni, L.I., Huang, A., Huang, L., & Leveille, S.G. (2008). Improving patient-clinician communication about chronic conditions: description of an internet-based nurse e-coach intervention. *Nursing Research*, 57, 107 – 112.

Allison, P.D. (1999). Logistic Regression Using the SAS System. Cary, NC: The SAS Institute, Inc.

American Diabetes Association (2002). Economic costs of diabetes in the U.S. in 2002. *Diabetes Care*, 26, 917 – 932.

Angrist, J.D. & Krueger, A.B. (2001). Instrumental variables and the search for identification: from supply and demand to natural experiments. *Journal of Economic Perspectives*, 15, 69 – 85.

Austin, P.C., Grootendorst, P., & Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a monte carlo study. *Statistics in Medicine*, 26, 734 – 753.

Austin, P.C. (2008). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine*, 27, 2037 – 2049.

Baser, O. (2006). Too much ado about propensity score models? comparing methods of propensity score matching. *Value in Health*, 9, 377 – 385.

Basu, A., Heckman, J.J., Navarro-Lozano, S., & Urzua, S. (2007). Use of instrumental variables in the presence of heterogeneity and self-selection: an application to treatments of breast cancer patients. Health Economics, 16, 1133 – 1157.

Baum, C.F. (2006). *An Introduction to Modern Econometrics Using Stata.* College Station, TX: Stata Press.

Berg, G.D. & Wadhwa, S. (2007). Health services outcomes for a diabetes disease management program for the elderly. *Disease Management*, 10, 4, 226 – 234.

Bodenheimer, T. (2000). Disease management in the American market. *BMJ*, 320, 563 – 566.

Bodenheimer, T., Lorig, K., Holman, H., & Grumbach, K. (2002). Patient self-management of chronic disease in primary care. *Journal of the American Medical Association*, 288, 2469 – 2475.

Brookhart, M.A., Schneeweiss, S. Rothman, K.J., Glynn, R.J., Avorn, J., Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*. doi:10.1093/aje/kwj149

Buntin, M.B. (2006). Rigorous disease management evaluation. *Journal of Evaluation in Clinical Practice*, 12, 2, 121 – 123.

Caliendo, M. & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, 22, 31 –72.

Center on an Aging Society (2004). *Disease Management Programs: Improving Health While Reducing Costs?* Retrieved February 22, 2007, from http://hpi.georgetown.edu/agingsociety/pdfs/management.pdf.

Centers for Disease Control and Prevention (2005a). *Chronic Disease Overview.* Retrieved October 12, 2007 from http://www.cdc.gov/nccdphp/overview.htm.

Centers for Disease Control and Prevention (2005b). National Diabetes Fact Sheet. Retrieved October 20, 2007 from http://www.cdc.gov/diabetes/pubs/factsheet.htm.

Choe, H.M., Mitrovich, S., Dubay, D., Haywood, R.A., Krein, S.L., & Vijan, S. (2005). Proactive case management of high-risk patients with type 2 diabetes mellitur by a clinical pharmacist: a randomized controlled trial. *American Journal of Managed Care*, 11, 253 – 260.

Christakis, D., Connell, F., Richardson, A., & Maciejewski, M. (2004). *Report of Disease Management Evaluation*, University of Washington.

Clark, N.M. & Dodge, J.A. (1999). Exploring self-efficacy as a predictor of disease management. *Health Education and Behavior*, 26, 72 – 89.

Coberly, C.R., McGinnis, M., Orr, P.M., Coberly, S.S., Hobgood, A., Hamar, B., Gandy, B., Pope, J., Hudson, L., Hara, P., Shurney, D., Clarke, J.L., Crawford, A. & Goldfarb (2007). Associated between frequency of telephonic contact and clinical testing for a large, geographically diverse diabetes disease management population. *Disease Management*, 10, 2, 101 – 109.

Cohen, J., Cohen, P., West, S.G., & Aiken, L.S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (3rd Ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Congressional Budget Office (2004). *An Analysis of the Literature on Disease Management Programs*. Retrieved September 29, 2007, from www.cbo.gov.

Cutler, T.W., Palmieri, J., Khalsa, M., & Stebbins, M. (2007). Eavluation of the relationship between a chronic disease care management program and California pay-for-performance diabetes care cholesterol measures in one medical group. *Journal of Managed Care Pharmacy*, 13, 578 – 588.

D'Agostino, R.B. & Kwan, H. (1995). Measuring effectiveness: what to expect without a randomized control group. *Medical Care*, 33, AS95 – AS105.

D'Agostino, R.B. (1998). Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine*, 17, 2265 – 2281.

Davidson, E.J. (2005). *Evaluation Methodology Basics: The Nuts and Bolts of Sound Evaluation*. Thousand Oaks, CA: Sage Publications.

Department of Medical Assistance Services (2005a). *Disease Management and Virginia's Medicaid Program* (House Document No. 90). Commonwealth of Virginia: Author.

Department of Medical Assistance Services (2005b). *Request for proposals 2005-06*. Commonwealth of Virginia: Author.

Department of Medical Assistance Services (2006). *Virginia Medicaid Healthy Returns^{SM} Disease Management Program*. Commonwealth of Virginia: Author.

Diamond, F. (1999). *DM's Motivation Factor Can Skew Study Results*. Retrieved April 10, 2008 from http://www.managedcaremag.com/archives/9906/9906.dmstudies.html.

Disease Management Association (n.d.). DMAA Definition of Disease Management. Retrieved August 29, 2008 from http://www.dmaa.org/dm_definition.asp.

Doran, T., Fullwood, C., Graville, H., Reeves, D., Kontopantellis, Hiroeh, U., & Roland, M. (2007). *Pay-for-Performance Programs in Family Practices in the United Kingdom*. Retrieved November 10, 2008 from http://content.nejm.org/cgi/reprint/335/4/375.pdf.

Ender, P. (n.d.). *Applied Categorical and Nonnormal Data Analysis: Instrumental Variables Regression*. Retrieved July 23, 2008 from http://www.gseis.ucla.edu/courses/ed231c/notes3/instrumental.html.

Ettner, S. (2004). *Methods for Addressing Selection Bias in Observational Studies*. Retrieved September 15, 2007 from http://www.ahrq.gov/fund/training/ettnertxt.htm.

Farrell, K., Wicks, M.N., & Martin, J.C. (2004). Chronic disease self-management improved with enhanced self-efficacy. *Clinical Nursing Research*, 13, 289 – 308.

Fetterolf, D. & Olson, M. (2008). Opt-in medical management strategies. *Disease Management*, 11, 37 – 46.

Fireman, B., Bartlett, J., & Selby, J. (2004). Can disease management reduce health care costs by improving quality? *Health Affairs*, 23, 63 – 75.

Florida Department of Health (2007). Chronic Disease Prevention. Retrieved April 14, 2008 from http://www.doh.state.fl.us/Family/chronicdisease/.

Foote, S. M. (2003). Population-based disease management under fee-for-service medicare. Retrieved February 18, 2008 from http://content.healthaffairs.org/cgi/reprint/hlthaff.w3.342v1.pdf.

Fortney, J., Booth, B., Zhang, M., Humphrey, J., & Wiseman, E. (1998). Controlling for selection bias in the evaluation of alcoholics anonymous as aftercare treatment. *Journal of Studies in Alcohol*, 59, 690 – 697.

Foster, E.M. & McLanahan S. (1996). An illustration of the use of instrumental variables: do neighborhood conditions affect a young person's chance of finishing high school? *Psychological Methods*, 1, 249 – 260.

Frank, KA., Sykes, G., Anagnostopoulos, D., Cannata, M., Chard, L., Krause, A., & McCrory, R. (2008). Does nbpts certification affect the number of colleagues a teacher helps with instructional matters? *Educational Evaluation and Policy Analysis*, 30, 3 – 30.

Freedman, D.A. (2005). *Statistical Models: Theory and Practice*. Cambridge, NY: Cambridge University Press.

Freedman, D.A. (2006). Statistical models for causation: what inferential leverage do they provide? *Evaluation Review*, 30, 691 – 713.

Funnell, M.M. & Anderson, R.M. (2002). Working toward the next generation of diabetes self-management education. *American Journal of Preventive Medicine*, 22, 3 – 5.

Gelman, A. & Hill, J. (2007). Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge, NY: Cambridge University Press.

Gillespie, J.L. & Rossiter, L.F. (2003). Medicaid disease management programs: findings from three leading us state programs. *Disease Management & Health Outcomes*, 11, 6, 345 – 361.

Gozalo, P. L. & Miller, S. C. (2006). Predictors of Mortality: hospice enrollment and evaluation of its causal effect on hospitalization of dying nursing home patients. *Health Services Research*, 42, 587 – 610.

Greene, W.H. (2003). Econometric Analysis (5th Ed.). Upper Saddle River, NJ: Prentice Hall.

Greenland, S. (2000). An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology*, 29, 722 – 729.

Gujarati, D.N. (2003). *Basic Econometrics, 4th Ed.* Boston,MA: McGraw Hill.

Guo, S., Barth, R., & Gibbons, C. (2008). Propensity score matching strategies for evaluating substance abuse services for child welfare clients. *Children and Youth Services Review*, 28, 357 – 383.

Hadley, J., Polsky, D., Mandelblatt, J.S., Mitchell, J.M., Weeks, J.C., Wang, Q., Hwang, Y.T., & the OPTIONS Research Team (2003). An exploratory instrumental variable analysis of the outcomes of localized breast cancer treatments in a medicare population. *Health Economics*, 12, 171 – 186.

Hadley, J. & Cunningham, P. (2004). Availability of safety net providers and access to care of uninsured persons. *Health Services Research*, 39, 1527 – 1546.

Hanley, J.A., Negassa, A., deB. Edwards, M.D., & Forrester, J.E. (2003). Statistical analysis of correlated data using generalized estimating equations: an orientation. American Journal of Epidemiology, 157, 364 – 375.

Haro, J.M., Kontodimas, S., Negrin, M.A., Ratcliffe, M., Suarez, D., & Windmeijer, F. (2006). *Applied Health Economics and Health Policy*, 5, 11 – 25.

Harris, K.M. & Remler, D.K. (1998). Who is the marginal patient? Understanding instrumental variables estimates of treatment effects. *Health Services Research*, 33, 5, 1337 – 1360.

Health Management Corporation (2007). *Disease Management Annual Report Presentation – December 5, 2007*. Richmond, VA: Author.

Health Management Corporation (2008). *Department of Medical Assistance Services: Integrated Account Report – August 6, 2008*. Richmond, VA: Author.

Hernan, M.A. & Robins, J.M. (2006). Instruments for causal inference: an epidemologist's dream? *Epidemiology*, 17, 4, 360 – 372.

Hill, J. (2008). Discussion of research using propensity-score matching: comments on 'a critical appraisal of propensity-score matching in the medical literature between 1996 and 2003 by peter austin, *statistics* in medicine. *Statistics in Medicine*, 27, 2055 – 2061.

Holahan, J. & Ghosh, A. (2005). Understanding the Recent Growth in Medicaid Spending, 2000 – 2003. Retrieved October 16, 2007 from http://content.healthaffairs.org/cgi/content/abstract/hlthaff.w5.52.

Hunter, D.J. & Fairfield, G. (1997). Managed care: disease management. *BMJ*, 315, 50 – 53.

Johnson, A. (2003). Disease Management: The Programs and the Promise. Retrieved April 14, 2008 from http://www.milliman.com/expertise/healthcare/publications/rr/pdfs/Disease-Mangement-Programs-Promise-RR05-01-03.pdf.

Johnston, K.M., Gustafson, P., Levy, A.R., & Grootendorst, P. (2008). Use of instrumental variables in the analysis of generalized linear models in the presence of unmeasured confounding with applications to epidemiological research. *Statistics in Medicine*, 27, 1539 – 1556.

Jones, A.S., & Richmond, D.W. (2006). Causal effects of alcoholism on earnings: estimates from NLSY. *Health Economics*, 15, 849 – 871.

Judd, C. H., & Kenny, D.A. (1981). *Estimating the Effects of Social Interventions.* Cambridge, MA: Cambridge University Press.

Kennedy, P. (2003). *A Guide to Econometrics* (5th Ed.). Cambridge, MA: The MIT Press.

Knight, K., Badamgarav, E., Henning, J.M., Hasselblad, V., Gano, A.D., Ofman, J.J., & Weingarten (2005). A systematic review of diabetes disease management programs. *The American Journal of Managed Care*, 11, 242 – 250.

Krein, S.L. & Klamerus, M.L. (2000). Michigan diabetes outreach networks: a public health approach to strengthening diabetes care. *Journal of Community Health*, 25, 495 – 511.

Landon, B. E., Hicks, L.S., O'Malley, A.J., Lieu, T.A., Keegan, T., McNeil, B.J., & Guadagnoli, E. (2007). Improving the management of chronic disease at community health centers. *The New England Journal of Medicine*, 356, 9, 921 – 934.

Landrum, M.B. & Ayanian, J.Z. (2001). Causal effect of ambulatory specialty care on mortality following myocardial infarction: a comparison of propensity score and instrumental variable analyses. Health *Services & Outcomes Research Methodology*, 2, 221 – 245.

Leigh, J.P & Schembri, M. (2004). Instrumental variables technique: cigarette price provided better estimate of effects of smoking on sf-12. *Journal of Clinical Epidemiology*, 57, 284 – 293.

Linden, A., Adams, J.L, & Roberts, N. (n.d.). Evaluation Methods in Disease Management: Determining Program Effectiveness. Retrieved May 14, 2008 from http://www.dmaa.org/pdf/Evaluation_Methods_in_DM.pdf.

Linden, A., Adams, J.L., & Roberts, N. (2005). Using propensity scores to construct comparable control groups for disease management program evaluation. *Disease Management Health Outcomes*, 13, 107 – 115.

Linden, A. (2006). What will it take for disease management to demonstrate a return on investment? new perspectives on an old theme. *The American Journal of Managed Care*, 12, 217 – 222.

Linden, A. and Adams, J.L. (2006). Evaluating disease management program effectiveness: an introduction to instrumental variables. *Journal of Evaluation in Clinical Practice*, 12, 2, 148 – 154.

Lindrooth, R.C., Hoerger, T.J., & Norton, E.C. (2000). Expectations among the elderly about nursing home entry. *Health Services Research*, 35, 1181 – 1202.

Lohr, K.N., Keeler, E.B., Calabro, T.A., & Brook, R.H. (1986). Chronic disease inn a general adult population: findings from the rand health insurance experiment. *The Western Journal of Medicine*, 145, 537 – 545.

Long, S.K., Coughlin, T., & King, J. (2005). How well does Medicaid work in improving access to care? *Health Services Research*, 40, 39 – 58.

Lorig, K., Stewart, A., Ritter, P., Gonzalez, V., Laurent, D., & Lynch, J. (1996). *Outcome Measures for Health Education and Other Health Care Interventions*. Thousand Oaks, CA: Sage Publications.

Lorig, K.R., Sobel, D.S., Ritter, P.L., Laurent, D., & Hobbs, M. (2001). Effect of a self-management program on patients with chronic disease. *Effective Clinical Practice*, 4, 256 – 262.

Love, T.E. (2003). Propensity Scores: What Do They Do, How Should I Use Them, and Why Should I Care? Retrieved May 14, 2008 from http://www.chrp.org/love/ASACleveland2003Propensity.pdf.

Love, T.E. (2004). *Using Propensity Score Methods Effectively.* Retrieved May 14 2008 from http://www.chrp.org/love/ASA_love_oct1 12004.pdf.

Lynn, J. & Adamson, D.M. (2003). *Living Well at the End of Life Adapting Health Care to Serious Chronic Illness in Old Age*. Retrieved November 13, 2008 from http://www.rand.org/pubs/white_papers/2005/WP137.pdf

Lunceford, J.K. & Davidian, M. (2004). Stratification and weighting via the propensity score in the estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, 23, 2937 - 2960.

Ma. S. (2008). Paternal race/ethnicity and birth outcomes. *American Journal of Public Health*, 98, 1 – 8.

MacDowell, M. and Wilson, T. (2002). *Framework for Assessing Causality in Disease Management Programs*. Retrieved September 29, 2007, from http://www.dmaa.org/pdf/FrameworkCausalityDM.pdf.

Malkin, D, Broder, M.S., & Keeler, E., (2000) Do longer postpartum stays reduce newborn readmissions? analysis using instrumental variables. *Health Services Research*, 35, 1071–1091

Mangione, C.M., Gerzoff, R.B., Williamson, D.F., Steers, W.N., Kerr, E.A., Brown, A.F., Waitzfelder, B.E., Marrero, D.G., Dudley, R.A., Kim, C., Herman, W., Thompson, T.J., Safford, M.M., & Selby, J.V., (2006). The association between quality of care and the intensity of diabetes disease management programs. *Annals of Internal Medicine*, 145, 107 – 116.

Marks, R., Allegrante, J.P., & Lorig, K. (2005). A review and synthesis of research evidence fro self-efficacy-enhancing interventions for reducing chronic disability: implications for health education practice (part 1). Health Promotion Practice, 6, 37 – 43.

Martens, E.P., Pestman, W.R., de Boer, A., Belitser, S.V., & Klungel, O.H. (2006). Instrumental variables: applications and limitations. *Epidemiology*, 17, 260 – 267.

Matheson, D., Wilkins, A., & Psacharopoulos, D. (2006). *Realizing the Promise of Disease Management: Payer Trends and Opportunities in the United States*. Retrieved November 13, 2008 from http://www.bcg.com/publications/files/Realizing_the_Promise_of_Disease_Management_Feb06.pdf.

Mattke, S., Seid, M., & Ma, S. (2007). Evidence for the effect of disease management: is $1 billion a year a good investment? *American Journal of Managed Care*, 13, 670 – 676.

McClellan, M., McNeil, B.J., & Newhouse, J.P. (1994). Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? *JAMA*, 272, 859 – 866.

McClellan, M., & Newhouse, J. (2000). Overview of the special supplement issue. *Health Services Research*, 35, 1061 – 1069.

McMillan & Schumacher (2006). *Research in Education: Evidence-Based Inquiry* (6th Ed.). Boston, MA: Pearson Education, Inc.

Meigs, J.B., Cagliero, E., Dubey, A., Murphy-Sheehy, P., Gildesgame, C., Chueh, H., Barry, M.J., Singer, D.E., & Nathan, D.M. (2003). A controlled trial of web-based diabetes disease management. *Diabetes Care*, 26, 3, 750 – 757.

Mendenhall, W & Sincich, T. (2003). *A Second Course in Statistics: Regression Analysis(6th Ed)*. Upper Saddle River, NJ: Pearson Education, Inc.

Mohr, L.B. (1995). *Impact Analysis for Program Evaluation* (2nd Ed.). Thousand Oaks, CA: Sage Publications.

Morgan, S.L. & Winship, C. (2007). *Counterfactual and Causal Inference: Methods and Principles for Social Research*. New York, NY: Cambridge University Press.

Morisky D.E., Kominski, G.F., Afifi, A.A., & Kotlerman, J.B. (2008). The effects of a disease management program on self-reported health behaviors and health outcomes: evidence from the "florida: a health state (fahs)" Medicaid program. *Health Education OnlineFirst*. doi: 10.1177/1090198107311279.

Mullahy, J. (1998). Much ado about two: reconsidering retransformation and the two-part model in health econometrics. *Journal of Health Economics*, 17, 247 – 281.

National Center for Health Statistics (2007). *International Classification of Diseases, Ninth Revision (ICD-9)*. Retrieved November 8, 2007, from http://www.cdc.gov/nchs/about/major/dvs/icd9des.htm.

Newgard, C.D., Hedges, J.R., Arthur, M., & Mullins, R.J. (2004). Advanced statistics: the propensity score – a method for estimating treatment effect in observational research. *Academic Emergency Medicine*, 11, 953 – 961.

Newhouse, J.P. & McCellan, M. (1998). Econometrics in outcomes research: the use of instrumental variables. *Annual Review of Public Health*, 19, 17 – 34.

Newhouse, J.P. (2005). Instrumental Variables in Health Services Research. In P. Armitrage and T. Colton (Eds.), *Encyclopedia of Biostatistics (2nd Ed.)* (2561 – 2565). Hoboken, NJ: John Wiley and Sons.

Normand, S.L.T., Sykora, K., Li, P., Mamdani, M., Rochon, P.A., & Anderson, G.A. (2005). Readers guide to critical appraisal of cohort studies: 3 analytical strategies to reduce confounding. *BMJ*, 330, 1021 – 1023.

Nuovo, J. (Ed.) (2007). *Chronic Disease Management*. New York, NY: Springer.

Odegard, P.S., Goo, A., Hummel, J., Williams, K.L., & Gray, S.L. (2005). Caring for poorly controlled diabetes mellitus: a randomized pharmacist intervention. *The Annals of Pharmacotherapy*, 39, 433 – 440.

Ofman, J.J., Badamgarav, E., Henning, J.M., Knight, K., Gano, A.D., Levan, R.K., Gur-Arie, S., Richards, M.S, Hasselbald, V., & Weingarten, S.R. (2004). Does disease management improve clinical and economic outcomes in patients with chronic diseases? A systematic review. *The American Journal of Medicine*, 117, 182 – 192.

O'Kane, M.E. (2007). *Performance-Based Measures: The Early Results Are In.* Retrieved November 10, 2008 from http://www.amcp.org/data/jmcp/Pages%203-61.pdf.

Peng, C.Y.J., Lee, K.L., & Ingersoll, G.M. (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research*, 96, 3 – 14.

Piecoro, L.T., Wang, L.S., Dixon, W.S., & Crovo, R.J. (1999). Creating a computerized database from administrative claims data. *American Journal of Health-System Pharmacists*, 56, 1326-1329.

Posner, M.A., Ash, A.S., Freund, K.M., Moskowitz, M.A., & Shwartz, M. (2002). Comparing standard regression, propensity score matching, and instrumental variables methods for determining the influence of mammography on stage of diagnosis. *Health Services and Outcomes Research Methodology*, 2, 279 -290.

Redman, B.K. (2005). The ethics of self-management preparation for chronic illness. *Nursing Ethics*, 12, 360 – 369.

Robins, J.M & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical; Association*, 90, 122 – 129.

Robins, J.M., Hernan, M.A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemology*, 11, 550 – 560.

Rosenbaum, P.R. & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41 – 55.

Rosenbaum, P.R. (1995). *Observational Studies*. New York, NY: Springer – Verlag.

Rosenbaum, P.R. (2005). Sensitivity Analysis in Observational Studies. In B.S. Everitt and D.C. Howell (Editors), Encyclopedia of Statistics in Behavioral Science (4th Ed.) (1809–1814). Hoboken, NJ: John Wiley and Sons.

Rothman, R.L. & Elasy, T.A. (2005). Can diabetes management programs create sustained improvements in disease outcomes? Canadian Medical Association Journal, 173, 1457 – 1466.

Sadur, C.N., Moline, N., Costa, M., Michalik, D., Mendlowitz, D., Roller, S., Watson, R., Swain, B.E., Selby, J.V., & Javorski, W.C. (1999). Diabetes management in a health maintenance organization. *Diabetes Care*, 22, 2011 – 2017.

Sales, A.E., Plomondon, M.E., Magid, D.J., Spertus, J.A., & Rumsfeld, J.S. (2004). Assessing response bias from missing quality of life data: the heckman method. *Health and Quality of Life Outcomes*, 2, 49.

Salkever, D.S., Slade, E.P., Karakus, M., Palmer, L., & Russo, P.A. (2004). Estimation of antipsychotic effects on hospitalization risk in a naturalistic study with selection on unobservables. *The Journal of Nervous and Mental Disease*, 192, 2, 119 – 204.

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W.H., & Shavelson, R.J. (2007). *Estimating Causal Effects Using Experimental and Observational Designs*. Washington, DC: American Educational Research Association.

Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin Company.

Shojania, K.G., Ranji, S.R., McDonald, K.M., Grimshaw, J.M., Sundaram, V., Rushakoff, R.J., & Owens, D.K. (2006). Effects of quality improvement strategies for type 2 diabetes on glycemic control: a meta-regression analysis. Journal of the American Medical Association, 296, 427 – 440.

Siu, A.M.H., Chan, C.H.C., Poon, P.K.K., Chui, D.Y.Y., & Chan, S.C.C. (2007). Evaluation of the chronic disease self-management program in a chinese population. *Patient Education and Counseling*, 65, 42 – 50.

Sprague, L. (2003). *Disease Management to Population-Based Health: Steps in the Right Direction?* Retrieved February 18, 2008 from http://nhpf.ags.com/pdfs_ib/IB791_DiseaseMgmt_5-16-03.pdf.

Staiger, D. & Stock, J.A. (1997). Instrumental variables regression with weak instruments. *Econometrica*, 65, 557 – 586.

Stevens, J.P. (2002). *Applied Multivariate Statistics* (4th Ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Stock, J.A. (n.d.). Instrumental Variables Regression (SW Ch. 10). Retrieved July 13, 2008 from http://ksghome.harvard.edu/~jstock/tb/ch10_slides_1.doc.

Stock, J.H. & Watson, M.W. (2007). Introduction to Econometrics (2nd Ed.). Boston, MA: Pearson Education, Inc.

Stukel, T.A., Fisher, E.S., Wennberg, D.E., Alter, D.A., Gottlieb, D.J., & Vermeulen, M.J. (2007). Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on ami survival using propensity score and instrumental variable methods. *JAMA*, 297, 278 – 285.

Sturmer, T., Joshi, M., Glynn, R.J., Avorn, J., Rothman, K.J., & Schneeweiss, S. (2003). R review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariate methods. *Journal of Clinical Epidemiology*, 59, 437.e1 – 437.24.

Suter, P., Hennessey, B., Harrison, G., Fagan, M., Norman, B., & Suter, W.N. (2008). Home-based chronic care. *Home Healthcare Nurse*, 26, 223 – 229.

Tabachnick, B.G. & Fidell, L.S. (2001). *Using Multivariate Statistics* (4th Ed.). Boston, MA: Allyn and Bacon.

Trochim, W.M.K. (2005). *Research Methods: The Concise Knowledge Base*. Mason, OH: Thomson.

U.S. Department of Health and Human Services (2007). Code of Federal Regulations, Title 45 – Public Welfare. Retrieved November 13, 2007 from www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.htm.

Villagra, V.G. & Ahmed, T. (2004). Effectiveness of a diabetes management program for patients with diabetes. *Health Affairs*, 23, 255 – 266.

Weiss, C.H. (1998). *Evaluation (2nd Ed)*. Upper Saddle River, NJ: Prentice Hall.

Weitzen, S., Lapane, K.L., Toledano, A.Y., Hume, A.L., & Mor, V. (2004). Principles for modeling propensity scores in medical research: a systematic review. *Pharmacoepidemiology*, 13, 841 – 853.

Wendel, J. & Dumitras, D. (2005). Treatment effects model for assessing disease management: measuring outcomes and strengthening program management. *Disease Management*, 8, 3, 155 – 168.

Williams, C. (2004). Medicaid Disease Management: Issues and Promises. Retrieved October 18, 2007, from http://www.kff.org/medicaid/upload/Medicaid-Disease-Management-Issues-and-Promises-Issue-Paper.pdf

Winship, C. & Morgan, S.L. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology*, 25, 659 – 706.

Wolf, J. (2006). The benefits of diabetes self-management education of the elderly veteran in the home care setting. *Home Healthcare Nurse*, 24, 645 – 651.

Wooldridge (2002). *Econometric Analysis of Cross Section and Panel Data.* Cambridge, MA: The MIT Press.

Wooldridge, J. M. (2006). *Introductory Econometrics: A Modern Approach* (3rd Ed.). Mason, OH: Thomson Higher Education.

Wunsch, H., Linde-Zwirble, W.T., Angus, D.C. (2006). Methods to adjust for bias and confounding in critical care health services research involving observational data. *Journal of Critical Care*, 21, 1 –7.

Wyant, T. & Parente, S.T., (n.d.). Use of Medicaid and Medicaid Administrative Claims Data in Litigation and Regulation. Retrieved April 14, 2008 from http://www.fcsm.gov/03papers/Wyant.pdf.

Yanovitzky, I., Zanutto, E., & Hornik, R. (2005). Estimating causal effects of public health education campaigns using propensity score methodology. *Evaluation and Program Planning*, 28, 209 – 220.

Zhang, N.J., Wan, T.H.T., Rossiter, L.F. Murawski, M.M., & Patel, U.B. (2008). Evaluation of chronic disease management on outcomes and cost of care for Medicaid beneficiaries. *Health Policy*, 86, 345 – 354.

Appendices

Appendix A: An Overview of the Inverse Propensity Score Weighting Method

Propensity scores (PS) are often used in social science research to estimate treatment effects through one or more of the following three methods: regression covariate adjustment, stratification, or matching. A less frequently used method of employing propensity scores involves using the inverse of the estimated propensity scores to develop a weight for each study subject. The objective of propensity score weighting is to develop weighted averages of the data that approximate what would have been derived through randomized experiments (Gelman & Hill, 2007). According to Austin (2008), propensity score weighting is rarely used in medical research. (Austin does not offer an explanation for why propensity score weighting is rarely used.) As a result, the researcher decided to briefly explore propensity score weighting in Appendix 1 using the dataset developed for the present study.[1]

As an alternative method of employing propensity scores, Robins and colleagues suggest deriving an average treatment effect estimate by conducting a simple regression of the outcome on the treatment using inverse propensity score weights (IPSW) of 1/PS and 1/(1 – PS) for the treatment and control groups, respectively (Frank et al., 2008).[2,3] Following this guidance, high intensity DM program participants (the treatment group) and standard intensity participants (the control group) were weighted accordingly. By using

[1] The researcher's objective in writing Appendix 1 was not to provide a detailed analysis of propensity score weighting. Instead, the objective was to briefly explore its use for possible consideration in a future study. Technically-oriented information on propensity score weighting can be found in sources such as Robins and Rotnitzky (1995), Robins, Hernan, and Brumback, (2000), Lunceford and Davidian, (2004), Austin and Mamdani, (2006), and Morgan and Winship (2007).

[2] IPSW can also be used in multiple regressions (Gozalo & Miller, 2007).

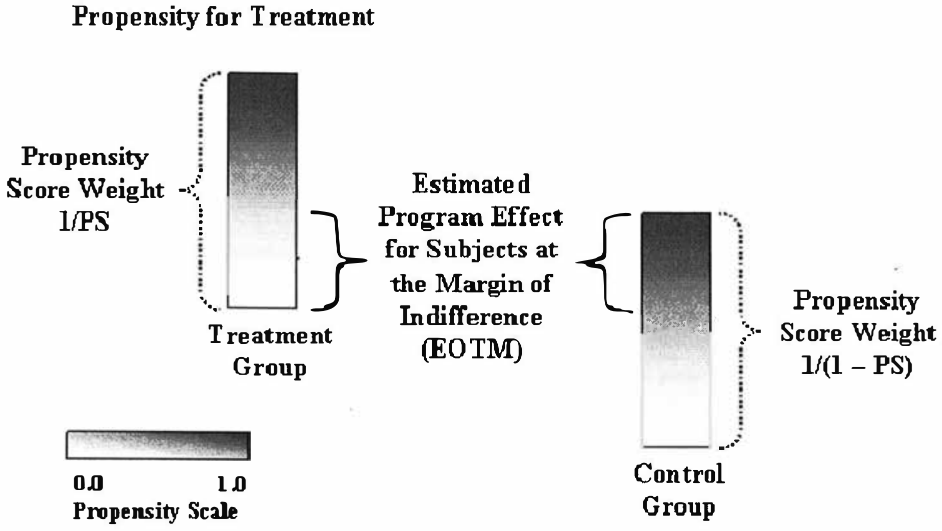
[3] Additional weighting procedures exist that can be used to estimate the treatment effect for the treated subjects or the treatment effect for the control subjects (Gelman & Hill, 2007; Frank et al., 2008). However, these procedures are not discussed in this appendix section.

this procedure, the high-intensity DM participants were weighted more the lower their propensity for enrollment in the high-intensity program, while the standard intensity participants were weighted more the higher their propensity for enrollment in the high-intensity program (Frank et al., 2008). The weights were calculated using SAS version 9.1, while the regression analyses were performed using STATA version 10.1.

The comparison produced using the IPSW method focuses the analysis on the strongest overlap in propensity scores. Specifically, it focuses on subjects who are most likely to respond to change: individuals who received the treatment, but had a low propensity for doing so, with individuals who did not receive the treatment, but had a high propensity for doing so (Frank et al., 2008). Through this process, IPSW mimics propensity score matching by essentially creating a hypothetical group of treatment subjects who are similar to the control subjects in terms of all characteristics except actual treatment assignment (Morgan & Winship, 2007; Ma, 2008). The estimate produced using IPSW is referred to as the estimated effect for individuals at the margin of indifference (EOTM) (Figure A1) (Frank et al., 2008).

There are several advantages of using the IPSW procedure over propensity score matching, stratification, and regression adjustment. In particular, the IPSW method offers a relatively intuitive process for controlling for overt bias that is easy to calculate and use in regression analyses. Moreover, the IPSW method can improve estimation efficiency because the full sample of subjects is retained in the analysis (however, not all subjects contribute equally to average effect estimates due to the weighting scheme) (Frank et al.,

Figure A1. Propensity Score Overlap for the Treatment and Control Groups


Propensity for Treatment
Propensity Score Weight 1/PS
Treatment Group
Estimated Program Effect for Subjects at the Margin of Indifference (EOTM)
Control Group
Propensity Score Weight 1/(1 – PS)
0.0
1.0
Propensity Scale

Source: Frank et al. (2008)

2008). In addition, the IPSW method can balance the treatment and control groups by producing equal covariate distributions between the study groups (Ma, 2008). The major concern with using this method is that some propensity score weights may have extreme values that exert undue influence on the average treatment effect estimates. This concern can be ameliorated, however, by examining the dataset to identify these weights and trimming their values so they do not unduly influence average treatment effect estimates (Frank et al., 2008).

For this analysis, extreme weights were identified as cases with standardized scores greater than 3.29 (Tabachnick & Fidell, 2001). Thirty-two weights met this criterion and were thus considered extreme. Using the procedure identified by Frank et al. (2008), these weights were trimmed to a value of one greater than the next most extreme weight in the dataset, which was 12.62 (Frank et al., 2008).[4] To test the sensitivity of the IPSW method to these extreme values, each outcome was regressed on the program participation variable weighted by both the trimmed and untrimmed weights. These simple regressions revealed that the trimmed and untrimmed weighted program coefficients were very similar for the ED visits and hospital days models, which may suggest that the extreme values did not actually exert much influence in the regressions. However, the trimmed weighted program coefficient for the diabetes-related cost model was higher (1.62) than the untrimmed weighted coefficient (1.56). This finding could suggest that the extreme values did exert some influence in the diabetes cost regression model.

The researcher next used the trimmed and untrimmed weights to assess covariate balance between the study groups. This process was accomplished using a series of simple linear probability models (i.e., the treatment variable was regressed on the covariate using OLS regression) (Cohen et al., 2003; Wooldridge, 2006).[5] The models were generated for each covariate, which was weighted separately using the trimmed and untrimmed weights. The results are presented in Table A1. The regressions revealed that the trimmed weights did not control for significant differences between study groups on the 2006 diabetes-related cost variable ($p = 0.015$), while the untrimmed weights did control for all significant differences between the study groups. (The untrimmed weighted results are

[4] The average IPSW was 1.96 (*SD* = 2.82). The range of the weights was 1.00 to 32.51. Weights greater than 11.62 were considered extreme based on having a standardized score greater than 3.29.

[5] Some observers may argue that logistic regression should have been used to assess covariate balance. The researcher decided against using logistic regression for two reasons: 1) Wooldridge (2006) indicates that linear probability models are often used in economics when the outcome is dichotomous, and 2) the logit command in STATA version 10.1 did not support analytic weights (i.e., weights that are inversely proportional to the variance of the observations), which was the procedure used to employ the inverse probability score weights in the present study. However, OLS regression in STATA supports using analytic weights.

similar to the results obtained using logistic regression to assess covariate balance, which are presented in Table 3 in Chapter 4.) Because the covariates were similarly balanced between study groups when using the trimmed and untrimmed weights, a decision was made to report the results from the regressions using both weights because: 1) the extreme untrimmed inverse propensity score weights may have exerted some influence in the diabetes-cost regressions and 2) the untrimmed weights appeared to more effectively control for covariate balance between study groups than the untrimmed weights.

*Table A1. Simple Linear Probability Regressions to Assess Covariate Balance Using IPSW (N = 1,627)**

Variables	Unadjusted *p* value**	Adjusted *p* value (IPSW Trimmed)**	Adjusted *p* value (IPSW Untrimmed)**
Female	0.546	0.528	0.211
Nonwhite Race	0.001	0.351	0.378
Western Region	0.001	0.367	0.489
Country of Origin US	0.101	0.701	0.224
English Language	0.508	0.096	0.127
US Citizen	0.135	0.777	0.137
Age	0.131	0.351	0.860
Hospital Days	0.000	0.211	0.313
ED Visits	0.000	0.873	0.414
Diabetes Related Costs	0.000	0.015	0.462

*Outcome variable is program participation (1 = high intensity participation and 0 = non-high intensity participation)
**$\alpha = 0.05$

The researcher next estimated the average effect of the high intensity DM program by regressing the study outcome variables on the program participation variable weighted using both trimmed and untrimmed inverse propensity scores.[6] For comparison purposes, additional multiple regression models were generated by regressing the outcomes on the program participation and quantitative propensity score variables as well as the program participation, age, gender, and quantitative propensity score variables. The estimated effects of the high intensity DM program that were derived from the regressions are resented in Table A2.

Four observations can be made about the information in Table A2. First, the propensity score weights do not appear to add much to the program effect estimates in this study because the weighted estimates are comparable to the unweighted estimates that were calculated in the multiple regressions. In other words, none of the four regression models in Table A2 are clearly dominant.

Second, the program effect estimates may be relatively stable because comparable estimates were derived in the four regression models. However, readers should not conclude that any of the models presented in Table A2 actually depict the "true" casual effect of the high intensity program due to omitted variable bias (i.e., propensity scores can only control for overt bias to the extent that important observable variables are included in the calculation process and hidden bias to the extent that the included variables are correlated with excluded unobserved variables).

[6] Simple regression was used because the inverse propensity score weights account for the covariates included in the propensity score models (Frank et al., 2008).

Table A2. *Estimated Effect of High Intensity DM Participation on the Study Outcomes*

Model	*Coefficient (SE)**	*p value**	*95% CI**
Models Weighted by IPSW (Trimmed)**			
Diabetes Related Costs	1.62 (0.12)	0.000	1.38 – 1.86
ED Visits	0.17 (0.05)	0.002	0.06 – 0.27
Hospital Days	0.37 (0.05)	0.000	0.26 – 0.48
Models Weighted by IPSW (Untrimmed)**			
Diabetes Related Costs	1.56 (0.12)	0.000	1.32 – 1.79
ED Visits	0.19 (0.05)	0.000	0.09 – 0.30
Hospital Days	0.34 (0.05)	0.000	0.23 – 0.45
Models Using Propensity Score Covariate**			
Diabetes Related Costs	1.53 (0.21)	0.000	1.13 – 1.94
ED Visits	0.19 (0.08)	0.013	0.04 – 0.35
Hospital Days	0.42 (0.07)	0.000	0.28 – 0.56
Models Using Age, Gender, and Propensity Score Covariates**			
Diabetes Related Costs	1.53 (0.20)	0.000	1.14 – 1.93
ED Visits	0.20 (0.08)	0.011	0.04 – 0.35
Hospital Days	0.42 (0.07)	0.000	0.28 – 0.56

*High intensity program effect coefficient, standard error, *p* Value, and 95% confidence interval for each outcome variable.
** Study outcomes are transformed: diabetes-related costs (log), emergency department visits (square root), and hospital days (square root).

Third, because the four regression models produced similar estimates, researchers faced with similar results could use that information to depict a combination of approaches that may provide useful boundaries on the magnitude of the true program effect estimate. However, the researchers must ensure that they aggressively controlled for both overt and hidden biases if using these propensity score methods to bolster their argument that they have derived true program effect estimates. In order to accomplish this, researchers would probably need access to multiple "information rich" datasets to identify as many variables as possible that may have directly or indirectly (through their correlations with important unobserved variables) accounted for preexisting differences between the study groups. These variables would then have to be included in the propensity score calculation process in order for the scores to control for these counfounding variables. Failure to do this would limit the ability of the researchers to argue persuasively that they have derived unbiased boundaries for the true effect of the program.

Finally, the effect estimates produced in the third set of models are almost identical to the effect estimates produced in the fourth set of models. This finding suggests that simply including a propensity score covariate in a regression of the outcome on a treatment variable may be just as effective as including a propensity score and other variables as covariates in the regression, but this interpretation is dependent upon whether the other variables were included in the propensity score calculation process. If the variables were included, then little may be gained by including them and the propensity scores as covariates in the regressions. If the variables were not included in the propensity score calculation, then it may be more feasible to include them in the regression as covariates; however, if the variables contribute to study group assignment, then they should be included in the propensity score calculation.

Based on the information presented in this appendix section, inverse propensity score weights appear to offer a viable means of estimating average treatment effects in DM evaluations that use observational data.[7] However, the determination of whether propensity score weights, matching, stratification, or covariate adjustment are preferable in a particular case will have to be made by the researchers based on their knowledge of the study topic and relevant statistical methodologies.

[7] Morisky et al. (2008) used inverse propensity score weights to adjust for attrition bias in the study they performed on the effects of Florida Medicaid's DM program on self-reported health behaviors and outcomes.

Appendix B: Simple Logistic Regression Analysis to Assess Covariate Balance Using the Quintile Propensity Score (N = 1,627)

Variables	Treatment (*n* = 229)*	Control (*n* = 1,398)*	Unadjusted p-value**	Adjusted p-value**
Female	166 (72.5%)	986 (70.5%)	0.546	0.695
Nonwhite Race	52 (22.7%)	198 (14.2%)	0.001	0.079
Western Region	170 (74.2%)	1,166 (83.4%)	0.001	0.096
Country of Origin US	225 (98.3%)	1,343 (96.1%)	0.110	0.249
English Language	226 (98.7%)	1,386 (99.1%)	0.511	0.615
US Citizen	226 (98.7%)	1,355 (96.9%)	0.147	0.248
Age	46.6 (14.25)	44.93 (15.70)	0.131	0.468
Hospital Days	4.90 (29.75)	0.44 (1.91)	0.000	0.000
ED Visits	2.36 (3.65)	1.50 (3.20)	0.001	0.083
Diabetes Related Costs	$7,473.33 ($16,962.42)	$1,060.48 ($2,380.31)	0.000	0.000

Researcher Vitae

[redacted]