Title

Studies in the Methods and Applications of Comparative Effectiveness Research: Hospice Stay Determinants, Human Papillomavirus Vaccination in Men, and Analytic Techniques for Evaluating Experimental and Non-experimental Data

Author

Sunkara, Srinivasu

Publication Date

2013

Peer reviewed|Thesis/dissertation

University of California
Los Angeles

Studies in the Methods and Applications of Comparative Effectiveness Research: Hospice Stay Determinants, Human Papillomavirus Vaccination in Men, and Analytic Techniques for Evaluating Experimental and Non-experimental Data

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Health Services

by

Srinivasu Ammisetty Sunkara

2013

Abstract

Studies in the Methods and Applications of Comparative Effectiveness Research: Hospice Stay Determinants, Human Papillomavirus Vaccination in Men, and Analytic Techniques for Evaluating Experimental and Non-experimental Data

by

Srinivasu Ammisetty Sunkara
Doctor of Philosophy in Health Services
University of California, Los Angeles 2013
Professor Robert M. Kaplan, Chair

The passage of the Affordable Care Act (ACA) of 2011 made comparative effectiveness research (CER) a major priority of national healthcare reform. Many health policy and academic leaders believe that direct comparison of competing healthcare interventions will lead to improvements in population health outcomes This dissertation contributes to the field by applying CER techniques to understand the determinants of hospice stay duration, the cost-effectiveness of male-female Human Papillomavirus (HPV) vaccination, and the optimal method for evaluating non-experimental and experimental data.

The research on hospice stay duration uses national-level data from the Dartmouth Atlas. It finds positive correlations between the logarithm of hospice days and hospice reimbursement ($p<0.001$), inpatient days and total physician visits ($p<0.001$), and a negative correlation between inpatient days and outpatient reimbursement ($p=0.001$). Multiple regression analysis demonstrated that nurse staffing numbers are a significant determinant of length-of-stay. In addition, simple regression analysis shows that total Medicare reimbursements go down by $117 for each day that a patient stays in

hospice.

The HPV paper aggregates results across 7 published papers. It finds substantial variability in the published incremental cost-effectiveness ratios (ICERs) for male-female HPV vaccination. There is also a lack of standardization in the disease states that are evaluated across these papers. Using fixed-effects analysis, male-female HPV vaccination is found to have a cost-effectiveness of $42,425/QALY. Comparatively, female HPV vaccination has a cost-effectiveness of $8,498/QALY.

The third paper compares the accuracy of linear regression, propensity score matching, and simple subtraction for approximating the results of a RCT that evaluated competing medications (Xalatan and Xalacom) that lower intraocular pressure (mmHg). The RCT measurement in the Xalatan-to-Xalacom difference was +0.97 mmHg. Comparatively, the results for the approximation results and the p-value for the RCT-approximation difference were: linear regression – 0.86 mmHg (p=0.953), nearest neighbor propensity scoring- -1.56 mmHg (p=0.28), kernel matching propensity scoring – +0.79 mmHg (N/A), radius matching propensity scoring – +0.77 mmHg (p=0.915), stratification propensity scoring – +1.03 mmHg (p=0.974), and simple subtraction- -1.10 mmHg (p=0.000). It finds all methods excepting simple subtraction and nearest neighbor propensity score matching produce results that are not significantly different from the RCT result.

Each of these three papers contributes to CER by applying quantitative methods to relevant policy questions: how best to approximate RCT results when these studies are not available, what is the cost-effectiveness of vaccination programs that include adolescent males, and why might there be differences across the nation in how long patients at the end-of-life spend in hospice? Much attention is

centered on how CER will enhance healthcare delivery, improve health outcomes, and expand healthcare access. It is hoped that these papers will help contribute to these goals, and further substantively benefit the lives of everyday people.

Committee Page

Abdelmonem A. Afifi

Ronald Hays

Arturo Vargas Bustamante

Robert M. Kaplan, Chair

2013

Dedication Page

Completion of this dissertation would not have been possible without the support of two people who are very close to my heart. I dedicate this dissertation to my parents, Kusuma and Prasad Sunkara. They have been there for me through thick and thin, and I love them dearly. They have taught me about life in ways that I will carry deep within me for the rest of my days. Their fundamental message has been: *'We are always here for you no matter what'*. I hope in my life that I can be the same rock for others as they have been for me.

Thank you, Amma and Nanna.

Table of Contents

Acknowledgements

Thank you first to Professor Richard Brown. Thank you for believing in me, Dr. Brown. You are greatly missed, but are forever remembered. I hope I made you proud.

My committee members helped steer this ship safely to shore. Thank you. It would not have been possible without each of you – Dr. Kaplan, Dr. Afifi, Dr. Hays, and Dr. Vargas Bustamante.

Thank you to my sister, Haritha (Sunkara) Naga. Akka sent me a card after I passed my dissertation defense: *"Dearest Vasu, so proud to see you achieve all of your goals...looking forward to see what's next"*. The future is still an open book Akka, but I know that I cannot wait to share the good times with you, Arathi, Kushal, and Karun.

Sincere thanks to Carina Carriedo, for a friendship that means so much to me.

Thanks to Professor Mary Coombs and Dr. Paul Torrens. You shared your time and invested your energy in me. I am greatly appreciative.

Josie Wei and Jessica Shim were indefatigable. You both provided crucial help and timely advice at so many critical junctures. Thank you for being there for me, and lending a helping hands at the times when I needed it most.

Thank you to Dr. Jay Enoch, for being a wonderful mentor to me from 18 years old to now. You are one in a million, Dr. Enoch.

Biographical Sketch

Vasu Sunkara completed his B.A. in Development Studies and Biochemistry at UC Berkeley in 2003. While at Berkeley, he was a National Merit Scholar and an Alumni Scholar. As a President's Undergraduate Fellow, he conducted research with the supervision of Dr. Daniel McFadden at the UC Berkeley Department of Economics. He received his M.D. from the University of Wisconsin School of Medicine and Public Health in 2009. As a PhD student at UCLA, he was a recipient of the Jeffrey L. Hanson Distinguished Service Award from the UCLA Graduate Student Association. In 2012, he was supervised by Dr. Amartya Sen at the Harvard Department of Economics, where he completed a year of postdoctoral training. Currently, he is a Family Medicine resident in Erie, Pennsylvania.

Introduction

Introduction

Comparative effectiveness research (CER) provides perspective on what treatments or interventions are most likely to benefit patients. Knowing which treatments are more effective should lead to better patient outcomes, and a better use of available healthcare resources. CER practices have been increasingly used at a local and federal level in the United States. This paper combines three original papers on comparative effectiveness methodology and application. One paper applies established statistical techniques to evaluate between-study treatment differences. The other two papers apply CER techniques to understand healthcare phenomena, including the determinants of hospice stay duration and the difference in benefit between male-female versus female-only HPV vaccination.

CER: An Overview

The Agency for Healthcare Research and Quality (AHRQ) and the American College of Physicians (ACP) provide two revealing descriptions on the definition of CER. The AHRQ states, "[*T*]*he core question of comparative effectiveness research (is) which treatment works best, for whom, and under what circumstances*." (National Research Council, page 36, 2009). The ACP states, "*Comparative effectiveness analysis evaluates the relative (clinical) effectiveness, safety, and cost of two or more medical services, drugs, devices, therapies, or procedures used to treat the same condition*." (National Research Council, page 35, 2009).

CER is a descendent of effectiveness and outcomes research that developed in the 1980s and 1990s. Due to a substantial growth in health outcomes and clinical trial data during this period, there

was a strong interest in making clinical treatment decisions more data-driven.

In forecasting the future of the field, Paul Ellwood described a multifaceted area of "outcomes management" that would more actively connect medical decisions with their projected impact on patients: "*Outcomes management is a technology of patient experience designed to help patients, payers, and providers make rational medical care-related choices based on better insight into the effect of these choices on the patient's life"*(Selby, 2010). Later, in the form of *"patient outcome research teams (PORTs)*", David Blumenthal forecast a deep and unprecedented impact of outcomes research on medicine: "*[PORTs] represent the coming of age of health services research as a useful, clinically relevant discipline, whereby the evaluative sciences for the first time are, with the help and support of Congress, [are] going to become directly relevant to clinical decision making in a way that they haven't before*" (Selby, 2010).

The Medicare Modernization Act (MMA) of 2003 was a decisive step in transitioning this research from academic circles into actual practice. Section 1013 of the MMA ("Research on Outcomes of Health Care Items and Services") stated the provision of $50 million for the explicit purposes of supporting "*scientific information needs and priorities*" that would improve outcomes and effectiveness of specified programs associated with Medicare: "*The legislation authorizes and appropriates $50 million for FY 2004 for the Secretary through AHRQ to conduct research to address the scientific information needs and priorities related to improving outcomes, clinical effectiveness and appropriateness of specified health services and treatments including prescription drugs identified by the Medicare, Medicaid, and State Children Health Insurance Programs and to improving the efficiency and effectiveness of these Programs. The Secretary is required to establish a process for*

developing research priorities" (CMS, 2004).

The MMA was recognized as an important step forward in reconciling medical evidence with healthcare practice. In a January 2005 article in *Health Affairs,* Carolyn M. Clancy and Kelly Cronin described the role of the MMA in aligning the most effective evidence with ultimate Medicare practice: "*MMA Section 1013 requires the HHS secretary to set priorities and target areas where evidence is needed to improve the effectiveness of services delivered "* (Clancy and Cronin, 2005). The authors further emphasized the impact of the MMA on medical decision-making in general, long-term healthcare spending, and health outcomes: "*Such evidence has the potential to greatly reduce out-of-pocket and government spending for new drug benefits. Section 1013 of MMA thus sets the stage for explicitly incorporating decisionmakers' needs in setting priorities for the effectiveness of health care interventions.*" (Clancy and Cronin, 2005).

From 2000 to 2010, there was a rapid transition from the potential benefits of outcomes research to the implementation of this research to fill the gaps in the medical evidence base. A June 2003 article in the *New England Journal of Medicine* (NEJM) underscored the importance of these concerns by reporting that nearly 50% of medical practice does not conform to clinical guidelines. The authors concluded, "*These deficits, which pose serious threats to the health and well-being of the U.S. public, persist despite initiatives by both the federal government and private health care delivery systems to improve care*" (McGlynn et al, 2003). The Congressional Budget Office (CBO) reported in 2007 that, "[o]*nly a limited amount of evidence is available about which treatments work best for which patients*" (Nabel, 2009). A 2009 *Journal of the American Medical Association* (JAMA) article found that approximately 50% of clinical practice recommendations of the American College of Cardiology

and the American Heart Association were not evidence-based (Nabel, 2009).

CER was popularized during this period, and increasingly became seen as a remedy for gaps in medical care. Indeed, in a November 2006 article in *Health Affairs*, Gail Wilensky promoted a new center for comparative effectiveness research as a means to optimize decision-making: "*Interest in objective, credible comparative clinical effectiveness information has been growing in the United States, both by those who support competitive behavior in health care and by those who support administered pricing...Finding politically acceptable ways to reduce the long-term growth rate in health care spending will be difficult. Within this context, learning how to "spend smarter," rather than relying on arbitrary mechanisms to limit spending, begins to look very appealing*" (Wilensky, 2006).

CER has similarly gained prominence in Health Services Research (HSR). Academy Health, the leading HSR academic organization, has worked diligently to promote understanding of CER. It provides support for two major CER projects – the Electronic Data Methods (EDM) Forum and the Multi-Payer Claims Database (MPCD). The EDM Forum for Comparative Effectiveness Research is a Agency for Healthcare Research and Quality (AHRQ) funded initiative to develop an infrastructure for evaluating comparative effectiveness using electronic clinical data. The MPCD is a Center for Medicare and Medicaid Services (CMS) funded initiative that will help apply comparative effectiveness analysis to longitudinal CMS data (AcademyHealth, 2008-2012).

Outside of Academy Health, CER is an important topic for other entities involved in HSR. The Veterans Affairs Health Services Research and Development (HSR&D) is already positioning itself to become a leader in CER in the United States. Its extensive utilization and cost data, combined with its

existent electronic database infrastructure, prime it for making an impact in the area. As the Director of the HSR&D, Dr. Seth Eisen, states, "*For clinicians, patients, and policymakers alike...CER has the potential to provide much needed evidence-based criteria for health care decision-making*" (Veteran Affairs Health Services Research & Development Services, May 2009).

At a federal level, a capstone moment for the entire CER field was the passage of the American Recovery and Reinvestment Act (ARRA) in February 2009, and the American Affordable Care Act (ACA) in March 2010. ARRA allocated $1.1 billion to comparative effectiveness research by funding the National Institutes of Health ($400 million allocation), the Agency for Healthcare Research and Quality (AHRQ) ($400 million), and the Office of the Secretary of the Department of Health and Human Services ($300 million allocation). This legislation also established the (now defunct) Federal Coordinating Council for Comparative Effectiveness Research (HHS, 2009).

The dual CER goals in the ARRA legislation were to: 1) "*Conduct, support, or synthesize research that compares the clinical outcomes, effectiveness, and appropriateness of items, services, and procedures that are used to prevent, diagnose, or treat diseases, disorders, and other health conditions*.", and 2) "*Encourage the development and use of clinical registries, clinical data networks, and other forms of electronic health data that can be used to generate or obtain outcomes data*." (HHS, 2009).

The passage of the ACA was a watershed moment for the American healthcare system, and CER. The legislation established the Patient-Centered Outcomes Research Institute (PCORI) to oversee high priority topics in CER. The objective of PCORI is to:"a*ssist patients, clinicians, purchasers, and*

policy-makers in making informed health decisions by advancing the quality and relevance of evidence concerning the manner in which diseases, disorders, and other health conditions can effectively and appropriately be prevented, diagnosed, treated, monitored, and managed through research and evidence synthesis." (Selby et al, 2012). With substantial funding and a wide mandate, the Institute has been described by Dr. Francis Collins, the Director of the National Institutes of Health (NIH), as the "*most significant thing*" in the ACA (Saslow, 2010).

Hospice Length-of-Stays in the United States

Using Medicare data, Fisher and Wennberg helped demonstrate the magnitude of healthcare variation in the United States through their studies with the Dartmouth Atlas (Welch et al, 1993). Hospice care is one important area of American healthcare that has wide variations of usage across the country. While the variations in length-of-stay have been studied in the literature, the drivers of this variation remain poorly understood. Better characterizing the primary drivers of hospice utilization will increase understanding on both the provider and patient side on how best to meet the patient's wishes of whether to enroll in hospice once it is determined that they qualify.

The first dissertation paper evaluates the main hospice-stay variable through kernel density histogram plotting. Thereafter, correlations between hospice stay and potential predictor variables are estimated – including proxies for healthcare reimbursement, quality-of-care indicators, and the number of providers. Then, a stepwise selection process is conducted to select a subset of significant predictor variables. An Ordinary Least Squares (OLS) model is specified to identify which of the selected variables from the Dartmouth Atlas have a significant association with hospice length-of-stay. Secondary testing of the OLS model is done to correct for autocorrelation between predictors, and to

check whether the model meets the homogeneity requirements of OLS.

This paper connects with the theme of CER because it evaluates the factors affecting differential utilization of a healthcare service. The utilization difference between different hospital referral regions (HRRs) warrants attention to how end-of-life care is practiced across the United States, This paper evaluates differential uptake of hospice services by patients in the end-of-life within the United States, . It further seeks to identify factors that might explain how these differences in the length of stay are affected.

The HPV Vaccine for Men: Evaluating Effectiveness through Past Studies

Studies of male HPV vaccination have consistently challenged whether it is cost-effective. As the number of vaccinated women increases, these studies have shown the incremental cost-effectiveness ratio (ICER) outside the $100,000 per QALY cost-effectiveness threshold. Despite this trend, male HPV vaccination has been actively promoted as an important public health intervention. Citing low female HPV vaccination coverage rates in the United States, the Advisory Committee on Immunization Practice (ACIP) stated that male HPV vaccination was needed. In October 2011, it recommended universal HPV vaccination of all adolescent males, starting at 12 years old (CDC, 2011; *NYTimes*, Oct. 28, 2011; *NYTimes*, Oct. 25, 2011) .

Given the ACIP policy recommendation, it is even more critical to understand the circumstances in which male HPV vaccination is and is not cost-effective. There are only a few systematic reviews of HPV vaccination in males. To the author's knowledge, there is not a single published meta-analysis. This is an important gap in the literature that should be resolved.

This paper compiles a selection of cost-effectiveness papers evaluating male HPV vaccination. It presents the relevant input and ICER values. It then provides a weighed ICER value based on the inverse variance meta-analysis method. Heterogeneity testing is performed to determine if the ICER values satisfy the assumption of being derived from the same sample.

CER and Observational Data: Statistical Considerations

The randomized clinical trial (RCT) is the best accepted method for determining the efficacy of a treatment. The double-blinded randomized allocation of subjects to separate arms of a study minimizes the threat of selection and observer bias. However, due to time and resource-limitations, conducting an RCT may be impractical. In these situations, observational data is the next best option. As D'Agostino and D'Agostino state, "*Observational, nonrandomized studies have a role when RCTs are not available, and, even when RCTs are available, to quantify effectiveness and other real world experiences.*" (D'Agostino and D'Agostino, 2007)

The challenge is that observational data is more likely to have biases that affect the treatment results. Controlling for these biases is an important means of calculating reliable CER treatment differences. As D'Agostino and D'Agostino conclude, "*There are many approaches for making statistical inferences from observational data. Some approaches focus on study design, others on statistical techniques. However, even with the best of designs, observational studies, unlike the RCTs, do not automatically control for selection biases. Therefore, statistical methods involving matching, stratification, and/or covariance adjustment are needed.*" (D'Agostino and D'Agostino, 2007). Lalonde and Dehejia and Wahba have evaluated the robustness of different statistical methods in correcting

biases in observational labor economics data (LaLonde, 1986; Dehejia and Wahba, 1999).

LaLonde determined that major limitations to the use of non-experimental methods included specification biases, gender-based differences in effect size, and significant differences in overall effect size. Dehejia and Wahba found that propensity score matching applied to non-experimental data could approximate experimental results under restricted conditions where the correct propensity score matching algorithm is selected, when there are a sufficient number of similar treatment and comparison units in the study, and whether there is selection bias occuring amongst observed covariates. Stukel and co-authors have also applied different statistical techniques to determine which one minimizes selection bias in estimates of the association between cardiac cathertization and long-term acute myocardial infarction (AMI) mortality. The authors found that instrumental variables was able to better account for unobserved variables compared to propensity score and multivariate risk techniques in assessing associations between catheterization and AMI mortality (Stukel et al, 2007).

The statistics paper in the dissertation seeks to contribute to CER by determining if statistical methods may be used to merge data across studies to obtain a relative treatment difference. There are few studies that seek to determine if there are circumstances where data from completely different datasets may be applied comparatively to obtain treatment results similar to an RCT. This study selects three papers evaluating the impact of two anti-glaucoma medications, Xalatan and Xalacom, on lowering intraocular pressure (IOP). Linear regression, propensity score matching, and simple subtraction are applied to two datasets. The calculated treatment difference using these methods is then compared to the treatment difference for the RCT, where the RCT has head-to-head testing of the two medications.

Conclusion

The hospice length-of-stay paper studies the determinants of hospice stay across the United States. The HPV meta-analysis paper calculates an aggregate cost-effectiveness value for male and female HPV vaccination using data from the United States and internationally. The accuracy of different statistical methods to proxy RCT results is studied in the statistics paper. The hospice length-of-stay paper uses multiple regression techniques, while the HPV meta-analysis applies geometric and arithmetic means to calculate weighted and unweighted ICER values. The statistics paper compares the accuracy of linear regression, propensity score matching, and simple subtraction in approximating the results of an RCT. Furthermore, each paper has a clear comparison: the hospice length-of-stay paper compares hospice stays across HRRs in the United States, the HPV meta-analysis paper compares the male-female and female-only ICER values across 7 published papers, and the statistics paper compares three different methods to approximate RCT results.

The three papers advance the field of CER. The hospice-stay paper uses well-established data to help identify important determinants of hospice length-of-stay. For families and patients struggling with end-of-life questions, this data provides guidance on what factors affect hospice placement. For policy-makers comparing ICU versus hospice stays for end-of-life patients, the paper identifies factors that may tilt the choice in one direction versus another. Aggregating values across multiple papers, the meta-analysis provides a deeper understanding of the cost-effectiveness of male-female HPV vaccination.

Substantial resources are being invested to vaccinate young male adolescents, but the

effectiveness of this intervention varies based on the paper one cites. Instead of allowing for this confusion, this paper provides a clearer idea of whether vaccination is indeed worthwhile. Finally, with the financial costs and time investments needed to complete a randomized clinical trial, it is understandable that researchers seek alternative methods to analyze existent observational data. The statistics paper provides further guidance on statistical techniques that researchers can use to get results when an RCT is not available.

In combination, these three papers help patients, researchers, and policy-makers get a clearer idea of the impact of important health interventions, and on how to measure such interventions when an RCT has not been done. In this way, there are practical benefits to each paper that others can extend through future work. In the process, CER can continue its promising rise in the areas of American and global healthcare policy.

References

Academy Health. "Projects & Initiatives: Comparative Effectiveness Research". Academy Health website. 2008-2012. Website: http://www.academyhealth.org/content.cfm?ItemNumber=2841&navItemNumber=2933

CDC. "Press Briefing Transcript: ACIP recommends all 11-12 year old males get vaccinated against HPV". Tuesday, October 25, 2011. Website: http://www.cdc.gov/media/releases/2011/t1025_hpv_12yroldvaccine.html

Clancy CM, Cronin K. "Evidence-Based Decision Making: Global Evidence, Local Decisions". *Health* Affairs. January 2005. Vol. 24. No. 1 (151-162). Website: http://content.healthaffairs.org/content/24/1/151.full?sid=527c7b53-183d-44fd-b471-13dd6178b8de

CMS. “CMS Legislative Summary: Medicare Prescription Drug, Improvement, and Modernization Act of 2003, Public Law 108-173”. Center for Medicare and Medicaid Services Office of Legistlation. April 15, 2004. Website: https://www.cms.gov/MMAUpdate/downloads/PL108-173summary.pdf

D'Agostino RB Jr, D'Agostino RB Sr. “Estimating treatment effects using observational data". *JAMA*. 2007 Jan 17;297(3):314-6. Website: http://jama.ama-assn.org/content/297/3/314.full

Dehejia RH, Wahba S. “Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation Training Programs”. *Journal of the American Statistical Association*. December 1999. Vol.94 (448): 1053-62. Website: http://www.jstor.org/stable/2669919

Department of Health and Human Services (HHS). “Comparative Effectiveness Research Funding”. U.S. Department of Health and Human Services Recovery Programs. Website: http://www.hhs.gov/recovery/programs/cer/

Harris G. “Panel Recommends Vaccination for Boys of 11”. New York Times. October 25, 2011. Website: http://www.nytimes.com/2011/10/26/health/policy/26vaccine.html?_r=1

LaLonde RJ. “Evaluating the Econometric Evaluations of Training Programs with Experimental Data”. *American Economic Review*. September 1986. Vol.76, No. 4, pp. 604-620. Website: www.jstor.org/stable/1806062

McGlynn EA, Asch SM, Adams J, Keesey J, Hicks J, DeCristofaro A, Kerr EA. “The quality of health care delivered to adults in the United States". *N Engl J Med*. 2003 Jun 26;348(26):2635-45. Website: http://www.nejm.org/doi/full/10.1056/NEJMsa022615#t=article

Nabel EG. “Comparative Effectiveness Research: National Institutes of Health”. PCAST Session on Health Reform and CER. August 6, 2009. Website: http://www.whitehouse.gov/files/documents/ostp/PCAST/Nabel%20Presentation.pdf

National Research Council. "What Is Comparative Effectiveness Research?." *Initial National Priorities*

for Comparative Effectiveness Research. Washington, DC: The National Academies Press, 2009. Website: http://books.nap.edu/openbook.php?record_id=12648&page=35

New York Times Editorial Board. “Editorial: For Their Own Good”. *New York Times*. October 28, 2011. Website: http://www.nytimes.com/2011/10/29/opinion/the-hpv-vaccine-is-for-their-own-good.html

Saslow R. “Conversations: NIH Director Francis S. Collins”. *Washingtonpost.com*. March 24, 2010. Website: http://www.washingtonpost.com/wp-dyn/content/article/2010/03/23/AR2010032304094.html

Selby JV, Beal AC, Frank L. “The Patient-Centered Outcomes Research Institute (PCORI) national priorities for research and initial research agenda". *JAMA*. 2012 Apr 18;307(15):1583-4.

Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vermeulen MJ. “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*. 2007 Jan 17;297(3):278-85. Website: http://www.ncbi.nlm.nih.gov/pubmed/17227979

Veteran Affairs Health Services Research & Development Services. *HSR&D FORUM*. May 2009. Website: http://www.academyhealth.org/files/publications/forum/VAFORUM0509.pdf

Welch WP, Miller ME, Welch HG, Fisher ES, Wennberg JE. "Geographic variation in expenditures for physicians' services in the United States". *N Engl J Med*. 1993 Mar 4;328(9):621-7.

Wilensky G. “Developing A Center For Comparative Effectiveness Information”. *Health Affairs*. November 2006. Vol. 25. No. 6 (572-585). Website: http://content.healthaffairs.org/content/25/6/w572.full#R1

Determinants of Hospice Stay Duration in the United States

Background

Medical care received at the end of life is a major driver of health care costs (Hogan et al, 2001). One factor that makes end-of-life care so expensive is the use of intensive care unit (ICU) resources, often in cases where the care does not have a substantial impact on the ultimate patient outcome. In lieu of staying in the ICU, hospice care is an alternative for patients nearing the end of their lives. Hospice focuses on patient comfort when medical treatment is unlikely to extend life or improve the patient's quality-of-life. While hospice is an option for many patients across the United States, its use varies based on different geographic and hospital-specific factors (Kaplan, 2011).

A variety of studies have attempted to determine what predisposing factors affect the use of hospice services (Goldsmith et al, 2008; O'Hare et al, 2010; Kelley et al, 2011). Goldsmith and co-authors found that hospital not-for-profit status, specific geographic location, medical school affiliation, and hospice ownership increased the likelihood of a hospital having an in-house palliative care program. In particular, the West and Midwest were more likely to have a palliative care program. Greater use of hospice care was associated with fewer days in the hospital ICU, less total inpatient care, and lower overall Medicare spending (Goldsmith et al, 2008). The amount of care from inpatient specialty physicians and in the ICU also seem to play a role in whether a patient elects for hospice services or not. O'Hare and co-authors found that patients living in areas with greater availability of hospital intensive care services were more likely to die in a hospital setting compared to those with lower access to intensive care (O'Hare et al, 2010).

Besides an investigation of the ultimate decision of entering hospice care at the end-of-life, research has further been done on the factors affecting the duration of hospice stays. Using data from

the Health and Retirement Study, Kelley and co-authors evaluated a host of disease-specific, demographic, and geographic factors influencing the number of days spent in hospice versus an inpatient hospital setting. They found that patients with dementia spent 3.02 fewer days in the hospital, while those with diabetes and chronic kidney disease had 2.37 and 2.40 additional hospital days, respectively. Patients living closer to family stayed 1.62 fewer days in a hospital, while those already enrolled in hospice stayed 1.88 fewer days in the hospital. Death in an inpatient setting was not associated with a patient's gender, educational level, completion of an advance directive, or cumulative net worth (Kelley et al, 2011).

Quality of care also seems to play an intriguing role in the area of end-of-life care. In a comparison of cost of end-of-life care in San Diego versus Los Angeles, Kaplan found that Los Angeles incurred higher costs and intensity of care for its patients compared to San Diego, but that San Diego scored better than Los Angeles on quality-of-care measures. This work helps extend the connection between quality-of-care and healthcare costs to the arena of end-of-life care (Kaplan, 2011). It aligns with the work reported in the Dartmouth Atlas linking quality-of-care to healthcare costs using Medicare data. Two major reports on hospice care have been published by the Dartmouth group: Trends and Variations in End-of-Life Care for Medicare Beneficiaries with Severe Chronic Disease and Quality of End-of-Life Cancer Care for Medicare Beneficiaries: Regional and Hospital-Specific Analyses (Goodman et al, 2011; Goodman et al, 2010). These two reports evaluate the patterns of end-of-life care, describing the locations where hospice stays are high versus low. The work echoed much of the past literature on the topic. Enrollment into hospice care varied greatly between hospital referral regions and between academic and non-academic hospitals. Patients in hospitals during the last days of life received a greater amount of life-supporting interventions such as endotracheal intubation and cardiopulmonary resuscitation (CPR) (Goodman et al, 2010).

Hospice stay duration has typically been evaluated using disease specific comparisons or by comparing rates between cities. However, it has not been evaluated across hospital referral regions in the United States. This is an important task, since factors affecting hospice stays at a local level may not be applicable at a national level. Using data from the Dartmouth Atlas, the current paper seeks to assess what factors affect the duration of hospice stays in the United States. This paper evaluates the association between hospice duration and total Medicare costs to determine whether costs do, in fact, go down with longer hospice stays.

Methods

The study used data obtained from the 2003-2007 Dartmouth Atlas of Healthcare (Dartmouth Atlas, 2012). These years were selected because they were the most recent years available as of 2012. Hospital Referral Regions (HRR) represent the 308 different urban and rural localities across the United States. HRR is defined as a collection of zip codes grouped together based on the major referring hospitals in the area (Dartmouth Atlas, 2012; Blue Cross/Blue Shield Website). More detail on the methods for forming HRRs can be found in the Research Methods section of the Dartmouth Atlas (Dartmouth Atlas, 2012). The primary outcome variable was the number of days spent in inpatient hospice care during the last 6 months of life. We also considered the reimbursement type, supply of providers, and quality ratings from the Center for Medicare and Medicaid Services (CMS). The data were downloaded from the Dartmouth Atlas website. For each variable, all the HRRs were selected. The data were analyzed using STATA 9.1.

The predictor variables were:

- Number of inpatient hospital days for *Medicare beneficiaries between the ages of 67 to 99 years old who were admitted to a hospital for a chronic condition, and died within 2 years of the*

admission (Dartmouth Atlas, 2012). These patients were enrolled in Medicare A and B, did not have managed care coverage, and had died within 2 years of the start time of measurement.

- Medicare reimbursements for inpatient hospital short stays - Reimbursement amount drawn from the CMHS inpatient short stay code E (Dartmouth Atlas, 2012).

- Medicare reimbursements for inpatient hospital long stays - *Data came from the Continuous Medicare History Sample (CMHS) long stay inpatient trailer (code D) for Medicare A and B beneficiaries ages 65-99 years old* (Dartmouth Atlas, 2012).

- Outpatient Medicare reimbursements - *Medicare reimbursements derived from the CMHS outpatient trailer, for patients aged 65 to 99 years old that were enrolled in both Medicare A and B* (Dartmouth Atlas, 2012).

- Total Medicare reimbursement - *Total Medicare spending (MedPAR, Home Health, Hospice, DME, Part B, and Outpatient) for patients who died within 2 years of a hospital admission for a chronic condition* (Dartmouth Atlas, 2012).

- Medicare reimbursement per enrollee - *Total reimbursement amount for both Medicare A and B services for enrollees aged 65 to 99 years old* (Dartmouth Atlas, 2012).

- Physician per-visit fee - *Physician reimbursements for evaluation and management (E&M) and consultations for patients ages 67-99 years old admitted for a chronic condition and died within 2 years of the admission* (Dartmouth Atlas, 2012).

- Number of total physician visits - *All specialist and primary care visits for patients who died within 2 years of an inpatient admission for a chronic condition* (Dartmouth Atlas, 2012).

- Center for Medicare and Medicaid Services (CMS) hospital quality score - *Hospital quality score derived from the CMS Hospital Compare program* (Dartmouth Atlas, 2012).

- Number of hospital-based physicians per 100,000 people - *Number of hospital-based physicians in Anesthesiology, Pathology, and Radiology divided by 100,000 people, as determined by the U.S. Census* (Dartmouth Atlas, 2012).

- Number of hospital-based nurses per 100,000 people - *The number of full-time equivalent (FTE) nurses working in a hospital setting divided by 1,000 people as tabulated by the U.S. Census* (Dartmouth Atlas, 2012).

Analysis

The dependent variable, the log transformation of hospice days, was evaluated across all HRR's. The log transformation was applied because the hospice variable had a strong right skew. The transformation helped reduce the skewness in the distribution (from 0.866 to -0.174) and the kurtosis (from 3.93 to 3.15). Bivariate correlations between the log hospice variable and the selected predictors was assessed. This provided an initial perspective on the association between the dependent and independent variables. Thereafter, a multiple linear regression model was specified. The specification process was done by a backward-forward stepwise algorithm. All of the independent variables were initially included in the model. Only total Medicare reimbursements was a mandatory predictor in the loghospice final model. Otherwise, variables were selected into the model if they met the inclusion criteria of $p<0.05$ and the exclusion criteria of $p>0.10$.

After the model was specified, the magnitude and p-value of each selected independent variable on hospice duration was assessed. The model was further assessed for regression diagnostics. Collinearity of selected variables was evaluated by the Variance inflation factor (VIF). Homoskedasticity of the regression residuals was assessed by the Breusch-Pagan/Cook-Weisberg and Cameron-Trivedi tests.

Results

Figure 1 is a map from the Dartmouth Atlas that shows the number of days individuals were in hospice care for the last 6 months of their lives. The data is for hospital referral regions (HRR) in the United States, and has a range of hospice stays between 4.9 and 35.5 days. There are scattered areas in the Pacific Northwest, North Central states, and the Northeast with no data. Noticeably higher hospice stays are observed in the South and Southwest compared to the West and Northeast.

Figure 2 also comes from the Dartmouth Atlas and is a map of the total Medicare reimbursements in the last 6 months of life across the United States. This map has substantially more areas where there are no data compared to Figure 1, particularly in the North Central states and the West. There are also fewer areas of especially high reimbursement rates – excepting California, the southeast of Texas, Alaska, and a portion of the Northeast. On initial assessment, it appears that areas with higher hospice stays in Figure 1 are associated with lower total Medicare reimbursements in Figure 2.

Figure 3 shows a histogram of the dependent variable, hospice days, while Table 1 provides the summary statistics for the hospice days and log hospice days variables. Kernel density plotting was used in addition to having the histogram boxes. The hospice days variable had values between 4.9 days and 35.5 days. The reduction in the skewness and kurtosis after log transformation is noted in the table.

Table 2 shows the bivariate correlations between the different variables. The correlations only show associations between variables. Regression analysis yielded p-values, and the significance of selected predictors on loghospice after controlling for other variables. As might be expected, there are negative correlations between loghospice and inpatient days (r=-0.250) and loghospice and inpatient short-stay reimbursements (r= -0.255). Hospice reimbursements was positively correlated with loghospice (r=0.794). Total medicare reimbursements was negatively correlated (r=-0.137) with

loghospice. As the number of loghospice days go up, the number of total physician visits go down (r=-0.141). More inpatient days were associated with lower outpatient reimbursements (r=-0.191).

Evaluating key bivariate relationships more closely, Figure 4 shows a positive curvilinear association between increasing hospice reimbursement and the log of hospice days. Comparatively, Figure 5 shows that outside of a negative association at low outpatient reimbursement levels, there appears to be no association between increasing outpatient reimbursement and the length of inpatient stays for reimbursement exceeding $250. Finally, Figure 6 shows a positive linear relationship between increasing physician visits and total inpatient days.

The regression model shows that for every 1 day increase in the inpatient stay, there is a -0.035 change in the log (hospice days) (significant at $p<0.01$). A 1 unit increase in the ratio of nurses per 100,000 people has a 0.061 increase in the log (hospice days) (significant at $p<0.01$). A change in total Medicare reimbursements has a beta-coefficient near 0 on the log (hospice days) (while being significant at $p<0.05$).

The variance inflation factor (VIF) was calculated to identify collinearity between variables. All the predictors had a VIF substantially less than 10, indicating lack of significant collinearity (UCLA ATS, 2012).[1] As a result, no variables were removed from the regression model.

Both the Cameron-Trivedi and the Breusch-Pagan tests statistically evaluate the assumption that the residuals from a regression model are homoskedastic. Table 5 shows that the homoskedasticity null hypothesis for the Cameron-Trivedi test is rejected with a p-value of 0.000. In contrast, Table 6 demonstrates that the homoskedasticity null hypothesis for the Breusch-Pagan test is not rejected. Since the two tests are contradicting each other, rejection of the homoskedasticity assumption is indeterminate. While no changes are made to the regression model, it is appropriate to be concerned

[1]

about the true significance of the results.

Table 7 evaluates the reverse association between hospice days on total Medicare reimbursements. Total Medical reimbursements is the dependent variable, while hospice days (untransformed) is the single predictor. The model results show that a 1 day increase in hospice stay decreases the total Medicare reimbursement by approximately $117. The association is significant at $p<0.05$, but the overall model has a low R-squared of 0.014.

Discussion

This paper provides evidence that total Medicare reimbursements *a priori* do not affect the hospice duration for an end-of-life patient. Total Medicare reimbursement includes multiple components, including outpatient, inpatient, and hospice reimbursements. Hence, it is reasonable to infer that the directional association of Medicare reimbursements on hospice duration will be non-significant. However, evaluating the *a posteriori* relationship shows a connection between increased hospice length-of-stay and lower total Medicare reimbursements. The $117 per day cost savings is substantial, and is in line with previous research by Goldsmith and co-authors showing lower Medicare costs with greater use of palliative care services. This research extends The Goldsmith et al finding across over 300 hospital referral regions in the United States.

After controlling for other variables, regression analysis shows that a higher number of hospital-based nurses is associated with a greater number of days the patient spends in hospice in the same communities. This is an unexpected finding since past work has emphasized the impact of ICU providers on reducing hospice length-of-stay. It is possible that this is a spurious relationship because patients in hospice have no direct connection with inpatient nurses.

However, it would be important to rule-out whether regions with a high proportion of inpatient nurses also facilitate patients entering hospice. In that case, it would be expected that the larger number of patients in hospice within these areas are being transferred from hospitals with a higher proportion of nurses. If this relationship is true, the relationship between inpatient nursing and hospice use may be more plausible. Nurses serve important roles in facilitating communication between patients, their families, and medical doctors. It is unclear whether nurses transmit patient preferences more directly to medical doctors so that doctors discontinue ICU care, or if nurses help facilitate better direct communication between patients and doctors. However the mechanism, further study of this pathway may be especially useful in understanding how hospice care decisions are made. Besides this finding, the research shows that other forms of reimbursement (such as hospice reimbursements and inpatient long-stay reimbursements) do not have a sizable effect on hospice length-of-stay.

The results also show that the higher the number of physician visits, the higher the number of inpatient days. This finding is potentially confounded by the fact that patients staying for longer inpatient stays have more specialists attending to them in the hospital. HRRs with fewer physician visits might be smaller hospitals without substantial inpatient intensive care unit (ICU) care. Stratifying by hospital size and hospital type (academic versus non-academic) might help elucidate this relationship. Nevertheless, it is noted that there is a small but significant correlation of -0.174 between total physician visits and log (hospice days).

This research extends previous work through the inclusion of data points from across the United States. The predictors cover such relatively diverse areas as reimbursements, quality-of-care, and provider supply. Another strength is that the data were collected over a longitudinal time-frame from 2004 to 2007. This increases the reliability of the data because more data points are available. Results due to exogenous factors particular to a single year are less likely to have an effect when multiple years

are considered.

There are also important limitations of this study. For example, the method does not assess other pertinent factors such as family influence or the effect of geographic regions on hospice days. Further assessment by regression of interactions between predictors could have been done. A main effects analysis was done instead because the further studying of interactions may add to spurious findings. Finally, the research evaluates hospital referral region, which might not be as good a unit of measurement as state-level data. The Dartmouth atlas uses aggregate data based on a random sampling of Medicare enrollees. It might not be considered generalizable for non-Medicare patients, or when evaluating care at a community level. Furthermore, error bounds are not available, so it is difficult to assess the precision of different data points. In addition, Dartmouth data draw from Medicare, is more quantitative in nature, and has little quality data from patients or providers.

Future research should evaluate the connection between hospice length-of-stay and total Medicare reimbursements. More details on how this mechanism acts would be helpful in identifying policies that serve patient interests as well as affect healthcare costs. Finally, work that evaluates interactions between different salient predictors will allow for a more sophisticated analysis of how patients are sorted into hospice care or acute inpatient care. In total, these additional areas of investigation will help promote a better understanding of how patients and providers make the critical choice of staying in a hospital at the end-of-life, versus being in a palliative setting. Ensuring that patients are making the best decision for themselves at this time is something that is critically important.

References

Goldsmith B, Dietrich J, Du Q, Morrison RS. “Variability in access to hospital palliative care in the United States." *J Palliat Med.* 2008 Oct;11(8):1094-102

Goodman DC, Fisher ES, Chuang C, Morden NE, Jacobson JO, Murray K, and Miesfeldt S. Quality of End-of-Life Cancer Care for Medicare Beneficiaries: Regional and Hospital-Specific Analyses. Dartmouth Institute for Health Policy and Clinical Practice. November 16, 2010. Website: http://www.dartmouthatlas.org/downloads/reports/Cancer_report_11_16_10.pdf

Hogan C, Lunney J, Gabel J, Lynn J. "Medicare beneficiaries' costs of care in the last year of life". *Health Aff (Millwood)*. 2001 Jul-Aug;20(4):188-95.

Kaplan RM. "Variation between end-of-life health care costs in Los Angeles and San Diego: why are they so different?" *J Palliat Med*. 2011 Feb;14(2):215-20. Erratum in: *J Palliat Med*. 2011 Sep;14(9):1081

Kelley AS, Ettner SL, Wenger NS, Sarkisian CA. "Determinants of death in the hospital among older adults". *J Am Geriatr Soc*. 2011 Dec;59(12):2321-5.

O'Hare AM, Rodriguez RA, Hailpern SM, Larson EB, Kurella Tamura M. "Regional variation in health care intensity and treatment practices for end-stage renal disease in older adults". *JAMA*. 2010 Jul 14;304(2):180-6.

UCLA Academic Technology Services (ATS). "Regression with STATA: Chapter 2 – Regression Diagnostics". Accessed October 28, 2012. Website: http://www.ats.ucla.edu/STAT/mult_pkg/faq/general/citingats.htm

The Dartmouth Atlas of Health Care. 2012. Website: http://www.dartmouthatlas.org

The Dartmouth Atlas of Health Care. "Research Methods". 2012. Website: http://www.dartmouthatlas.org/downloads/methods/research_methods.pdf

Figure 1. Duration of Hospice Stays Mapped across the United States

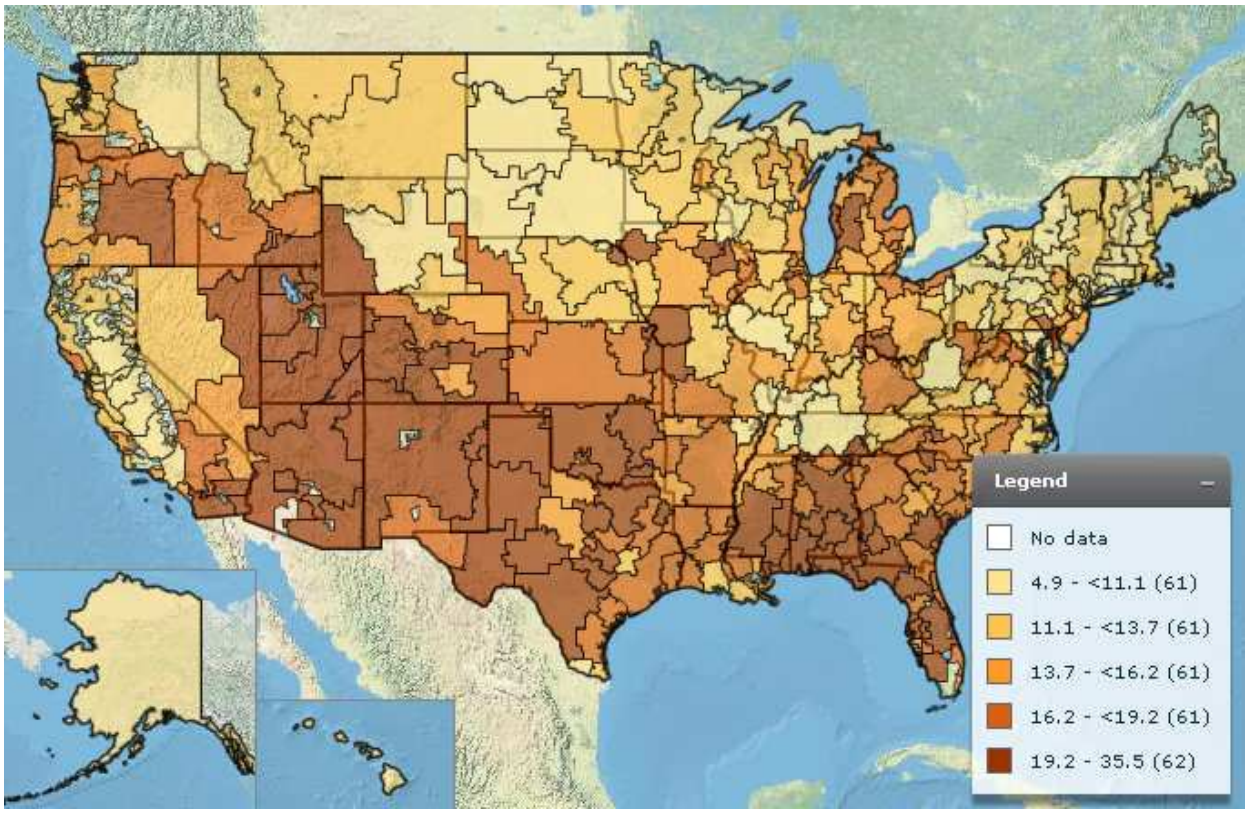

Figure 2. Total Medicare Reimbursements across the United States in the Last 6 Months of Life

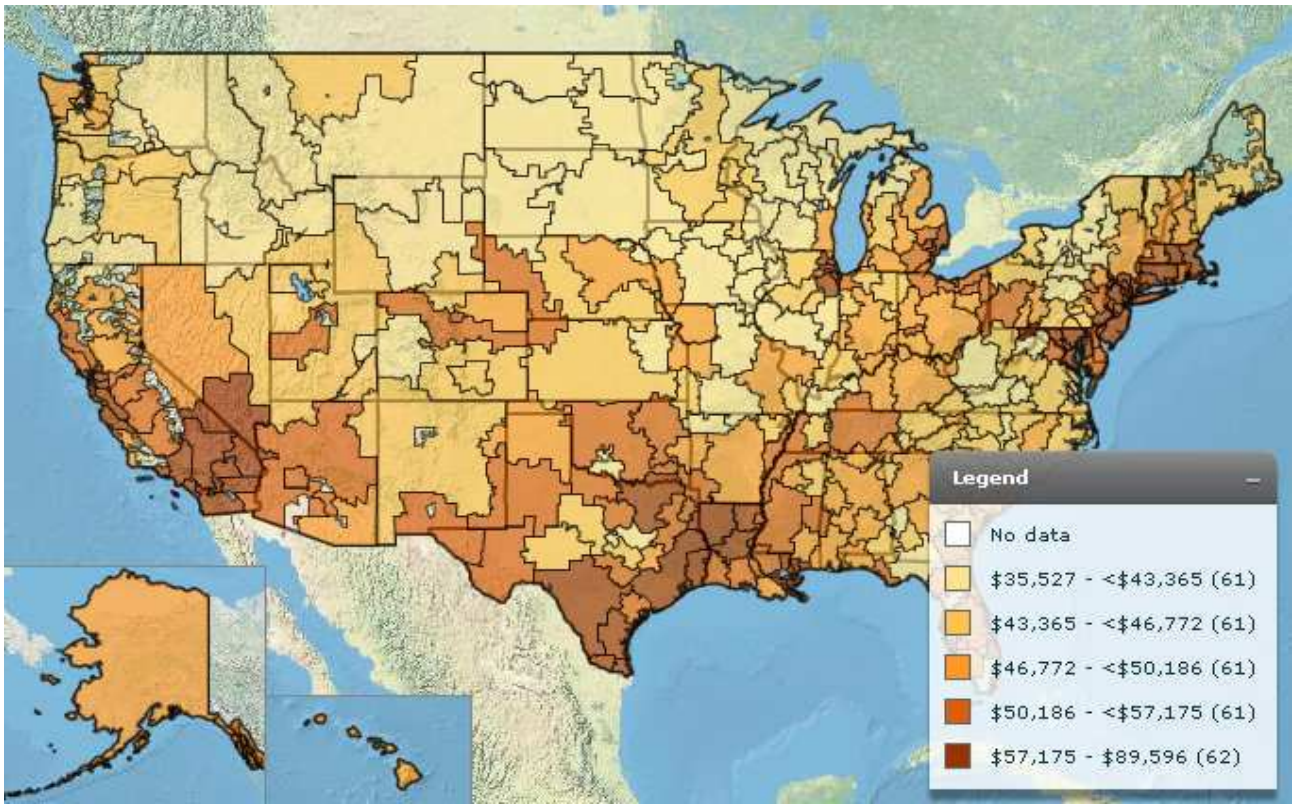

Figure 3. Histogram of Hospice Days with Kernel Density Plotting (Epanechnikov Option)

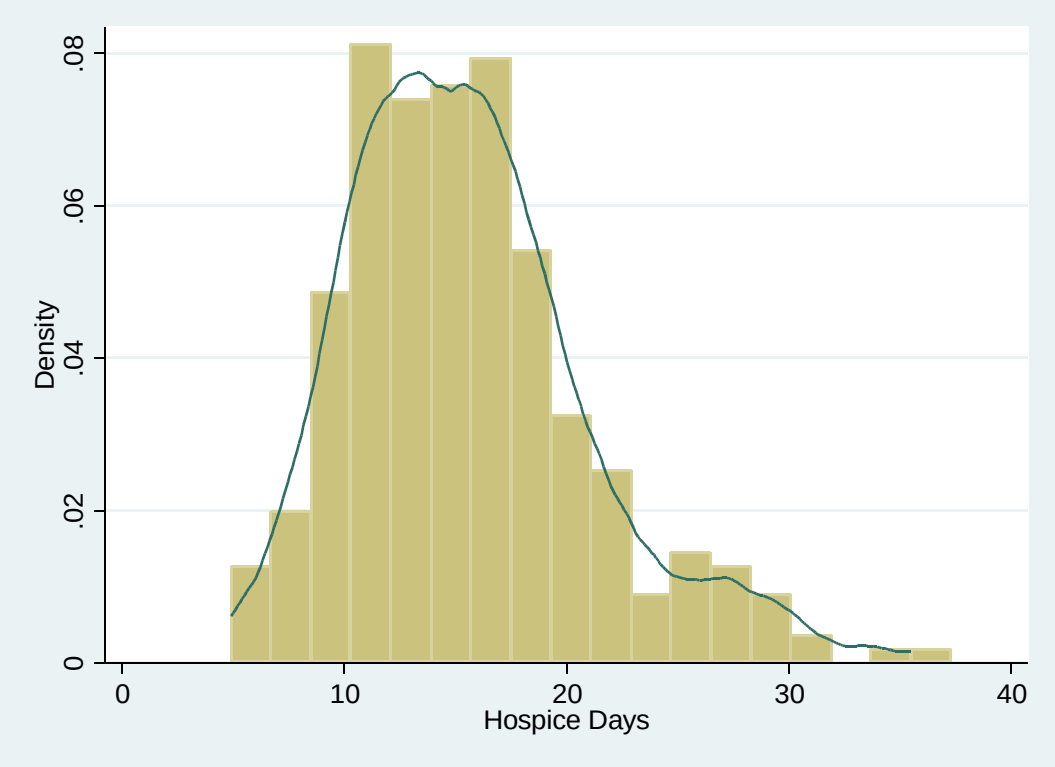

Figure 4. Scatterplot of Log(Hospice Days) vs Hospice Reimbursement (n=307)

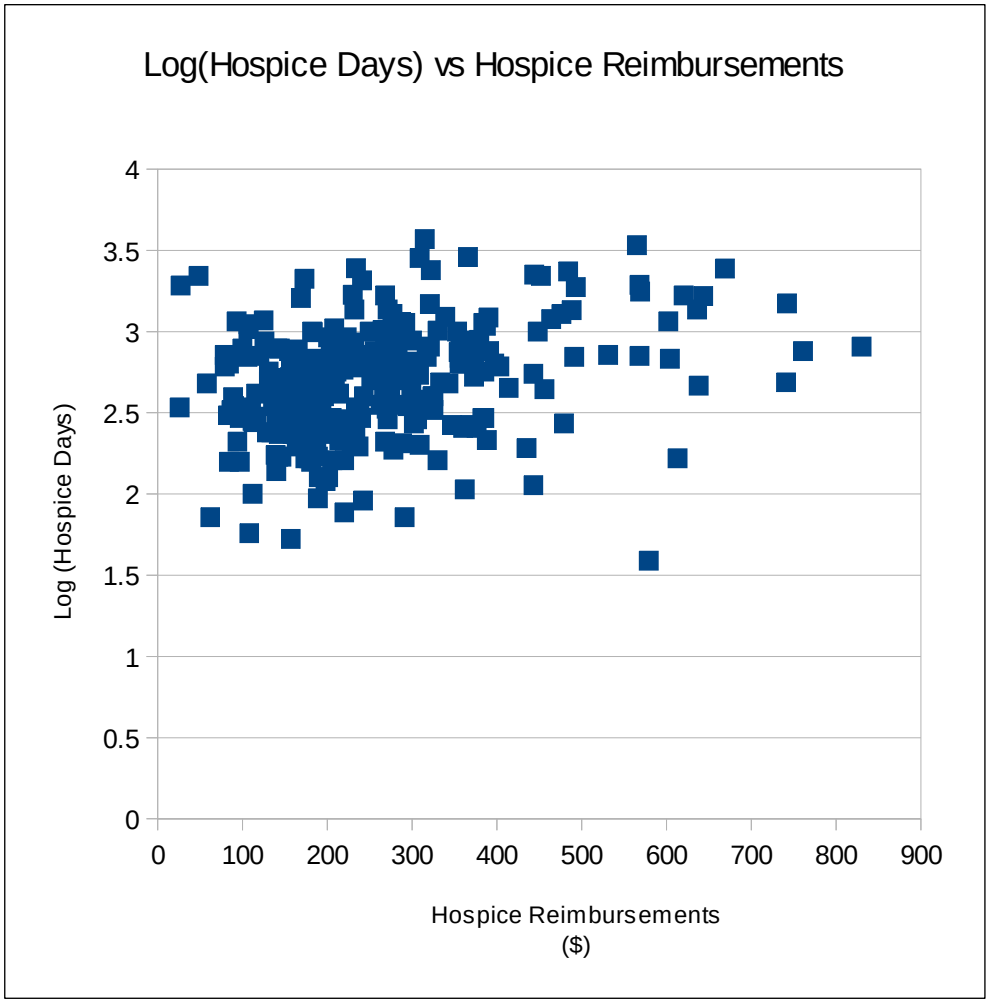

Figure 5. Scatterplot of Inpatient Days vs Outpatient Reimbursement (n=307)

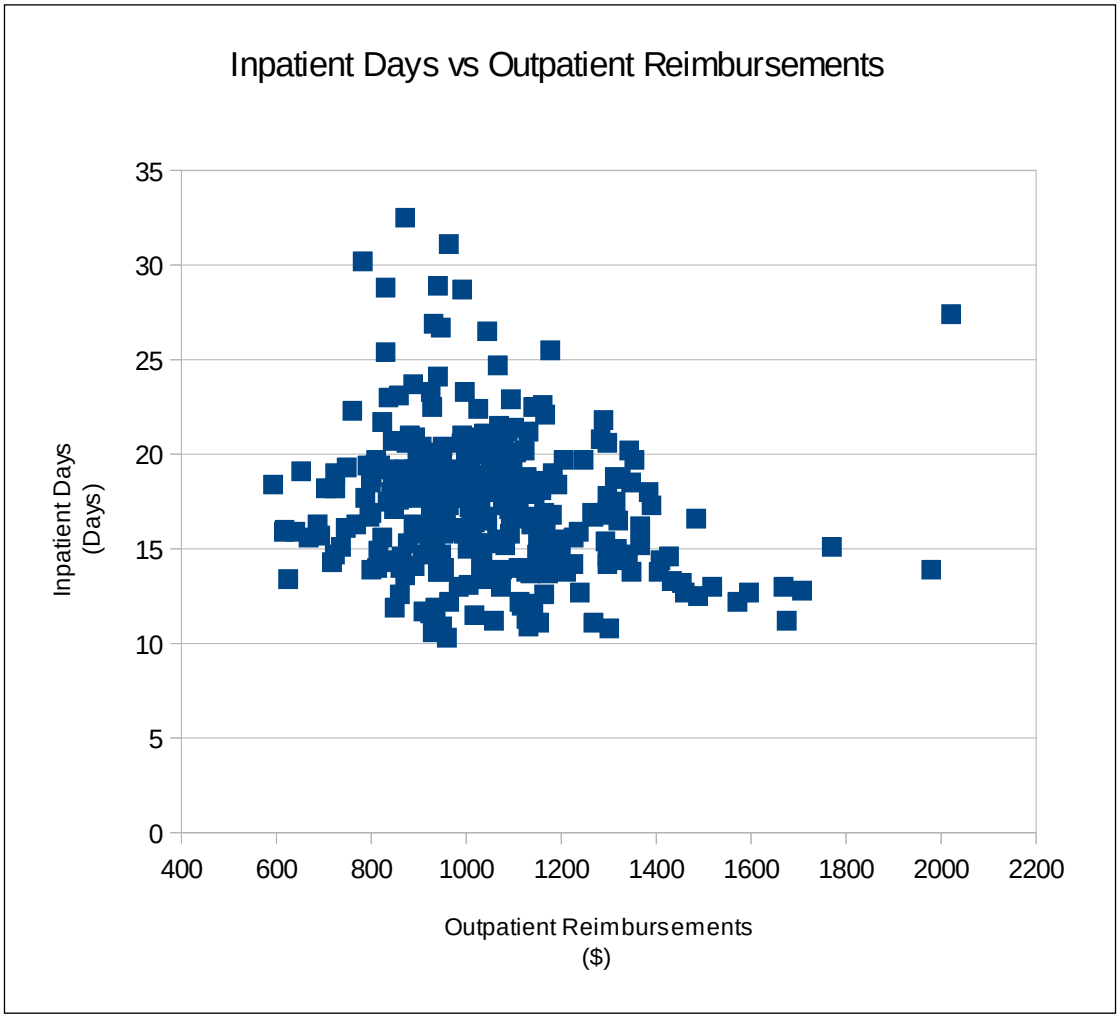

Figure 6. Scatterplot of Inpatient Days vs Total Physician Visits (n=307)

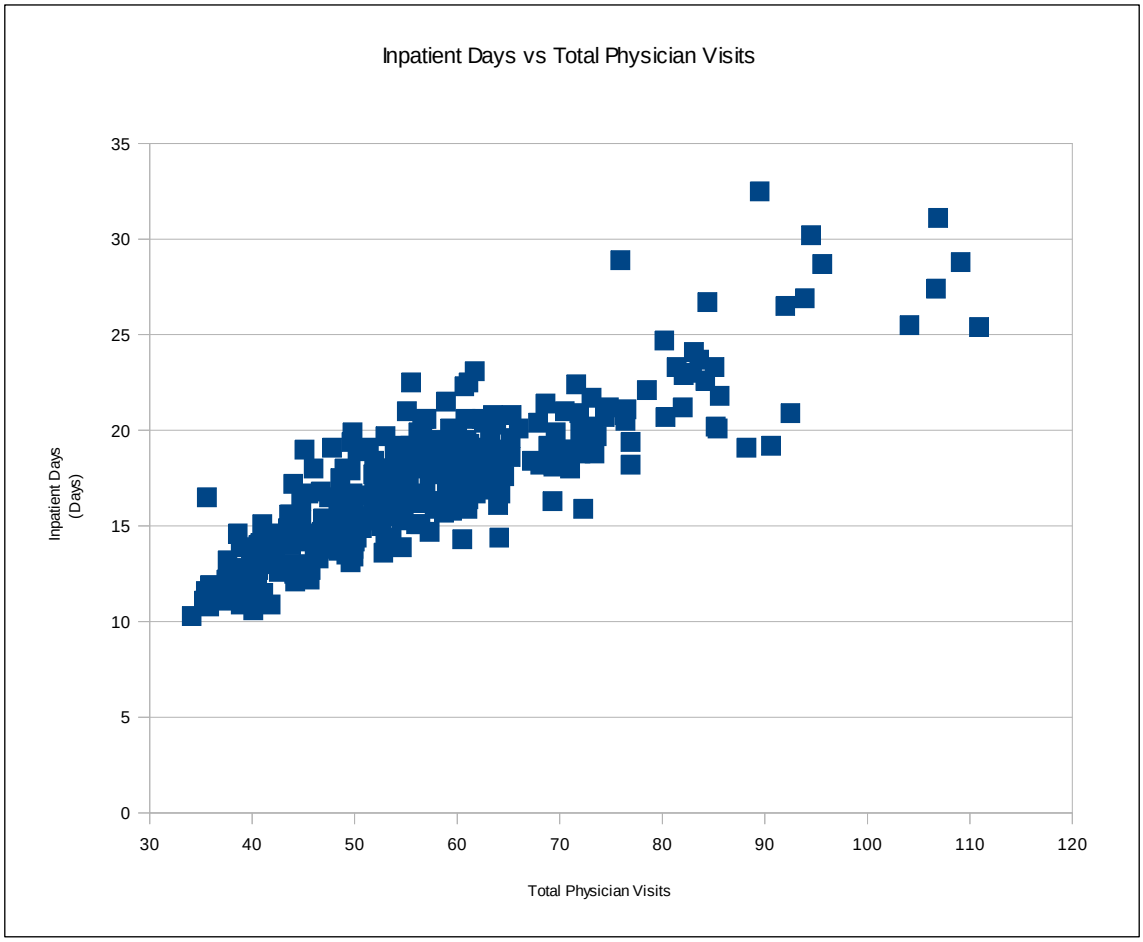

Table 1. Summary Statistics of Hospice Days

	Observations	Mean	Std. Deviation	Skewness	Kurtosis
Log Hospice Days	**308**	**2.69**	**0.35**	**-0.174**	**3.15**
Hospice Days	**308**	**15.58**	**5.39**	**0.866**	**3.93**

Table 2. Bivariate Correlation Matrix for All Variables (with p-values included in

paranthesis)

	Loghospice	Inpatient days	Average copay	Hospice reimburse	Outpatient reimburse	Inpatient short-stay reimburse	Inpatient long-stay reimburse	Total MD visits	Hospital-based nurses	Hospital-based MDs	CMS Hospital Rating	Total Reimburse
Loghospice	1.000											
Inpatient days	-0.250 (0.000)	1.000										
Average copay	-0.024 (0.677)	0.697 (0.000)	1.000									
Hospice reimburse	0.794 (0.000)	-0.098 (0.102)	0.057 (0.348)	1.000								
Outpatient reimburse	-0.236 (0.000)	-0.191 (0.001)	-0.377 (0.000)	-0.277 (0.000)	1.000							
Inpatient short-stay reimburse	-0.255 (0.000)	0.746 (0.000)	0.615 (0.000)	-0.155 (0.010)	-0.072 (0.207)	1.000						
Inpatient long-stay reimburse	-0.070 (0.233)	0.110 (0.061)	0.131 (0.026)	-0.034 (0.578)	0.118 (0.044)	0.050 (0.397)	1.000					
Total MD visits	-0.141 (0.014)	0.795 (0.000)	0.877 (0.000)	-0.029 (0.634)	-0.206 (0.000)	0.705 (0.000)	0.061 (0.304)	1.000				
Hospital-based nurses	-0.042 (0.461)	0.171 (0.003)	-0.262 (0.000)	-0.067 (0.265)	0.341 (0.000)	0.089 (0.119)	0.200 (0.001)	-0.165 (0.004)	1.000			
Hospital-based MDs	-0.069 (0.229)	0.111 (0.053)	0.274 (0.000)	-0.139 (0.021)	-0.091 (0.112)	0.157 (0.006)	-0.153 (0.009)	0.322 (0.000)	-0.180 (0.002)	1.000		
CMS Hospital Rating	-0.140 (0.014)	-0.084 (0.143)	-0.084 (0.143)	-0.222 (0.000)	0.163 (0.004)	-0.034 (0.552)	-0.197 (0.001)	0.037 (0.518)	-0.010 (0.861)	0.254 (0.000)	1.000	
Total Reimburse	-0.137 (0.017)	0.259 (0.000)	0.259 (0.000)	-0.043 (0.475)	-0.023 (0.684)	0.288 (0.000)	0.006 (0.915)	0.270 (0.000)	-0.077 (0.179)	0.050 (0.386)	-0.018 (0.752)	1.000

Table 3. Parameter Estimates for Multiple Regression Model (p-value and standard error, 'SE', included in paranthesis; n=272 and R-squared=0.72)

Variable	Beta Coefficient
Total Reimbursement	-0.000 (p=0.034; SE=0.0000)
CMS Hospital Rating	0.011 (p=0.053; SE=0.0057)
Inpatient days	-0.035 (p=0.000; SE=0.0052)
Hospital-based nurses	0.061 (p=0.006; SE=0.0222)
Average copay	0.000 (p=0.003; SE=0.0000)
Inpatient long-stay reimburse	0.000 (p=0.005; SE=0.0001)
Hospice reimburse	0.002 (p=0.000; SE=0.0001)
Outpatient reimburse	-0.000 (p=0.095; SE=0.0001)
Constant	1.459 (2.66; SE=0.5487)

Table 4. Variance Inflation Factor (VIF) Results for Multiple Regression Model Variables

Variable	VIF
Total Reimburse	1.09
CMS Hospital Rating	1.16
Inpatient days	2.77
Hospital-based nurses	1.64
Average copay	3.08
Inpatient long-stay reimburse	1.21
Hospice reimburse	1.2
Outpatient reimburse	1.41

Table 5. Cameron and Trivedi Decomposition Test for Multiple Regression Model Variables

Measure	Chi-square	p-value
Heteroskedasticity	124.26	0
Skewness	25.74	0
Kurtosis	2.63	0.11
Total	152.63	0.00

Table 6. Pagan/Cook-Weisberg Heteroskedasticity Test for Multiple Regression Model

Chi-square	0.28
P-value	0.6

Table 7. Simple Regression Results (for Dependent Variable=Total Medicare Reimbursement; for n=306, R-squared=0.014, *=significant at 0.05, and **=significant at 0.01)

Variable	Beta Coefficient
Hospice days	-117.79 (-2.08)*
Constant	28742.61 (30.70)**

Meta-analysis of Male and Female HPV Vaccination

Introduction

Nearly 10 million people in the United States between the ages of 15 and 24 years old are infected with HPV. HPV is believed to be an important risk factor for cervical cancer, a disease that is diagnosed in 12,000 women annually in the USA, with 4,000 women dying from it each year (CDC, 2007). The disease is largely preventable if a person is provided a three-dose vaccine ahead of time (CDC, 2012).

The two FDA-approved HPV vaccines are the bivalent vaccine Cervarix and the quadrivalent vaccine Gardasil. The quadrivalent vaccine provides protection from 4 strains of HPV, while the bivalent vaccine provides coverage against 2 strains (HPV 16 and 18) (CDC, 2007). The quadrivalent vaccine provides greater protection against HPV-related sequelae – protecting against nearly 100% of cervical pre-cancerous lesions, and also providing coverage against anal, vaginal, and vulvar cancers (CDC, 2012). Comparatively, the bivalent vaccine protects against 93% of cervical pre-cancers. The vaccines have a long duration of action, and initiate an antibody response in nearly 100% of adolescent females (CDC, 2012).

The quadrivalent and bivalent vaccines are indeed effective. However, there are serious concerns about vaccine compliance rates in the United States. Laz and co-authors found that only 3% of adolescents received all 3 vaccine doses by 11 to 12 years old. Full three dose compliance was only

19% by the ages of 13 to 17 years old. Parents are an important factor in adolescent compliant rates. Parental factors affecting vaccine non-compliance included: safety concerns (19%), questions on the need for the vaccine (26%), and questions about the vaccine itself (17%) (Laz et al, 2012). At the same time, Hirth and co-authors found that the likelihood of completing the three doses decreased as the age of vaccine recipients increased (Hirth et al, 2007).

Despite the compliance issues, female HPV vaccination has been widely seen as an effective intervention. Based on this, the ACIP recommended vaccination with the quadrivalent HPV vaccine for all male adolescents ages 11 to 12 years (CDC, October 2011; NY Times, 2011; Harris, 2011). In justifying its decision, the ACIP stated: *"Vaccination of males would provide direct benefits and likely would reduce HPV 6, 11, 16, and 18 transmission, and resulting infection, disease, and cancers in females (through herd immunity)."*(CDC, December 2011). Advocates for male HPV vaccination contend that it reduces the transmission of HPV to women, and also reduces the risk of men getting male-specific HPV disease such as penile cancer (Fontenot and Morelock, 2012).

As more and more adolescents receive the HPV vaccine across the United States, much attention has been centered on the cost-effectiveness of mass vaccination. In this paper, cost-effectiveness is defined as the comparison of interventions to determine which produces the same level of output at the lowest cost (World Bank IEG, 2007). While the cost-effectiveness of female HPV vaccination has been extensively studied, there are considerably fewer research studies evaluating the cost-effectiveness of male HPV vaccination. For example, a Pubmed search of '*female hpv cost-effectiveness*' returns 371 results compared to the 77 results for '*male hpv cost-effectiveness*' (Pubmed Search, 2012). This is not wholly surprising, since a disproportionate amount of the disease burden that

HPV alleviates is related to cervical cancer.

Systematic reviews are one method to evaluate cost-effectiveness. Reflecting the dearth of cost-effectiveness literature on the topic, there are only a few systematic reviews that look at the cost-effectiveness of male-female HPV vaccination compared to female-only HPV vaccination (Seto et al, 2012; Marra et al, 2009; Kim et al, 2008). These include studies by Seto et al and Marra et al that evaluate incremental cost-effectiveness ratios (ICERs) to make cost-effectiveness comparisons (Seto et al, 2012; Marra et al, 2009). For background, the ICER is a ratio that compares the cost of one practice to that of the current gold standard, and then divides this by the health effect of the practice to the current gold standard. In other words, it offers the incremental cost-effectiveness beyond the current standard of care. The health effect is defined by the Quality Adjusted Life Year (QALY). The formula is: ICER = (Cost of practice 1 – Cost of gold standard)/(Utility of practice 1 – Utility of practice 2). Cost-effective thresholds vary, but have been identified in the literature as between ($50,000-$100,000)/QALY, as well as ($90,000-$120,000)/QALY (Sarnnaliev; Kim and Goldie, 2009). Interventions that have ICER values below this cost-effectiveness threshold are designated as relatively cost-effective, while those above the threshold are considered relatively less cost-effective. A cost-effective the intervention is believed to have a good return on health outcomes for the amount of money spent (World Bank IEG, 2007).

The 2009 Seto et al study identified 4 studies. Outside of 1 paper that showed an ICER of US$440,000/QALY for adding males to a female-only program, the remaining studies had an ICER between US$17,000/QALY -$42,000/QALY. The 2012 Marra et al study had 4 more studies than Seto et al. For these additional papers, Marra and co-authors found that 1 study focused on cervical

outcomes, a second looked at cervical outcomes and genital warts, while the other 2 looked at cervical and non-cervical disease. In line with the 2009 paper, the 2012 study concluded that adding males was not cost-effective. It referred to Kim and Goldie's male-female ICER estimate of between US$114,000/QALY to $350,000/QALY as a baseline value. Beside this, the authors critiqued Elbasha and Dasbach's use of higher model input costs in obtaining their lower male-female ICER values (between US$25,664/QALY and $46,978/QALY).

While systematic reviews are valuable, they do not compare to a meta-analysis. Rigorous evaluation of each paper by meta-analytic techniques provides a more robust ICER value, weighted to the variability in the ICER estimate. This is valuable because it places relative weights on individual papers, and accounts for the variability within and between papers. Having an aggregate ICER value for male HPV vaccination will increase precision since investigators will now have a benchmark value in assessing male HPV vaccination, compared to multiple values that are confusing. On the other hand, if the analysis is not done in an aggregate way, mistakes might be made on the cost-effectiveness of HPV vaccination. For example, policy-makers might designate male HPV vaccination as cost-effective or not based on results that have large error bounds. In effect, policy decisions could be made without appropriately weighting studies for the variability in the reported data. This is especially relevant since California Governor Jerry Brown approved HPV vaccination for adolescents without parental approval. Advocates saw this measure as increasing access to critical preventive care, while critics contended it promoted sexual activity and was not cost-effective (Conley, 2011). With accurate male-female HPV cost-effectiveness data, the currently unsuccessful attempts to mandate HPV vaccination across the state might be completely changed (NCSL, 2012).

Paper Objectives

Currently there is a limited amount of male-female HPV vaccine cost-effectiveness research. The current exploratory analysis adds value by aggregating the individual paper results and providing a rough benchmark estimate of the cost-effectiveness between vaccination strategies. It is acknowledged that there are few papers available to complete a meta-analysis leading to particularly robust inference. Nevertheless, the value of aggregating the individual data was considered to be important despite this sample size limitation.

This analysis calculates the ICER values of a female-only vaccination strategy compared to a male and female vaccination strategy. The published male-female HPV papers are critically assessed. The end result is a single aggregate value that is more useful than multiple values from different papers. The main research questions are:

1) *When aggregating across papers, what is the fixed-effects incremental cost-effectiveness ratio (ICER) value for male and female HPV vaccination compared to female-only HPV vaccination?*
2) *Are the ICER results robust after completing sensitivity analysis?*

It is hypothesized that female-only HPV vaccination will have an ICER below the $50,000/QALY threshold, while male and female HPV vaccination will have an ICER above the $100,000/QALY cutoff. Past research shows female-only vaccination to relatively be cost-effective at this cutoff level, with less variability across papers in these results. Comparatively, results for male-female vaccination are substantially more variable across papers, and do not generally meet the $50,000/QALY cutoff level.

Comparison of Methods

Meta-analysis is a critical method for aggregating results across different studies. Factors that affect the quality of the meta-analysis include heterogeneity in the study design, disease data, and patient populations in the studies. To assess for robustness, one may conduct subgroup analysis where patient subgroups are specifically studied. For example, one might evaluate results for women only, or in smokers. Another technique is to complete sensitivity analysis, where certain studies are excluded from analysis (Davis and Crombie, 2001). As is discussed in two articles from *Medical Decision Making* and an open-source Cochrane Review article, there are multiple options for among meta-analysis techniques (Ades et al, 2005; Cochrane Collaborative, 2002).

Meta-analysis estimation methods include fixed effects, random effects, or Bayesian analysis. The fixed effects model assumes that all the studies have the same single "actual" treatment effect value, and that any heterogeneous error between study values is systematically random. As a result, there is no accounting of between-study error that might arise as a result of different sampling methods or varying target populations (Montori et al, 2008). In contrast, the random effects model assumes that each study has its own "actual" treatment effect value. The resulting variability in the estimates between studies is not interpreted as random, but rather follows a normal distribution (Ades et al, 2005; Hasselblad and McCrory, 1995; Cochrane Collaborative, 2002).

However, Montori, Ioannidis, et al note that random effects models are more likely to produce extreme results when the sample size is small (Montori et al, 2008). For studies that are similar, a fixed effects weighting can be done, based on the variance of the effect size between studies. On the other

hand, in random effects, the results are unweighted to account for between-study variability. Bayesian analysis assumes the ICER data has a prior and posterior probability distribution. The prior distribution is random, while the posterior distribution is a function of the prior distribution and the likelihood function for the model's parameters.

Fixed effects is advantageous when the data is homogenous and the studies are assumed to be the same. It also is better when a single true effect size is expected for all studies. The fixed effects model does not work as well with heterogeneity in the effect size of the data because the studies are weighted according to sample size. Random effects is better when the data is heterogeneous and there is a substantial difference between the studies (Borenstein et al, 2009). Random effects is also advantageous when the results are extended to different sorts of studies, where a common effect size is not suitable. Random effects is less appropriate for studies with small sample sizes. In addition, it is not good for homogenous data since the studies are weighted equally, and not by sample size (Borenstein et al, 2009). Bayesian analysis is advantageous in that it assumes a range of statistical distributions for the ICER values, increasing the chance for more efficient results. However, it is disadvantageous because it relies on *a priori* information that may be limited (Schmid, 2001).

In this paper, a fixed effects model was chosen. The trade-off for using fixed effects is that it assumes a single true ICER value for male-female HPV vaccination. Random effects assumes a distribution of true male-female ICER values, based on the conditions of the study. The cost of using fixed-effects is that the final ICER value may be skewed as a result of weighting studies by their sample size instead of assuming an additional component of variability. The benefit is that fixed-effects provides a better estimate for small samples. The major limitation for this paper is the small number of

published papers. Only 7 papers are used for the analysis. Based on this serious sample size limitation, fixed effects is ultimately used. Bayesian analysis is not used because of the limited data that is available for male-female HPV vaccination, and the disadvantage of setting *a priori* assumptions in such circumstances.

Methods

A Pubmed search was conducted using the search terms: *men hpv vaccination* and *"cost-effectiveness"*. In addition, the terms '*QALY*' and *'boys'* were used to expand the set of possible papers. Additional papers were identified as references in the 2012 Seto et al and 2009 Marra et all systemic reviews of male HPV vaccination. Inclusion criteria were: ICER values for male and/or boy HPV vaccination; ICER values for female-only HPV vaccination; and data from English-language publications. Both domestic and international papers were selected. In total, 8 papers were identified and 7 were used. The 8th paper was removed because it did not report cost-effectiveness in QALYs. It is acknowledged that 7 papers is a small sample size for calculation purposes. However, based on no previous precedent for male HPV meta-analysis, the study was still seen as contributing new information despite the small number of studies.

The selected papers are: Kim and Goldie, *BMJ*, 2009 (Kim and Goldie, 2009); Elbasha et al, *EID*, 2007 (Elbasha et al, 2007); Chesson et al, *Vaccine*, 2011 (Chesson et al, 2011); Jit et al, *BMJ*, 2008 (Jit et al, 2008); Kim et al, *BJC*, 2007 (Kim et al, 2007); Elbasha and Dasbach, *Vaccine*, 2010 (Elbasha and Dasbach, 2010); Taira et al, *EID*, 2004 (Taira et al, 2004); and Insinga et al, *Vaccine*, 2007 (Insinga et al, 2007). The Kim et al, *BJC*, 2007 paper is excluded from meta-analysis because it uses Years of Life Saved (YLS) as its main unit of cost-effectiveness analysis instead of quality adjusted life

years (QALYs). The two international papers were Jit et al, *BMJ* 2008 (United Kingdom) and Insinga et al, *Vaccine* 2007 (Mexico).

Each paper had an ICER value for female-only HPV vaccination and male-female HPV vaccination. The search was not restricted by year of publication year. Papers ranged in publication date from 2004 to 2011. To standardize ICER values, currency was converted to US dollars when necessary (ie. Jit et al, *BMJ*, 2008; Insinga et al, *Vaccine*, 2007), and then standardized to 2011 US dollars using the consumer price index (CPI) (US Department of Labor, 2012). Each paper had a set of ICER values for male-female and female-only vaccination. The mean of these values were calculated to obtain each paper's average male-female and female-only ICER value. A fixed-effects estimate was then derived across all papers. There was now a set of average male-female and female-only ICER values. The variance was calculated for the two sets of data. The inverse variance was obtained by taking the reciprocal of the variance for each strategy (1/variance). Next, the average ICER value for each paper was multiplied by its inverse variance (effect size (ES) * inverse variance (IV)). For each vaccination strategy, the sum of ES*IV across all papers was calculated. This value was divided by the sum of the IV across all papers to obtain the fixed-effects estimate for each strategy.

Meta-analysis is traditionally done for primary studies alone (Montori et al, 2008). This paper applies ICER values derived from simulations of primary data. We make this assumption because each paper attempts to evaluate the same phenomenon, despite the use of primary data from other sources and the application of simulation models with differing assumptions. To further assess this assumption, the comparability of ICER values across vaccination strategy might be evaluated using Bayesian, random-effects and fixed-effects approaches. Ultimately it might be a question of comparing the meta-

analysis to a well-designed HPV vaccination 'gold standard' cost-effectiveness study that has a large sample size. The meta-analysis results could then be compared to this reference study.

Data

Table 1 summarizes the papers used in the analysis. The papers did not uniformly measure the same disease states. Pre-cancerous lesions, CIN I to III, and genital warts were the most frequent disease state evaluated, with 5 of the 7 papers having data. Aggregate cervical cancer incidence and juvenile-onset recurrent respiratory papillomatosis (JORPP) was measured in 4 of the 7 papers. Non-cervical disease was measured in 3 of the 7 papers. Finally, HPV 16 and 18 incidence values were collected for at least 2 of the 7 papers. Few papers had data for the most prevalent disease states of cervical cancer, CIN I-III, and genital warts. Elbasha et al and Taira et al only had incidence rates for HPV 16 and HPV18. The Jit et al paper has HPV 16/18 and CIN I-III rates, but not rates for cervical cancer or genital warts. Further assumptions have to be made on the conversion rate from HPV infection to CIN or cervical cancer in these models – leading to presumably more flawed predictions. Only the Chesson et al and Elbasha and Dasbach papers have incidence data for CIN I-III, cervical cancer, and genital warts. However, the presence of this data does not lend insight into which data is most accurate. This is a clear issue since there is a difference between papers on the rates for cervical cancer (4.2 to 12.5), and genital wart rates (155 to 459).

As observed in Table 2, there is a lack of uniformity in what is measured in the different papers. This makes it difficult to assess which papers are the best. Disease states that were commonly assessed include: CIN, Cervical Cancer (Untreated and Treated), Genital Warts, Male Cancers (Penile Cancer),

and Rare Cancers (Vaginal Cancers, JORPP). There is not much difference between papers for their QALY values for genital warts and CIN. There is more variation for how cervical cancer is categorized. Certain papers describe treated versus untreated cervical cancer, by its stage. Other papers provide a QALY range for cervical cancer. The most limited papers are those by Insinga et al and Elbasha et al. They only provide data for CIN I-III and genital warts. There is no data for cervical cancer. There is also no variation between the disease states, with all states having a value of 0.97. Jit et al provides more comprehensive data for CIN I-III, cervical cancer, and genital warts; including standard deviations for the values. Chesson et al and Kim and Goldie provide data for CIN I-III, genital warts, cervical cancer, male-specific cancer, and rare cancers.

As Table 3 demonstrates, cost data varied between papers by the currency unit and year. The Insinga et al paper had currency in terms of Mexican pesos, while the Jit et al paper used British pounds. The vaccine cost was between $200 and $600 for 4 of the 7 papers. 2 of the 7 papers had a vaccine cost below $100. 3 of the 7 papers had cervical cancer costs between $30,000 and $55,000. 3 of the 8 papers had CIN I to III costs between $1000 and $4500. 1 of these 3 papers had the same value for CIN I, II, and III while the other 2 had increasing costs for the higher grade CIN. 1 of the 7 papers (Jit et al, 2008) had cost data for different stages of cervical cancer, with and without treatment.

Even after controlling for inflation, the cost data had a good amount of variability between the papers. The main issue was the cost projections for treatment of the different stages of cervical cancer. The papers were divided into three cost ranges: $30,000-$50,000 (Kim and Goldie; Elbasha and Dasbach, 2010; $15,000-$25,000 (Jit et al; Taira et al); and then $7,000-$7,800 (Insinga et al). The Chesson et al paper had a single value for cervical cancer treatment ($35,693), while the Elbasha et al,

2007 had starkly different values for female-only ($3,000-$9,000) versus male and female vaccination ($35,000-$65,000). Somewhat surprisingly, two papers (Elbasha et al, 2005; Taira et al) did not have cost data for genital warts, a common condition that the HPV vaccine can help prevent. Two of the papers, Insinga et al; Jit et al, used data from Mexico and the United Kingdom, respectively. Excepting these two papers, it is difficult to assess which papers are the best – since it is hard to say whether the $15,000-$25,000 range approximate the 'true' cost of cervical cancer treatment compared to the $30,000-$50,000 range. That being said, the Elbasha and Dasbach paper is the most comprehensive, with stage-specific cost data for not only cervical cancer, but also vulvar and vaginal cancer; besides having cost data for rare HPV-associated diseases.

As Table 4 shows, 1 of the 7 papers had ICER values ranging across increasing vaccine efficacy. Two of the 7 papers had values across increasing vaccine coverage rates. One of the 7 papers had values across increasing vaccine duration. For all 7 of the papers, female-only vaccination was more cost-effective than male add-on vaccination. At low coverage rates (<30%), male add-on vaccination was cost-effective ($23,600 in Chesson et al, 2011 and $110-$9370 in Kim et al, 2007). At higher coverage rates (>75%), male add-on vaccination was not cost-effective ($184,300 in Chesson et al, 2011 and $9,110-$136,910 in Kim et al, 2007). In comparison, female-only vaccination across the most stringent assumptions was still cost-effective ($33,868 for 10 year protection in Jit et al, 2008, and $27,370 for 50% efficacy in Kim and Goldie, 2006). Because there is such variety in how the ICER is assessed, it is difficult to determine what ICER values are most accurate. For example, Kim and Goldie calculate ICER across vaccine efficacy rates, while Chesson et al compare ICER data across vaccine coverage rates. Jit et al obtain ICER data for the U.K. and range it across different vaccine durations. Besides different scales for evaluating ICER, there is the added challenge of the accuracy of the

collected data: with the assumption that later studies draw from more comprehensive and accurate cervical cancer data than earlier studies.

As Table 5 shows, sample size differed significantly across the studies. The Kim and Goldie paper had a total sample of 282 million, while the Chesson et al paper had 281 million. This is substantially larger than the 80,000 to 4.1 million individuals in the other 5 studies. Based on this major difference, these two papers disproportionately skew the weighted ICER values. As well, the source for QALYs varied between papers. Some would draw from previous work (Insinga et al, 2007; Elbasha et al, 2007), while others drew from a standardized set of measures (Kim and Goldie, 2009; Elbasha and Dasbach, 2010). Chesson et al drew from 3 different sources, leading to a presumptively more robust analysis. Otherwise, there is much overlap between the models in the different papers – from having dynamic transmission models, to having similar age groups evaluated, to the timeframe of study. Based on the collected data, the Chesson et al paper seems to be the most comprehensive, since it includes more HPV strains than the Kim and Goldie paper.

Discussion

In evaluating the papers, 3 of the original 7 papers had cost-effectiveness ratios below <$50,000/QALY. These 3 papers had overlapping authors (Elbasha et al, 2007; Elbasha and Dasbach, 2010; Insinga et al, 2007). The ICER values for adding males versus having a female-only program were, respectively: $41,803/QALY: $4,666/QALY, $25,664/QALY:$3,282/QALY, and $7,075/QALY:$2,719/QALY. These papers did not include non-cervical disease in the analysis. The QALY values were disproportionately high compared to the other papers. The QALY range was 0.87-

0.97 between these 3 papers. In comparison, the Kim and Goldie (2009) paper had a QALY range of 0.48-0.91. Limited incidence data were available for these papers. Indeed, there is substantial heterogeneity in the disease states that are accounted for between papers.

At high coverage (>75%) and moderate efficacy levels (>50%), male HPV vaccination did not meet the conventional threshold for cost-effectiveness. The ratios between adding males and female-only vaccination for 3 of the papers (Kim and Goldie, 2009; Chesson et al, 2011; Kim et al, 2007; and Taira et al, 2004) were, respectively: $114510/QALY:$20990/QALY (at 90-100% vaccine efficacy), $184300/QALY:$10500/QALY (at 75% coverage), $136,910/YLS:$4180/YLS (at a vaccine coverage of 90% and a vaccine price of $400), and $442039/QALY:$14583/QALY. When coverage rates, and/or vaccine dose were low, the cost-effectiveness of male HPV vaccination generally increased. In the Chesson et al (2011) paper, reducing vaccine coverage to 20% changed the ICER from $184300/QALY to $23600/QALY. In Kim et al (2009), a reduction in vaccine efficacy from 90-100% to 50% changed the ICER from $114510/QALY to $164580/QALY. Overall, the fewer women that were vaccinated (either in terms of limited coverage, efficacy, or duration of protection), the higher the cost-effectiveness of vaccinating males. This held in cases where non-cervical disease was included in the analysis (Kim et al, 2009; Chesson et al, 2011), as well as in cases where only cervical disease was evaluated (Jit et al, 2008; Taira et al, 2004).

There is a substantial range in ICER values between papers. For male-female vaccination, the ICER range is $36,361/QALY to $380,284/QALY – a 10.5 magnitude difference. For female-only vaccination, ICERs range between $3,236/QALY to $46,187 – a 14.3 magnitude difference. The ICER is higher in male-female vaccination than female-only vaccination for all papers. At the same time, per

the magnitude difference, there is more variability in the ICER numbers for the female-only values than the male-female numbers. This may be because more data are available for female vaccination compared to male vaccination. Males having a lower risk of HPV-related cancer is the likely rationale for the higher male-female ICER value. However, there are differences in the model assumptions that makes the difference in ICER between male-female and female-only vaccination quite striking. Indeed, the proportionate difference between male-female versus female-only vaccination ranges between 4.45 to 12.9 – reflecting the heterogeneity between papers.

Based on the fixed-effects calculations, the aggregate ICER values for female-only vaccination is $8,498/QALY, while male-female vaccination is $42,425/QALY. There is an approximately 4-fold difference between male-female and female-only cost-effectiveness. Removing the international papers in the sensitivity analysis shows that female-only vaccination remains cost-effective. There is an approximate $2,700 increase in the aggregate ICER value when the two international papers are removed – with the aggregate ICER value changing to $11,269/QALY. Comparatively, there is a nearly $28,000 difference in the male-female ICER value with the international papers excluded. Based on these results, the initial hypotheses of female-only vaccination being cost-effective at the $50,000/QALY cost threshold is satisfied. Likewise, male-female vaccination is not cost-effective at the $50,000/QALY cost-effectiveness threshold.

The strengths to this analysis is that it aggregates multiple papers to obtain an aggregate ICER value for male-female and female-only HPV vaccination. This paper also uses multiple methods to provide robustness to these results, including weighted/unweighted techniques and arithmetic/geometric means. With the extensive data on incidence, costs, QALYs, and model

assumptions, there is additionally more context to these results. The heterogeneity in the measured disease states and the difference in the U.S. and international cost measures are two areas where the final results are likely influenced. limitations of this analysis include its small sample size. Only 7 papers were included in the assessment, making substantial variability in the ultimate findings a legitimate concern. There is limited information in many of the papers on how the models are explicitly structured. Without this information, it is harder to assess how model assumptions and input data directly affects the ICER values. Instead, qualitative assessments are made based on the difference between input values and the ICER data.

Conclusion

This paper evaluates the cost-effectiveness of male-female HPV vaccination using meta-analysis techniques. Future research that applies additional meta-analysis techniques on these data would be beneficial. Bayesian analysis of the data could be compared to the current analysis reported here. It was not done here because limited data are available for the cost-effectiveness of male HPV vaccination, placing limitations on *a priori* probability distributions. Result robustness can be further assessed by including more papers in the analysis, and increasing the amount of data. More sophisticated analysis can be conducted by, for example, studying the effect of different determinants on the ICER using meta-regression. Finally, it will be helpful to apply these findings to calculate the cost-benefit ratio of universal male-female HPV vaccination in the United States. The total costs and total benefits for mass vaccination is especially useful data when drawn from these aggregate statistics.

The Affordable Care Act (ACA) is a major factor when considering the total costs of mass HPV vaccination. For those with private insurance, the legislation mandates that all vaccines that are ACIP

recommended must be available to individuals without a copay or deductible (Healthcare.gov, 2010). Those without insurance and under 19 years old can receive the vaccine free-of-cost through the federal Vaccines for Children (VFC) program. The three dose regimen costs between $300 and $400. The cost subsidization by private insurers and the federal government will play an important role in increasing vaccine access, and simultaneously be a source of major expenses for both entities over the ensuing years (KaiserEdu.gov, May 2012).

Transmission of HPV is a major concern across the United States. The HPV vaccine is a powerful tool in preventing HPV-related sequelae. With mass vaccination for both adolescent males and females now part of official ACIP policy, it is more important than ever to determine the cost-effectiveness of vaccination. Meta-analysis is one compelling way to do so. Based on these results, and using the $100,000/QALY threshold as a cut-off, it is recommended that adolescent males be vaccinated with the HPV vaccine. The results of this research may be extended to not only determine cost-effectiveness, but also to calculate the overall costs of mass vaccination. In so doing, HPV infection prevention may be coordinated with greater precision – leading to even better outcomes for young men and women across the United States and larger world.

References

Ades AE, Lu G, Higgins JP. "The interpretation of random-effects meta-analysis in decision models". *Med Decis Making*. 2005 Nov-Dec;25(6):646-54.

Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. "Fixed Effect Versus Random Effects Models" Intro to Meta Analysis. January 2009. Website: http://www.meta-analysis.com/downloads/Meta-analysis%20fixed%20effect%20vs%20random%20effects.pdf

CDC. 'HPV Vaccine Information for Clinicians – Fact Sheet'. Last updated: July 12, 2012. Website: http://www.cdc.gov/std/hpv/STDFact-HPV-vaccine-hcp.htm

CDC. Morbidity and Mortality Weekly Report (MMWR). March 23, 2007. Volume 56. RR-2. Website: http://www.cdc.gov/mmwr/pdf/rr/rr5602.pdf

CDC. "Press Briefing Transcript: ACIP recommends all 11-12 year old males get vaccinated against HPV". Tuesday, October 25, 2011. Website: http://www.cdc.gov/media/releases/2011/t1025_hpv_12yroldvaccine.html

CDC. "Recommendations on the Use of Quadrivalent Human Papillomavirus Vaccine in Males - Advisory Committee on Immunization Practices (ACIP)". Morbidity and Mortality Weekly Report (MMWR). December 23, 2011. Website: http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6050a3.htm

Chesson HW, Ekwueme DU, Saraiya M, Dunne EF, Markowitz LE. "The cost-effectiveness of male HPV vaccination in the United States". *Vaccine*. 2011 Oct 26;29(46):8443-50.

Elbasha EH, Dasbach EJ. "Impact of vaccinating boys and men against HPV in the United States". *Vaccine*. 2010 Oct 4;28(42):6858-67. Epub 2010 Aug 14.

Elbasha EH, Dasbach EJ, Insinga RP. "Model for assessing human papillomavirus vaccination strategies". *Emerg Infect Dis*. 2007 Jan;13(1):28-41.

Harris G. "Panel Recommends Vaccination for Boys of 11". *New York Times*. October 25, 2011. Website: http://www.nytimes.com/2011/10/26/health/policy/26vaccine.html?_r=1

Hasselblad V, McCrory DC. "Meta-analytic tools for medical decision making: a practical guide". *Med Decis Making*. 1995 Jan-Mar;15(1):81-96.

Healthcare.gov. "The Affordable Care Act and Immunization". Healthcare.gov website. September 14, 2010. Website: http://www.healthcare.gov/news/factsheets/2010/09/affordable-care-act-immunization.html

Insinga RP, Dasbach EJ, Elbasha EH, Puig A, Reynales-Shigematsu LM. "Cost-effectiveness of quadrivalent human papillomavirus (HPV) vaccination in Mexico: a transmission dynamic model-based evaluation". *Vaccine*. 2007 Dec 21;26(1):128-39. Epub 2007 Nov 20.

Jit M, Choi YH, Edmunds WJ. "Economic evaluation of human papillomavirus vaccination in the United Kingdom". *BMJ*. 2008 Jul 17;337:a769.

KaiserEdu.org. "HPV Vaccines and Cancer in the U.S." KaiserEdu.org. May 2012. Website:

http://www.kaiseredu.org/Issue-Modules/HPV-Vaccines-and-Cancer-in-the-US/Background-Brief.aspx

Kim JJ, Andres-Beck B, Goldie SJ. "The value of including boys in an HPV vaccination programme: a cost-effectiveness analysis in a low-resource setting". *Br J Cancer*. 2007 Nov 5;97(9):1322-8. Epub 2007 Oct 9.

Kim JJ, Brisson M, Edmunds WJ, Goldie SJ. "Modeling cervical cancer prevention in developed countries". *Vaccine*. 2008 Aug 19;26 Suppl 10:K76-86. Website: http://www.ncbi.nlm.nih.gov/pubmed/18847560

Kim JJ, Goldie SJ. "Cost effectiveness analysis of including boys in a human papillomavirus vaccination programme in the United States". *BMJ*. 2009 Oct 8;339:b3884.

New York Times Editorial Board. "Editorial: For Their Own Good". *New York Times*. October 28, 2011. Website: http://www.nytimes.com/2011/10/29/opinion/the-hpv-vaccine-is-for-their-own-good.html

Marra F, Cloutier K, Oteng B, Marra C, Ogilvie G. "Effectiveness and cost effectiveness of human papillomavirus vaccine: a systematic review". *Pharmacoeconomics*. 2009;27(2):127-47. Website: http://www.ncbi.nlm.nih.gov/pubmed/19254046

Montori V, Ioannidis J, Cook DJ, Guyatt G. Fixed-effects and random-effects models. In: In: Guyatt G, et al, editors. Users' guides to the medical literature: a manual for evidence-based clinical practice. 2nd ed. New York: McGraw-Hill Medical; JAMA & Archives Journals; 2008. p. 555-62.

Pubmed Search. Accessed: October 31, 2012.

Sarnnaliev M. "Cost-effectiveness Analysis (CEA)". Children's Hospital Boston.

Schmid CH. "Using Bayesian Inference to Perform Meta-Analysis". *Eval Health Prof*. June 2001. vol. 24. no. 2. 165-189. Website: http://ehp.sagepub.com/content/24/2/165.abstract

Seto K, Marra F, Raymakers A, Marra CA. "The Cost Effectiveness of Human Papillomavirus Vaccines: A Systematic Review". *Drugs*. 2012 Mar 13. Website: http://www.ncbi.nlm.nih.gov/pubmed/22413761

Smolders B, Lemmens R, Thijs V. "Lipoprotein (a) and stroke: a meta-analysis of observational studies". *Stroke*. 2007 Jun;38(6):1959-66. Epub 2007 May 3. Website: http://stroke.ahajournals.org/content/38/6/1959.full

Taira AV, Neukermans CP, Sanders GD. "Evaluating human papillomavirus vaccination programs". *Emerg Infect Dis*. 2004 Nov;10(11):1915-23.

The Cochrane Collaborative. "Combining Studies". Cochrane Open Learning Materials. Version 1.1. November, 2002. Website: http://www.cochrane-net.org/openlearning/html/mod12-3.htm

U.S. Labor Department. "Consumer Price Index". Website: ftp://ftp.bls.gov/pub/special.requests/cpi/cpiai.txt

Appendix

Table 1. Incidence Data

It is noted that each row represents the incidence rate percent for the specific disease state. Values either come as single estimates or in ranges. They may be stratified between men (M), women (W), or women and men (W+M). Each column references the specific paper.

Incidence	**Kim and Goldie, *BMJ*, 2009 (per 100,000)**	**Elbasha et al, *EID*, 2007**	**Chesson et al, *Vaccine*, 2011 (per 100,000)**	**Elbasha and Dasbach, *Vaccine*, 2010 (Model predictions only)**	**Taira et al, *EID*, 2004**	**Insinga et al, *Vaccine*, 2007 (per 100,000)**	**Jit et al, *BMJ*, 2008**
HPV 16 (% Incidence)	N/A	2.4% (W) 1.7% (M)	N/A	N/A	<18 yrs: 2.6% (W), 3.5% (M)	N/A	3.30%
HPV 18 (% Incidence)	N/A	2.4% (W) 1.7% (M)	N/A	N/A	<18 yrs: 0.9% (W), 1.2% (M)	N/A	1.10%
HPV 16/18 (% Incidence)	N/A	N/A	N/A	N/A	N/A	0.0075% (Upper Bound: F) 0.006% (Upper Bound: M)	N/A
CIN I	N/A	N/A	0.459%	0.052%	N/A	N/A	3.2%

(% Incidence)							
CIN II (% Incidence)	N/A	N/A	0.288 %	0.122% (CIN II and III)	N/A	N/A	0.86%
CIN III (% Incidence)	N/A	N/A	0.117 %	0.122% (CIN II and III)	N/A	N/A	1.1%
Incide-nce	**Kim and Goldie, *BMJ*, 2009 (per 100,000)**	**Elbasha et al, *EID*, 2007**	**Chesson et al, *Vaccine*, 2011 (per 100,000)**	**Elbasha and Dasbach, *Vaccine*, 2010 (Model predictions only)**	**Taira et al, *EID*, 2004**	**Insinga et al, *Vaccine*, 2007 (per 100,000)**	**Jit et al, *BMJ*, 2008**
CIN II/III (% Incidence)	N/A	N/A	N/A	N/A	N/A	0.100%	N/A
Cervical Cancer (% Incidence)	0.0042% - 0.0628%	N/A	0.0125%	0.0065%	N/A	0.0048%	N/A
Vulvar Cancer (% Incidence)	0.0002% - 0.0196%	N/A	0.0048%	0.00046%	N/A	N/A	N/A
Vaginal Cancer (% Incidence)	0.0001% -0.006%	N/A	0.0015%	0.00024%	N/A	N/A	N/A
Penile Cancer (% Incidence)	0.0000% - 0.0076%	N/A	0.0028%	0.00079%	N/A	N/A	N/A

Anal Cancer (% Incidence)	0.0-0.0056% (W) 0.0001% - 0.0043% (M)	N/A	0.0053%	0.00249% (W+M)	N/A	N/A	N/A
Incide-nce	**Kim and Goldie, *BMJ*, 2009 (per 100,000)**	**Elbash-a et al, *EID*, 2007**	**Chesson et al, *Vaccine*, 2011 (per 100,000)**	**Elbasha and Dasbach, *Vaccine*, 2010 (Model predictions only)**	**Taira et al, *EID*, 2004**	**Insinga et al, *Vaccine*, 2007 (per 100,000)**	**Jit et al, *BMJ*, 2008**
Oral Cancer (% Incidence)	0.0002% - 0.0139% (W) 0.0001-0.0177% (M)	N/A	N/A	0.00573% (head/neck)	N/A	N/A	N/A
Oropharyngeal Cancer (% Incidence)	0.0-0.0019% (W) 0.0-2.9 (M)	N/A	0.0052%	0.00573% (head/neck)	N/A	N/A	N/A
Genital warts (% Incidence)	0.0007% - 0.0062% (W) 0.00013%-0.00501	N/A	0.459%	0.316% (W+M) 0.162% (W) 0.155% (M)	N/A	0.155 %	0.49%

	%(M)						
JORRP (% Incidence)	0.0043%	N/A	0.00074%	0.00072%	N/A	N/A	N/A

Table 2. QALY Data

It is noted that all QALY values are point-in-time wellness for each disease state, excepting those values where treatment (Tx) or no treatment (no Tx) are expressly specified.

QALY	**Kim and Goldie, *BMJ*, 2009**	**Elbasha et al, *EID*, 2007**	**Chesson et al, *Vaccine*, 2011 (lifetime QALYs lost)**	**Jit et al, *BMJ*, 2008**	**Elbasha and Dasbach, *Vaccine*, 2010**	**Taira et al, *EID*, 2004**	**Insinga et al, *Vaccine*, 2007**
HPV 16/18	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Low Grade SIL	N/A	N/A	N/A	N/A	N/A	0.97	N/A
High Grade SIL	N/A	N/A	N/A	N/A	N/A	0.97	N/A
Cervical Cancer, F/U -I	N/A	N/A	N/A	N/A	N/A	0.9	N/A
Cervical Cancer, F/U	N/A	N/A	N/A	N/A	N/A	0.62	N/A

-II							
Cervical Cancer, F/U -III	N/A	N/A	N/A	N/A	N/A	0.62	N/A
Cervical Cancer, F/U -IV	N/A	N/A	N/A	N/A	N/A	0.62	N/A
CIN I	N/A	0.97	0.959	0.988 (SD=0.031)	0.91	N/A	0.97
CIN II	N/A	0.97	0.943	0.935 (SD=0.0051)	0.87	N/A	0.97
CIN III/CIS	N/A	0.97	0.940	0.946 (SD=0.051)	0.87	N/A	0.97
Cervical Cancer	0.48-0.76	N/A	0.58	N/A	N/A	N/A	N/A
Survivors: Local Cancer	N/A	N/A	0.73	N/A	N/A	N/A	N/A
QALY	**Kim and Goldie, *BMJ*, 2009**	**Elbasha et al, *EID*, 2007**	**Chesson et al, *Vaccine*, 2011 (lifetime QALYs lost)**	**Jit et al, *BMJ*, 2008**	**Elbasha and Dasbach, *Vaccine*, 2010**	**Taira et al, *EID*, 2004**	**Insinga et al, *Vaccine*, 2007**
Survivors: Regional Cancer	N/A	N/A	0.67	N/A	N/A	N/A	N/A
Survivors: Distant Cancer	N/A	N/A	0.55	N/A	N/A	N/A	N/A
Non-Survivors: Local Cancer	N/A	N/A	0.64	N/A	N/A	N/A	N/A

Non-Survivors : Regional Cancer	N/A	N/A	0.49	N/A	N/A	N/A	N/A
Non-Survivors : Distant Cancer	N/A	N/A	0.55	N/A	N/A	N/A	N/A
No Tx: Cx – I	N/A	N/A	N/A	0.65 (SD=0.082)	0.76	N/A	N/A
No Tx: Cx – II	N/A	N/A	N/A	0.56 (SD=0.071)	0.76	N/A	N/A
No Tx: Cx – III	N/A	N/A	N/A	0.56 (SD=0.071)	0.67	N/A	N/A
No Tx: Cx – IV	N/A	N/A	N/A	0.48 (SD=0.061)	0.48	N/A	N/A
Tx: Cx – I	N/A	N/A	N/A	0.90 (SD=0.066)	N/A	0.79	N/A
Tx: Cx – II	N/A	N/A	N/A	0.85 (SD=0.077)	N/A	0.62	N/A
Tx: Cx – III	N/A	N/A	N/A	0.85 (SD=0.077)	N/A	0.62	N/A
QALY	**Kim and Goldie, *BMJ*, 2009**	**Elbasha et al, *EID*, 2007**	**Chesson et al, *Vaccine*, 2011 (lifetime QALYs lost)**	**Jit et al, *BMJ*, 2008**	**Elbasha and Dasbach, *Vaccine*, 2010**	**Taira et al, *EID*, 2004**	**Insinga et al, *Vaccine*, 2007**
Tx: Cx – IV	N/A	N/A	N/A	0.62 (SD=0.061)	N/A	0.62	N/A
Vulvar Cancer	0.68	N/A	0.65	N/A	N/A	N/A	N/A
Vaginal Cancer	0.68	N/A	0.54	N/A	N/A	N/A	N/A

Penile Cancer	0.68	N/A	0.66	N/A	N/A	N/A	N/A
Anal Cancer	0.68	N/A	0.65 (F) 0.64 (M)	N/A	N/A	N/A	N/A
Oral Cancer	0.68	N/A	N/A	N/A	N/A	N/A	N/A
Oropharyngeal Cancer	0.68	N/A	0.47 (F) 0.49 (M)	N/A	N/A	N/A	N/A
Genital warts	0.91	0.97	0.9807 (F) 0.9789 (M)	0.961 (SD=0.017)	0.91	N/A	0.97
JORRP	0.69	N/A	0.77	N/A	0.8	N/A	N/A

Table 3. Cost Data

Costs	Kim and Goldie	Elbasha et al	Chesson Et al	Jit Et al	Elbasha, Dasbach	Taira Et al	Insinga Et al
HPV 16/18	N/A	N/A	N/A	N/A	N/A	40929	N/A
Cervical screen	N/A	N/A	N/A	108.9	112	81	16.64
Colposcopy	N/A	N/A	N/A	427.68	187	N/A	44.55
Hysterectomy	N/A	N/A	N/A	N/A	N/A	7883	N/A
precancerous les	N/A	N/A	N/A	657.36	N/A	N/A	N/A
S1 Cx Tx	N/A	N/A	N/A	13113.54	30059	14979	6985.64
S2 Cx Tx	N/A	N/A	N/A	21601.8	30059	21811	6985.64
S3 Cx Tx	N/A	N/A	N/A	20946.42	32171	21811	7895.09
S4 Cx Tx	N/A	N/A	N/A	21849.3	51527	24004	7729.27
Warts Tx	N/A	N/A	N/A	265.32	515 (F); 607 (M)	N/A	181.82
Vaccine Cost	N/A	$360 ($300-500) ludes administra	$500 ($360-600)	118.8-159.4	133	300 (3 doses) 100 (booster)	240.00
hool Admin. Co	N/A	N/A	N/A	7.0488	N/A	N/A	N/A
vider Admin. Co	N/A	N/A	N/A	19.8	N/A	N/A	N/A
LSIL	N/A	N/A	N/A	N/A	N/A	630	N/A
HSIL	N/A	N/A	N/A	N/A	N/A	1218	N/A
CIN I	N/A	N/A	1959	N/A	1764	N/A	1441.00
CIN II	N/A	N/A	3642	N/A	3955	N/A	1441.00
CIN III	N/A	N/A	4135	N/A	3955	N/A	1441.00
Cervical Cancer	29540-45540	2422, 9900 (F+CU); 36,161, 65,810 (F/M+CU)	35693	N/A	30059 (Local) 32171 (Regional 51527 (Distant)	N/A	N/A
Vulvar Cancer	20430	N/A	19697	N/A	12380 (Local) 26740 (Regional 28217 (Distant)	N/A	N/A
Vaginal Cancer	23440	N/A	26756	N/A	27848 (Local) 24512 (Regional 24512 (Distant)	N/A	N/A
Penile Cancer	17110	N/A	18528	N/A	18528	N/A	N/A
Anal Cancer	31300	N/A	33894	N/A	32902	N/A	N/A
Oral Cancer	37370	N/A	N/A	N/A	40463 ed 'Head/Neck C	N/A	N/A
pharyngeal Can	37370	N/A	40463	N/A	40463 ed 'Head/Neck C	N/A	N/A
Genital warts	430	N/A	568	N/A	515 (F); 607 (M)	N/A	181.82
JORRP	62010	N/A	137308	N/A	214952	N/A	N/A

Table 4. ICER Data (Without Full Sensitivity Data Included)

ICER	Kim and Goldie, *BMJ*, 2009	Elbasha et al, *EID*, 2007	Chesson et al, *Vaccine*, 2011	Jit et al, *BMJ*, 2008	Elbasha and Dasbach, *Vaccine*, 2010	Taira et al, *EID*, 2004	Insinga et al, *Vaccine*, 2007
Male and Females	$114,510 (90-100% efficacy) $164580 (50% efficacy)	$41,803	$184,300 (75% coverage) $41,400 (30% coverage) $23,600 (20% coverage)	10 yr protection: $113,846 20 yr protection: $172,892 Lifetime protection: $520,255	$25,664	$442,039	$7,075
Female Only	$20,990 (90-100% efficacy) $27,370 (50% efficacy)	$4,666	$10,500 (75% coverage) $7,200 (30% coverage) $5,700 (20% coverage)	10 yr protection: $33,868 20 yr protection: $22,474 Lifetime protection: $15,094	$3,282	$14,583	$2,719

Table 5. Model Assumptions Data

Model assumptio -ns	**Kim and Goldie, *BMJ*, 2009 (2006 US $)**	**Elbasha et al, *EID*, 2007 (2005 US $)**	**Chesson et al, *Vaccine*, 2011 (2008 US $)**	**Jit et al, *BMJ*, 2008 (Pounds; 1 Pound=$1.98)**	**Elbasha and Dasbach, *Vaccine*, 2010 (2008 US $)**	**Taira et al, *EID*, 2004**	**Insinga et al, *Vaccine*, 2007 $US/QALY**
Dynamic or Static?	Dynamic Transmissio -n	Dynamic	Dynamic Transmissio -n	Dynamic Transmission	Dynamic Transmission	Dynamic Transmissio -n	Dynamic Transmission
HPV Type	16, 18	6,11,16,18	6,11,16,18	Vaccine #1: 6, 11, 16, 18 Vaccine #2: 16, 18	4, 6,11,16,18	16, 18	6,11,16,18
Effectiveness Unit	QALY	QALY	QALY	QALY	QALY	QALY Per Life-Year	QALY
QALY Source	Gold et al, 1998.	Dasbach et al, 2006.	Institute of Medicine (IOM), 2000. Kulasingam et al, 2007. Myers et al, 2004.	Parametric fitting using Monte Carlo sampling	Gold et al, 1998.	Unclear	Elbasha et al, 2007.

Model assumpti-ons	**Kim and Goldie, *BMJ*, 2009 (2006 US $)**	**Elbasha et al, *EID*, 2007 (2005 US $)**	**Chesson et al, *Vaccine*, 2011 (2008 US $)**	**Jit et al, *BMJ*, 2008 (Pounds; 1 Pound=$1.98)**	**Elbasha and Dasbach, *Vaccine*, 2010 (2008 US $)**	**Taira et al, *EID*, 2004**	**Insinga et al, *Vaccine*, 2007 $US/QAL-Y**
Discounting Rate	3%	3.00%	N/A	3.5%	3%	N/A	3.00%
Perspective	Societal	Healthcare (USA)	Societal	Healthcare (UK)	N/A	N/A	Healthcare system (Mexican)
Population Size	138,595,702 (M) 143,742,929 (W) 282,338,631 (Total)	100,000 people (equal M:W ratio)	135,858,651 (M) 145,463,047 (W) 281 321 698 (Total)	Base case: 80,000 infants/yr	100,000 people (equal M:W ratio)	Base case: 2 million girls, 2.1 million boys	100,000 people
Model assumptions	Kim and Goldie, *BMJ*, 2009 (2006 US $)	Elbasha et al, *EID*, 2007 (2005 US $)	Chesson et al, *Vaccine*, 2011 (2008 US $)	Jit et al, *BMJ*, 2008 (Pounds; 1 Pound=$1.98)	Elbasha and Dasbach, *Vaccine*, 2010 (2008 US $)	Taira et al, *EID*, 2004	Insinga et al, *Vaccine*, 2007 $US/QALY
Population Characterist-ics	W: 12 yrs only M: 12 yrs only C/E: 20 yr-	Base case: 12 yr old girls, 12 yr old boys	W: 12-26 yrs M: 12 yrs only	Base case: 12 yr old girls Alternative: 12 yr old boys and girls Alternative: 13, 14	W: 9-26 yrs M: 9-26 yrs	Base case: 12 yr old girls, 12 yr old boys	12 years or >

	old W for screening	<u>Alternative:</u> 12-24 yr catch-up (W and M)		yr old girls		<u>Booster:</u> At 22 years old for girls	
Model assumpti-ons	**Kim and Goldie, *BMJ*, 2009 (2006 US $)**	**Elbasha et al, *EID*, 2007 (2005 US $)**	**Chesson et al, *Vaccine*, 2011 (2008 US $)**	**Jit et al, *BMJ*, 2008 (Pounds; 1 Pound=$1.98)**	**Elbasha and Dasbach, *Vaccine*, 2010 (2008 US $)**	**Taira et al, *EID*, 2004**	**Insinga et al, *Vaccine*, 2007 $US/QAL-Y**
Timeframe	100 yr timeframe from post-10 yr vaccination	100 years	100 years	100 years	100 years	Lifetime, not further specified	100 years
Vaccine coverage rate	50% and 75%	<u>Base case:</u> 90% <u>Alternative:</u> 50%	20%, 30%, 75%	Base-case: 80%	W: 50% (9-12 yrs), 85% (18 yrs), 90% (26 yrs)	W: 70%	70% (12-24 yrs, over 5 yrs)
Vaccine effectiveness	HPV-targeted Dx: 100% (W), 85% (M) HPV-Associated Dx: 100% (W) 90% (M)	100% (HPV-specific) and 90% (“associated disease”)	95% (W) 90% (M)	100.00% (W and M)	95% (after 3 doses) 85% (after 2 doses)	90% against HPV 16/18	90.00% 95.2%-incident cervical Dx 98.9%-genital warts
Model assumptions	Kim and Goldie,	Elbasha et al,	Chesson et al,	Jit et al, *BMJ*, 2008 (Pounds; 1	Elbasha and Dasbach, *Vaccine*, 2010	Taira et al, *EID*,	Insinga et al, *Vaccine*,

	BMJ, 2009 (2006 US $)	*EID*, 2007 (2005 US $)	*Vaccine*, 2011 (2008 US $)	Pound=$1.98)	(2008 US $)	2004	2007 $US/QALY
Vaccine duration	20 years and lifelong	Base case: lifelong Alternative: 10 years	Lifelong	10 yrs, 20 yrs, and lifelong	32 years	10 years, followed by booster	10 years
Model assumpti-ons	**Kim and Goldie, *BMJ*, 2009 (2006 US $)**	**Elbasha et al, *EID*, 2007 (2005 US $)**	**Chesson et al, *Vaccine*, 2011 (2008 US $)**	**Jit et al, *BMJ*, 2008 (Pounds; 1 Pound=$1.98)**	**Elbasha and Dasbach, *Vaccine*, 2010 (2008 US $)**	**Taira et al, *EID*, 2004**	**Insinga et al, *Vaccine*, 2007 $US/QAL-Y**
Herd immunity or not?	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Cross-immunity?	Yes	No	No	Yes	No	No	No

Table 6. Meta-Analysis Results

Vaccination Strategy	Cumulative ICER ($/QALY)	Sensitivity Analysis ICER ($/QALY)
Female-only	8,498	11,269
Male-Female	42,425	70,248

I. Individual Paper Assessments:

Kim and Goldie, *BMJ*, 2009	Elbasha Et al, *EID*, 2007	Chesson Et al, *Vaccine*, 2011	Jit Et al, *BMJ*, 2008	Elbasha and Dasbach, *Vaccine*, 2010	Taira Et al, *EID*, 2004	Insinga Et al, *Vaccine*, 2007
F Only	F Only	F only	F only	F only	F only	F12 only
40310	2964	18300	67059	22113	88572	2719
40310	997	14000	52367	20248	108685	2498
28940	7553	8800	41200	10041	284904	2832
28940	5241	5700	33638	6802	27600	3454
20990	2094	21300	269931	5270	28181	3048
20990	4273	16500	44499	3418	27556	8153
14540	3116	10500	37335	3282		2651
14540	2636	7200	32506		M+F	3296
	3449	28400	23475	M+F	51646	3690
M+F	4666	22000	254038	195322	116413	12651
290290	2422	14500	29886	178908	285776	
382860	9900	10500	27001	69038	57795	M+F
190030	7739		26144	62293	388368	1663
255210	21121	M+F	22788	46978	40865	6319
114510	4187	69600	209561	27511		40835
154420	5403	52100		25664		16702
90870	4922	29700	M+F			22340
123940	4221	23600	225415			8140
	5269	121700	342326			33229
		89100	1030105			21693
	M+F	50800				41486
	41803	41400				
	33469	741300				
	61250	436000				
	82700	229600				
	54755	184300				
	39990					
	40269					
	23862					
	45506					
	36161					
	65810					
	83714					
	54928					
	51436					
	43930					
	43974					
	36235					
	100418					

II. Fixed-Effects Calculation for Full Data:

A. Female-only Data

Paper	Effect Size	Variance	Weight (1/Variance)	Effect Size * Weight
Kim and Goldie	28,914	128,694,908	7.77 E -9	0.000225
Elbasha et al	6,086	24,496,382	4.08 E -8	0.000248
Chesson et al	14,872	46,620,998	2.14 E -8	0.000319
Elbasha and Dasbach	10,603	67,530,498	1.48 E -8	0.000157
Taira et al	114,297	14,660,817,585	5.82 E -11	7.79 E -6
Insinga et al	5,092	14,015,580	7.13 E -8	0.000363
Jit et al	92,308	10,748,495,370	9.30 E -11	0.000009
Cumulative Values	--	--	1.56 E -7	0.001329
Aggregate ICER Value				8498

II. Fixed-Effects Calculation for Full Data:

B. Male-Female Data

Paper	Effect Size	Variance	Weight (1/Variance)	Effect Size * Weight
Kim and Goldie	22,054	12,500,103,780	7.99 E -11	1.76 E -5
Elbasha et al	59,118	510,613,883	2.0 E -9	1.16 E -4
Chesson et al	173,171	46,353,086,507	2.16 E -11	3.74 E -6
Elbasha and Dasbach	90,234	5,439,890,358	1.84 E -10	1.66 E -5
Taira et al	190,164	31,233,460,385	3.20 E -11	6.09 E -6
Insinga et al	24,196	278,121,546	3.60 E -9	8.70 E -5
Jit et al	629,551	2.64 E 11	3.79 E -12	2.39 E -6
Cumulative Values	--	--	5.88 E -9	0.00025
Aggregate ICER Value				42,425

III. Sensitivity Analysis – Fixed-Effects Calculation with International Papers Removed (Jit et al and Insinga et al papers excluded)

A. Female-only Data

Paper	Effect Size	Variance	Weight (1/Variance)	Effect Size * Weight
Kim and Goldie	28,914	128,694,908	7.77 E -9	0.000225
Elbasha et al	6,086	24,496,382	4.08 E -8	0.000248
Chesson et al	14,872	46,620,998	2.14 E -8	0.000319
Elbasha and Dasbach	10,603	67,530,498	1.48 E -8	0.000157
Taira et al	114,297	14,660,817,585	5.82 E -11	7.79 E -6
Cumulative Values	--	--	6.82 E -11	0.000957
Aggregate ICER Value				11,269

III. Sensitivity Analysis – Fixed-Effects Calculation with International Papers Removed (Jit et al and Insinga et al papers excluded)

B. Male-Female Data

Paper	Effect Size	Variance	Weight (1/Variance)	Effect Size * Weight
Kim and Goldie	22,054	12,500,103,780	7.99 E -11	1.76 E -5
Elbasha et al	59,118	510,613,883	2.0 E -9	1.16 E -4
Chesson et al	173,171	46,353,086,507	2.16 E -11	3.74 E -6
Elbasha and Dasbach	90,234	5,439,890,358	1.84 E -10	1.66 E -5
Taira et al	190,164	31,233,460,385	3.20 E -11	6.09 E -6
Cumulative Values	--	--	2.28 E -9	0.000160
Aggregate ICER Value				70,248

The Application of Quantitative Methods to Merge Data between Studies

Introduction

This study poses the following research question: *Can statistical methods be used on two different studies to proxy the results of a third study that is a randomized clinical trial (RCT)?* The paper considers three different approaches: multiple linear regression, propensity score matching, and a simple subtraction technique. A cohort study and an evaluator-blinded randomized study evaluating ocular hypertension with one of two medications, Xalatan or Xalacom, are compared to a third RCT that tests Xalatan against Xalacom.

The unique feature of this paper is that it does not apply statistical methods to observational data drawn from the same population as the randomized clinical trial. All three studies have different patient populations and study designs, but will be evaluated to see if the two studies can replicate the results of the double-blinded RCT. This might be possible to some extent since the studies share the same outcome measure and primary independent variables of age and gender. However, they differ in the treatment timeframe and geographical site of sampling. It is assumed that the data in these studies are not correlated and were drawn analogous to an independent random draw. However, it will be an important limitation to consider the effect of the differences in study design on the final results. As well, unobserved variables and the effect of omitted variable bias is another limitation when evaluating the final results.

The pooling of data to obtain between-study treatment results has a long history (e.g., meta-analysis). William Cochran was a pioneer in this area from the 1930s through the 1950s. In a signature paper, he sought to merge data from studies measuring the same outcome variable, but with different treatment conditions in each study (Cochran, 1954).

In Cochran's analysis, there was an overarching question of whether the between-study difference in the value of replicates violated the assumption of homogeneity of variance. The classic Bartlett test was used to determine whether there was a violation of the homogeneity assumption. If there was not, then the mean values from each study could be arithmetically averaged to obtain the cumulative between-study value. However, if homogeneity was violated, then a weighting had to be done for the mean and standard error of each study, with further inclusion of a correction factor, *f*. Cochran concluded that weighting generally was unnecessary in this study (Cochran, 1954).

Heterogeneity is routinely evaluated in meta-analysis (Higgins and Thompson, 2002; Thompson and Sharp, 1999; Olkin, 1995).

There have been several evaluations of different studies proxying the results of a randomized study including LaLonde (1986), and Dehejia and Wahba (1999 and 2002). Using data from the National Supported Work (NSW) Demonstration, LaLonde (1986) concluded that non-experimental methods were prone to specification biases, gender-based differences in effect size, and significant differences in overall effect size compared to an experimental control. He recommended the use of longitudinal data and a two-step estimation technique originated by Heckman to reduce selection bias.

For the NSW example, the wages that workers earn might be affected by the workers selected to participate in the study—that is, a correlation between the workers that participate and the overall wage measurement. The conditional error from the wages equation can be used as an additional regressor in the wage equation to approximate the conditional error associated with participation. The multiple time points in longitudinal data reduce the risk of misspecification found with cross-sectional data.

Dehejia and Wahba applied propensity scores to determine the difference in calculated NSW treatment effect between the original study and the observational study proxy. They reported a non-experimental treatment difference in labor earnings for people undergoing labor training versus not. These results were comparable to the NSW experimental treatment results (Dehejia and Wahba, 1999). The sensitivity of these non-experimental results were affected by level of variable overlap between the studies is a contributing factor to the accuracy of the results (Dehejia and Wahba, 1999; Pfizer, 2006). Specifically, when the amount of overlap between the experimental and control units is high, there is less variability between different propensity score matching methods, and a smaller chance of misspecification.

Data

Three studies were identified either through a PubMed search or the ClinicalTrials.gov data registry site. Glaucoma studies were selected solely on the basis of data availability and its satisfaction of the inclusion criteria. There was no prior established preference for ophthalmalogic-specific data. The inclusion criteria were that the three studies had: the same outcome measure, and that the 'gold standard' study was a randomized clinical trial, where at least one of the comparison studies was an observational study.

All three studies evaluated the effectiveness of glaucoma treatment in a sample of patients using change in intraocular eye pressure (IOP) in mmHg from the start to the end of each study.

The first study (designated 'Study 1') was the “gold standard” study. It was a double-blinded parallel assignment randomized clinical trial of Xalatan (also labeled 'XT' henceforth) versus Xalacom (also labeled 'XC' henceforth). The study had a sample size of 289 patients that were drawn from 55 treatment centers in Japan. Each patient received one of the medications once daily for 8 weeks. The change in IOP was measured at 4 week and 8 week time intervals (Pfizer, 2006).

The second study (designated 'Study 2') was a 12 week parallel-assignment evaluator-blinded (but not patient-blinded) randomized study of Xalatan versus Dorzolamide/Timolol (also labeled 'D' henceforth; where Dorzolamide is a carbonic anhydrase inhibitor, while Timolol is a beta-blocker). There was an initial sample of 300 patients drawn from 25 centers across Europe. Of the original 300 patients, 270 patients were randomly allocated to one of the two treatment arms. Patients had a diagnosis of ocular hypertension or open-angle glaucoma. They were resistant to traditional beta-blocker therapy. Patients were further evaluated for their IOP three times at 4 weeks, and three times at 12 weeks (Miglior et al, 2010). With blinding occuring only on the evaluator side, this study is noted to have a weaker study design than Study 1.

The third study (designated 'Study 3') was a prospective non-interventional cohort study that included patients diagnosed with glaucoma or ocular hypertension. The study lasted for 3 years, and had an initial sample size of 28,812 patients, drawn from approximately 385 offices of different practicing ophthalmologists. The IOP of patients in each treatment arm was measured at baseline

(time=0), and annually thereafter. Over the course of the 3 year study, 2,487 patients were lost to attrition, leading to a final sample size of 26,835 patients. The study had four arms: a Xalatan arm, a Xalacom arm, a Beta-blocker arm, and an Other arm (where 'Other' represents no anti-hypertensive treatment, and is also labeled 'O' henceforth) (Pfizer, 2009).

Methods

Multiple quantitative techniques were tested on the Study 2 and Study 3 data to determine which ones most accurately replicate the results from Study 1. As Table 1 shows, the overall methods that were tested were: simple subtraction, linear regression, and propensity score matching (nearest neighbor, stratification, radius, and kernel methods). Each propensity score estimate calculates the Average effect of Treatment on the Treated (ATT). The ATT is calculated by first determining the difference between the treatment and a placebo control group within a single balanced block. The ATT is the average of the difference for all blocks satisfying the balancing test (Becker and Ichino, 2002).

Simple Subtraction Method:

The first method, simple subtraction, involves subtracting the published mean of Study 2 from the published IOP mean of Study 3. The logic behind this method is that the difference represents the net difference between Xalatan and Xalacom when not controlling for between-study differences in participant characteristics. Also, this method does not seek to minimize the within-study error through estimation techniques such as ordinary least squares or propensity score matching.

However, the mean is just one data point. More data is needed to do linear regression or propensity score matching. The mean and standard deviation of Study 2 and Study 3 are used to generate a normal distribution with 200 observations (100 treatment observations and 100 control observations) for Study 2, and 200 observations for Study 3. Four equations are generated to obtain four sets of data. Each set has 100 observations, and represents one arm of either Study 2 or 3. The 200 observation number was selected based on it being a reasonable simulation sample size per faculty input.

Simulation Process

Study 2:

- *Given a Mean and Standard Deviation of Ocular Pressure (mmHg) for each arm of Study 2*
- *Generate a normal distribution in STATA with a sample size of 100 observations per study arm, with the given mean and standard deviation for Study 2*
- *Use the 200 total observations as the data for linear regression and propensity score matching to be described below.*

Study 3:

- *Given a Mean and Standard Deviation of Ocular Pressure (mmHg) for each arm of Study 3*
- *Generate a normal distribution in STATA with a sample size of 100 observations, with the given mean and standard deviation for Study 3*
- *Use the 200 total observations as the data for linear regression and propensity score matching*

The data simulation involved setting the number of observations to 100 for each study arm, and

then generating the distribution of parameters for the 100 observations. Setting the number of observations in each study arm is done with the following STATA code: 'SET OBS 100'. Next, the mean and standard deviation for each arm of Study 2 and Study 3 are incorporated into a STATA 'GEN' command to obtain a set of 100 observations for each study arm.

The simulation equation for the treatment arm ('XT") of Study 2 is: GEN YXT2=-9.7+0.2*INVNORM(UNIFORM(), where 9.7=Mean ocular pressure of treatment arm in Study 2 (Xalatan), and 0.2= Standard deviation of ocular pressure for treatment arm in Study 2 (Xalatan). The simulation equation for the control arm ('Dorzolamide/Timolol') of Study 2 is: GEN YDT2=-9.5+0.3*INVNORM(UNIFORM()), where 9.5=Mean ocular pressure of control arm in Study 2 (Dorzolamide/Timolol), and 0.2= Standard deviation of ocular pressure for control arm in Study 2 (Dorzolamide/Timolol).

The simulation equation for the control arm ('Other') of Study 3 is: GEN YO3=66.9+13.3*INVNORM(UNIFORM(), where 66.9=Mean ocular pressure of control arm in Study 3 ('Other'), and 13.3= Standard deviation of ocular pressure for control arm in Study 3 ('Other'). The simulation equation for the treatment arm ('Xalacom') of Study 3 is: GEN YXC3=66.5+12.7*INVNORM(UNIFORM()) where 66.5=Mean ocular pressure of treatment arm in Study 3 (Xalacom), and 13.3= Standard deviation of ocular pressure for treatment arm in Study 3 (Xalacom).

The newly-derived Y-variables (YXT2, YDT2, YO3, YXC3) are then merged for each study. The variable Y2 is the merging of YXT2 and YDT2, and has 200 observations. Likewise, Y3 merges

YO3 and YXC3, and also has 200 observations.

Y2 and Y3 are set as the dependent variables in separate regression equations in order to determine the IOP difference between the experimental and control arms of Study 2 and Study 3. The linear regression approach determines what the difference in ocular pressure is between the treatment and control arms of a single study, after controlling for the age and gender of the study participants. This is done by setting-up a dummy variable, dummy2 and dummy3 (defined below), in each equation, controlling for age and gender.

The STATA code for the equation for Study 2 is: `regress y2 dummy2 age2 gender2.` The code for Study 3 is: `regress y3 dummy3 age3 gender3.`

The regression equations are formally specified below. The dependent variables, Y_{study_2} and Y_{study_3}, are the mean ocular pressure (in mmHg) for each observation in Study 2 and Study 3, respectively. Dummy variables, X_{dummy_2} and X_{dummy_3}, are equal to 1 when the observations are from the treatment group for Study 2 and Study 3, respectively. They are equal to 0 when the observations are from the control group. The variables, X_{age_2} and X_{age_3}, are continuous variables that represent the age of each observation in Study 2 and Study 3, respectively. Variables, X_{gender_2} and X_{gender_3}, are dichotomous variables for each observation, where X_{gender}=1 when the observation is female, and X_{gender}=0 when the observation is male.

Within-Study Regression Equations:

Study 2: $Y2 = \beta_{dummy\_2} * dummy2 + \beta_{age\_2} * age2 + \beta_{gender\_2} * gender2 + \varepsilon_{\_2}$

Study 3: $Y3 = \beta_{dummy\_3} * dummy3 + \beta_{age\_3} * age3 + \beta_{gender\_3} * gender3 + \varepsilon_{\_3}$

dummy2: dummy2=1 where treatment=1 (Xalatan=1, Dorzolamide/Timolol=0)

dummy2=0 where treatment=0 (Xalatan=0, Dorzolamide/Timolol=1)

dummy3: dummy3=1 where treatment=1 (Xalacom=1, 'Other – No antihypertensive'=0)

dummy3=0 where treatment=0 (Xalacom=0, 'Other – No antihypertensive'=1)

β_{dummy_2} and β_{dummy_3} represent the within-study treatment difference for Study 2 and Study 3. Each represents the difference in mean ocular pressure between the treatment and control observations, after adjustment for the age and gender of each observation. To calculate the overall between-study treatment difference for linear regression, one subtracts β_{dummy_2} from β_{dummy_3}.

Between-Study Treatment Difference:

$\beta_{dummy\_2} - \beta_{dummy\_3}$=Between-Study Treatment Difference

The next set of analyses is done by propensity score matching. Linear regression calculates the best-fit line based on the smallest sum of squares of differences between the actual and predicted observations. Propensity score matching generates predicted probabilities that an observation would be in one group versus another. Linear regression calculates the within-study treatment difference by

adjusting for the additive effect of the age and gender variables. Propensity score matching controls for these variables by calculating a propensity score for each of the 200 observations in the study. The score is derived from an equation including the dummy variable indicator of x, age variable, and gender variable. It is used to match similar scores for observations between study group arms, according to different matching methods. In this study, the nearest neighbor, radius, kernel, and stratification matching methods were performed, with at most 5% of cases being dropped for any treatment or control group (Becker and Ichino, 2002; D'Agostino, 1998; Guo et al, 2004).

In the propensity score approach, the 200 observations of either Study 2 or Study 3 must first have a propensity score assigned to each observation. First we perform a logistic regression in which the dependent variable is the observation's dummy variable value (1=treatment; 0=control) and where the independent variables are the age (a continuous variable) and gender of the observation (1=female; 0=male). The propensity score is the probability of being in the treatment group based on the computed logistic regression. The raw data for this calculation is a 200 X 3 dataset, with the three columns being the values for the dummy variable (1=treatment; 0=control), age variable, and gender variable for each observation. Once the propensity score is calculated for each observation, a fourth column appears in the dataset.

The next step is matching scores for observations in Study 2, and for the observations in Study 3. The four different matching techniques (nearest neighbor, stratification, radius, and kernel matching), are used to match observations where the dummy=1 (treatment) to the observations where the dummy=0 (control), based on the pairs that have close propensity scores. If this can be done, the 'balancing test' is successful. The difference in the Y-values for each set of paired observations will then

be calculated, aggregated across all pairs in Study 2, and across all pairs in Study 3. The aggregate difference in the Study 2 and Study 3 Y-values is the between-study difference.

First, propensity scores are derived for the dummy variables in Studies 2 and 3, dummy2 and dummy3 respectively. The *pscore.ado* program (Becker and Ichino, 2002) first specifies a probit or logit equation. Thereafter, the sample is repeatedly divided into intervals, until the average propensity score and mean for the treatment and control groups are equivalent (Becker and Ichino, 2002). If the means are not equivalent between groups (cannot satisfy the 'Balancing Test'), then the model has to be made less restrictive by not including just first order covariates in the probit or logit equation, but also higher order covariates, such as interaction terms. Besides *pscore,* the command *comsup* activates the common support function, while the number of equally-spaced bins used in the balancing test are specified using the *numblo* command (Becker and Ichino, 2002). The Study 2 propensity score equation is specified using the *pscore* command: *pscore d2 a2 g2, pscore(mypscore1) blockid(myblock1) comsup numblo(5) level(0.005) logit.* Likewise, the Study 3 propensity score equation is: *pscore dummy3 age3 gender3, pscore(mypscore) blockid(myblock) comsup numblo(5).*

After the propensity scores are derived, the specific balancing tests are performed. The tests are nearest neighbor (ATTND), kernel (ATTK), radius (ATTR), and stratification (ATTS). Each test performs balancing within each block after the propensity score has been calculated for all blocks using the *pscore* command. The ATTND command matches experimental and control propensity scores based on nearest neighbor matching, a random selection of an experimental propensity score, and identification of the closest control propensity score within the same block. The ATTK command has a matching method that calculates a distance between observations based on a weighting formula using a

kernel function. The ATTR command matches the experimental propensity score with a control score limited to a certain pre-established distance (the radius). The ATTS command matches based on stratification, where propensity scores are divided into quintiles (strata), and the experimental and control propensity scores within the same strata are matched (Becker and Ichino, 2002; D'agostino, 1998; Guo et al, 2004).

The ATTND command is: *attnd y2 dummy2 age2 gender2, comsup bootreps(100) dots logit and attnd y3 dummy3 age3 gender3, comsup bootreps(100) dots logit.* The ATTK command is: a*ttk y2 d3 a3 g3, comsup bootreps(100) and attk y3 d3 a3 g3, comsup bootreps(100).* The ATTR command is: *attr y2 d3 a3 g3, comsup bootreps(100) dots logit and attr y2 d3 a3 g3, comsup bootreps(100) dots logit.* The ATTS command is: *atts y2 d2 a2 g2, pscore(mypscore1) blockid(myblock1) bootstrap* and *atts y3 d3 a3 g3, pscore(mypscore) blockid(myblock) bootstrap.*

The ATT is determined for Study 2 and Study 3 for each matching method. The final treatment difference between Study 2 and Study 3 is determined by the expression: Study 3 ATT – Study 2 ATT. This net ATT difference is calculated for each method and represents the “between-study treatment difference.” The 'between-study' results from linear regression, simple subtraction, and propensity score matching are compared to the within-study difference in Study 1. The percent difference between the Study 1 result and the experimental results are compared.

Data

Figure 1 provides the characteristics of each study. Notable differences include the fact that both Study 1 and Study 2 are randomized, but only Study 1 is a double-blinded study. In addition, while the

sample size for Study 1 and Study 2 are roughly comparable at between 240-290 participants, Study 3 has a sample size that is much larger at 9,830 participants. Participant age is roughly similar for Study 2 and Study 3, but age in Study 1 was stratified into age groups. Finally, while Study 2 and Study 3 have roughly equal gender ratios, Study 3 has a greater proportion of males in in the 'Other' category. The treatment difference between Xalatan and Xalacom in Study 1, Study 2, and Study 3 were respectively: 0.97 mmHg, -0.20 mmHg, and 0.90 mmHg. More information about each individual study is available in the Appendix.

Results

Simple Subtraction Method

Study 2 Difference – Study 3 Difference: -0.2-0.9=**-1.10 mmHg**

The first approach that is evaluated is the simple subtraction method. In this method, the published IOP number from Study 2 is simply subtracted from the published IOP number in Study 3 without controlling for participant age or gender. For example, in Study 3, the published IOP value for Xalacom (0.2) is subtracted from the published value for 'Other' (-1.1), where 0.2-(-1.1) = 0.9 mmHg. Furthermore, no assumptions are made about the probability distribution for the published IOP values. Based on this, the simple subtraction method calculates the -1.1 mmHg difference between Xalatan and Xalacom.

The second approach is the linear regression approach. In this method, two separate regression equations are specified. The variables 'Age' and 'Gender' are independent variables along with a dummy

variable for treatment. The dummy variable equals 1 for the treatment-group (Xalatan or Xalacom), and 0 for the control group (Dorzolamide/Timolol or 'Other Medications – No antihypertensive'). As Figure 2 shows, the Study 2 dummy variable shows a -0.17 mmHg treatment difference between Xalatan and Dorzolamide/Timolol. There is a -1.03 mmHg difference between Xalacom and 'Other Medications'. The net treatment difference between Xalatan and Xalacom is calculated to be 0.86 mmHg.

The final approach is the application of three propensity score methods. Again, the variables 'Age' and 'Gender' were independent variables. Kernel, radius, and nearest neighbor matching are selected. All three methods successfully balanced the covariates. The average treatment effect on the treated (ATT) is calculated by each method for each study. The ATT is then differenced between studies for each method to obtain the overall treatment difference. Figure 3 shows that the nearest neighbor, kernel, and radius matching result in a Xalatan to Xalacom treatment difference of -1.56 mmHg, 0.788 mmHg, and 0.77 mmHg, respectively. The large difference of the nearest neighbor method in particular is in-line with research showing it as one of the poorer propensity score balancing methods (Huber et al, 2010). The method is included here for the purposes of comparison.

Figure 4 shows the inter-study treatment differences for all the tested methods. Simple subtraction and nearest neighbor are furthest from the gold standard reference with values of -1.1 mmHg and -1.56 mmHg, respectively. Linear regression, kernel matching, and radius matching are all fairly close to the gold standard with values of 0.86 mmHg, 0.79 mmHg, and 0.77 mmHg, respectively.

The standard errors of each method are next determined in order to calculate the t-statistic

between the gold standard and each method. If the difference between the gold standard value and the experimental method value is sufficiently close, then the values will be assumed to be from the same population distribution. However, if the standardized difference has a p-value below 0.05, then the experimentally-derived value is significantly different from the gold standard value. The two values cannot be assumed to arise from the same population distribution.

Figure 5 provides the full table of pooled standard errors from the gold standard and each experimental method. The simple subtraction errors are obtained directly from the published results. The linear regression standard error is the calculated standard error for the dummy variable in each study's regression equation. The propensity score standard error is the calculated standard error for the ATT in each study.

Two sample t-tests with unequal variances are performed to evaluate each experimental method against the gold standard. As Figure 6 illustrates, only the simple subtraction method yields a result that is significantly difference from the gold standard value. The null hypothesis is rejected, and it is assumed that the simple subtraction between-study value arises from a different distribution than the gold standard value. The linear regression, nearest neighbor, radius matching, and stratification matching procedures all have p-values>0.05. For a reason that could not be determined, no numerical value was produced on attempting to calculate the kernel matching-based value. STATA was not contacted regarding this. A 'not available' designation (N/A) was included in Table 7. The null hypothesis of their values being from the same distribution as the gold standard cannot be rejected.

Conclusion

Previous research has shown the limited accuracy of replicating RCT results through the application of quantitative methods on non-experimental data. This paper evaluates the application of simple subtraction, linear regression, and propensity score matching to obtaining a treatment difference from two non-equivalent studies. It then compares the derived result from the actual treatment difference obtained from an RCT.

In the simple scenario using only two independent variables, this study finds that certain methods are more accurate than others in replicating RCT results. In particular, linear regression, nearest neighbor matching, stratification matching, kernel matching, and radius propensity score matching methods did not reject the null hypothesis of the merged study results being from the same population as the Study 1 results. Stratification matching was the closest to the gold standard at 1.03 mmHg, approximately 6.2% away from the RCT result. The linear regression result of a Xalatan to Xalacom treatment difference of 0.86 mmHg was 11.3% away from the RCT result of 0.97 mmHg.

Kernel and Radius matching were even less accurate in their predictions, despite the fact that the results were relatively close to one another. The kernel matching method calculated a 0.79 mmHg treatment difference between the two medications, an 18.6% difference from the RCT result. Radius matching calculated a 0.77 mmHg treatment difference between Xalatan and Xalacom, a 20.6% difference from the RCT value. With a calculated value of -1.56 mmHg, nearest neighbor was 261% away from the RCT value. It was the furthest of all the methods that still rejected the null hypothesis.

In contrast to the above mentioned methods, the simple subtraction method did reject the null hypothesis. Like nearest neighbor matching, it predicted a negative value for the treatment difference

between Xalatan and Xalacom. The opposite sign was due to the method predicting a positive value for the Study 3 treatment difference (1.1 mmHg). In contrast, linear regression, kernel matching, stratification matching, and radius matching all computed negative values for Study 3 (-1.03 mmHg, -0.953 mmHg, and -0.941 mmHg).

The cumulative results appear to demonstrate the feasibility of two non-equivalent studies proxying for the results of a third experimental study. Simple subtraction and nearest neighbor matching do not match within-study observations effectively. A cautious observation is that the accuracy of the quantitative method is directly based on the robustness of controlling or matching across covariates within the study. Robustness here depends on successful convergence for the propensity score matching algorithm, but can be further evaluated by simulation testing. In terms of its implications, the requirement for successful convergence limits how extensively the technique may be used, especially when there are many covariates between experimental and placebo groups.

While the results appear tentatively promising, there are important limitations to this work. Especially important is there are difficulties in generalizing these results. Reasons for this include that there are only two independent variables in Studies 2 and 3. Linear regression methods in particular are expected to perform poorly as there is an increase in the number of non-overlapping between-study variables. On the other hand, propensity score methods generate an index of all independent variables to create the propensity score; thereby limiting the effect of a single non-overlapping variable in skewing the results.

Another limitation is that the Study 2 and 3 results were assumed to be simulated with a normal

distribution assumption. Applying linear regression to a normally distributed dependent variable helps ensure that ordinary least squares (OLS) is an appropriate estimator. However, it is possible that there may be study-specific characteristics that may make the distribution non-normal. This study does not make this assumption, and as a result, may produce results that are biased towards the tested estimation techniques.

Assuming that there is no treatment time-specific difference in study results is another important qualifier to these findings. The patient treatment periods were substantially different between Study 3 and the other two studies. Patients were treated for 8 weeks in Study 1, 12 weeks in Study 2, and 1 year in Study 3. It is possible that the treatment-specific effect is different both at 8 and 12 weeks, as well as at the further 1 year time-point. In this case, it would not be appropriate to assume a constant treatment effect over time, but would instead be necessary to weight the treatment results by the treatment period.

The treatment time difference might have a minor effect if treatment effect beyond 8 weeks is relatively constant.

There are unequal sample sizes in the original data between Study 3 versus Study 1 and 2. The smaller samples sizes in Study 1 and Study 2 might be insufficient to capture the true effect size. Furthermore, the average age of participants may be markedly lower in Study 1 versus Study 2 and Study 3. The reference treatment effect size in Study 1 may be skewed based on age-dependent factors. A final limitation is that Study 3 has only two of the four study groups, the 'Xalacom' and 'Other' groups, used in the ultimate analysis. It is possible that performing one-to-one testing in this setting leads to biased results, where the 'Other' category is not the optimal control. As well, Study 2 is a

randomized trial, not an observational study. It is tentatively inferred that ambiguities on the nature of the variability between studies may be clarified to some extent with a few of these tested methods. However, in differentiating between systematic and random between-study variability, the level of covariate overlap between studies would be critical in exactly determining the robustness of the calculated between-study effect size.

Strengths to this work are that this study compares the results from multiple quantitative approaches. Significance testing is done as well to determine whether the method-specific calculation is significantly different from the gold standard. The selected studies meet the criteria of having the same Y-variable of ocular pressure (in mmHg), and Study 1 being a double-blinded RCT while Study 2 and Study 3 have weaker study designs. Finally, the three studies are independent from one another further, limiting the effect of correlated independent variables biasing the treatment difference results.

Future work might further evaluate the robustness of these findings for different medications. Further areas of extension include testing these methods for small samples, and across different time-intervals. It would be especially helpful if the precise bounds by which this method achieves unbiased estimates are determined. Simulation tests would be an excellent means for determining these bounds.

Overall, this study demonstrates that in certain limited circumstances, statistical methods applied to non-experimental and non-double-blinded experimental data can approximate the results of a double-blinded RCT. Propensity score matching is conjectured to perform better than the other tested methods because it is used to explicitly balance the covariates between a placebo control and experimental group, in the process randomizing the study. If further validated, this approach might be

especially useful in comparative effectiveness research and in clinical trial studies.

References

Becker S and Ichino A. “Estimation of average treatment effects based on propensity scores”. *Stata Journal*. Vol. 2(4) (2002), pages 358-377. Website: http://www.stata-journal.com/article.html?article=st0026

Cochran WG. “The Combination of Estimates from Different Experiments”. *Biometrics* , Vol. 10, No. 1 (Mar., 1954), pp. 101-129. Website: http://www.jstor.org/stable/3001666

D'Agostino Jr, RB. “Tutorial in Biostatistics: Propensity Score Methods for Bias Reduction in the Comparison of a Treatment to a Non-Randomized Control Group”. *Statist. Med*. 1998:17, 2265-2281 .Website: http://web.pdx.edu/~nwallace/EPA/Dagostino1998.pdf

Dehejia RH, Wahba S. “Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation Training Programs”. *Journal of the American Statistical Association*. December 1999. Vol.94 (448): 1053-62. Website: http://www.jstor.org/stable/2669919

Dehejia RH, Wahba S. “Propensity Score-Matching Methods for Nonexperimental Causal Studies”. *The Review of Economics and Statistics*. February 2002, 84(1): 151-61. Website: http://www.nber.org/~rdehejia/papers/matching.pdf

Guo S, Barth R, and Gibbons C. “Introduction to Propensity Score Matching: A New Device for Program Evaluation”. Annual Conference of the Society for Social Work Research. January 2004. Website: http://ssw.unc.edu/VRC/Lectures/PSM_SSWR_2004.ppt

Higgins JP, Thompson SG. "Quantifying heterogeneity in a meta-analysis". *Stat Med.* 2002 Jun 15;21(11):1539-58.

Huber M, Lechner M, Wunsch C. "How to Control for Many Covariates? Reliable Estimators Based on the Propensity Score". IZA: Institute for the Study of Labor. Discussion Paper No. 5268. October 2010. Website: http://ftp.iza.org/dp5268.pdf

Miglior S, Grunden JW, Kwok K, Xalacom/Cosopt European Study Group. "Efficacy and safety of fixed combinations of latanoprost/timolol and dorzolamide/timolol in open-angle glaucoma and ocular hypertension." *Eye*. 2010 Jul; 24(7): 1234-42. Website: http://www.nature.com/eye/journal/v24/n7/fig_tab/eye2009307t3.html#t3-fn4

Olkin I. Statistical and theoretical considerations in meta-analysis. *J Clin Epidemiol*. 1995 Jan;48(1):133-46; discussion 147.

Pfizer. "A Study Comparing Xalacom And Xalatan In Patients With Primary Open Angle Claucoma (POAG) Or Ocular Hypertension (OH)". Data first received September 29, 2006. Data last updated August 3, 2009. Website: http://clinicaltrials.gov/ct2/show/results/NCT00383019?term=xalatan&recr=Closed&rslt=With&type=Intr&rank=2&sect=X6015&view=results#outcome1

Pfizer. "Non-Interventional Study in Patients With Ocular Hypertension And Open Angle Glaucoma Treated With Xalatan and Xalacom (XCHANGE). ClinicalTrials.gov Website. Data received November 11, 2009. Data last updated February 18, 2010. Website:

http://www.clinicaltrials.gov/ct2/show/results/NCT01012245?term=xalatan&recr=Closed&rslt=With&type=Obsr&rank=2&sect=X6015#outcome1

Thompson SG, Sharp SJ. "Explaining heterogeneity in meta-analysis: a comparison of methods". *Stat Med*. 1999 Oct 30;18(20):2693-708.

Appendix

Study Tables

Table 1.Specified Method for Each Evaluated Quantitative Technique

Type	Method
Simple Subtraction	Direct subtraction of between-study published mean values
Linear Regression	Data simulation and subtraction of between-study dummy variables
Propensity Scores: Radius Matching	Data simulation and subtraction of between-study ATT
Propensity Scores: Kernel Matching	Data simulation and subtraction of between-study ATT
Propensity Scores: Nearest Neighbor Matching	Data simulation and subtraction of between-study ATT
Propensity Scores: Stratification Matching	Data simulation and subtraction of between-study ATT

Table 2. Study Characteristics

	Study 1	**Study 2**	**Study 3**
Experimental Structure	Double-blinded Randomized Parallel Assignment Study (Randomized)	Randomized Evaluator-Blinded Controlled Trial (Randomized)	Prospective Non-Interventional Cohort (Observational)
Sample Size	144 (XC)/145 (XT)/ 289 (T)	238	434 (XC)/9336 (O)/9830 (T)
Participant Age	XT: 18-65 yrs: 78 obs, >65 yrs: 66 obs XC: 18-65 yrs: 77 obs, >65 yrs: 68 obs	XT: 65.8 yrs D: 66.6 yrs	XC: 66.5 yrs O: 66.9 yrs
Participant Gender	XT: 70 obs (F), 74 obs (M) XC: 74 obs (F), 71 obs (M)	XT: 67 obs (M) 68 obs (F) D: 54 obs (M), 58 obs (F)	XC: 383 obs (F), 363 obs (M) O: 7493 obs (F), 5203 obs (M)
Treatment Difference (mmHg)	0.97	-0.20	0.90

Table 3. Linear Regression Method:

	Dummy Coefficient Value
Study 2	-0.17
Study 3	-1.03

S2-S3=-0.1689-(-1.0268)= **0.86**

Table 4. Propensity Score Method:

	Kernel	Radius	Nearest Neighbor	Stratification
Study 2	-0.17	-0.17	-0.197	-0.18
Study 3	-0.94	-0.95	1.359	-1.21

Interstudy Difference (Propensity Score-Derived)

Nearest Neighbor: S2-S3=(-0.197)-1.359=**-1.56 mmHg**

Kernel Matching: S2-S3=-0.165+0.953= **0.788 mmHg**

Radius Matching: S2-S3=-0.171+0.941=**0.77 mmHg**

Stratification Matching: S2-S3=-0.18+1.21=**1.03 mmHg**

Table 5. Aggregate Results: Different Approaches

Approach	Inter-study Treatment Difference (mmHg)
Gold Standard	0.97
Linear Regression	0.86
Propensity Scores: Nearest Neighbor	-1.56
Propensity Scores: Kernel Matching	0.79
Propensity Scores: Radius Matching	0.77
Propensity Scores: Stratification Matching	1.03
Simple Subtraction	-1.10

Table 6. Table of Standard Errors:

	Gold Standard	**Simple Subtraction**	**Linear Regression**	**Nearest Neighbor PScore**	**Kernel PScore**	**Radius PScore**	**Stratification PScore**
1st Study Standard Error	0.17	0.167	1.85	0.061	N/A	0.04	0.04
2nd Study Standard	0.17	0.048	0.039	2.344	N/A	1.85	1.84

Error							
Pooled Standard Error	0.24	0.174	1.85	2.344	N/A	1.85	1.84
Mean difference	0.97	-1.1	0.86	-1.56	0.79	0.77	1.03
Total n	289	10,040	200	307	395	395	394

Table 7. Significance Testing of the Difference between the Method-Specific Calculated IOP and the Gold Standard

Approximation Type	Null Hypothesis: Difference between Gold Standard and Selected Test is equal to 0 [2]

[2] The alternative hypothesis is rejected at alpha<0.05

Simple Subtraction	p-value=0.0000 (reject H_0)
Linear Regression	p-value=0.9530 (do not reject H_0)
Nearest Neighbor PScore	p-value=0.28 (do not reject H_0)
Kernel PScore	N/A
Radius PScore	p-value=0.9147 (do not reject H_0)
Stratification PScore	p-value=0.9742 (do not reject H_0)

Study Backgrounds

a. Study 1

Study 1 had a total of 289 participants divided between the Xalacom and Xalatan group. The 8 week medication treatment timeframe was selected for this paper. The Xalacom-treatment group had a -2.59 mmHg average reduction in ocular pressure over this period. There was a -1.62 mmHg reduction in ocular pressure in the Xalatan-treatment group. The standard errors for the Xalacom and Xalatan treatment groups were relatively small at 0.17 mmHg and 0.17 mmHg, respectively. Subtracting the treatment difference between groups results in Xalatan-treated patients having a 0.97 mmHg higher ocular pressure than their Xalacom-treated counterparts.

Study 1 Actual Results3

	Xalacom Group	**Xalatan Group**
Number of Participants Analyzed [units: participants]	144	145
Change of Intraocular Pressure (IOP) From Baseline to Week 8 [units: mmHg] Least Squares Mean (95% Confidence Interval)	-2.59 (-2.92 to -2.25)	-1.62 (-1.96 to -1.28)

Study 1 Difference: Xalatan-Xalacom: -1.62-(-2.59) = 0.97 mmHg

b. Study 2

[3] Study 1 Reference: Pfizer. "A Study Comparing Xalacom And Xalatan In Patients With Primary Open Angle Claucoma (POAG) Or Ocular Hypertension (OH)". Data first received September 29, 2006. Data last updated August 3, 2009. Website: http://clinicaltrials.gov/ct2/show/results/NCT00383019?term=xalatan&recr=Closed&rslt=With&type=Intr&rank=2&sect=X6015&view=results#outcome1

Study 2 had a total sample of 238 patients. The Xalatan-treated group had 121 patients while the control group, receiving treatment with Dorzolamide/Timolol, had 117 patients. The ocular pressures were measured 4 times daily over a 12 week period. The multiple time interval measurements were averaged to obtain an overall 12-week treatment difference. The Xalatan-treated group had an ocular pressure change of -9.725 mmHg, while the control Dorzolamide/Timolol-treated group had a -9.525 mmHg ocular pressure reduction. Xalatan-treated patients had a -0.200 mmHg ocular pressure than the control group. T-test calculation of the between-group treatment difference was not significant for all time intervals at the 12 week period.

Study 2 Actual Results4

[4] Study 2 Reference: Miglior S, Grunden JW, Kwok K, Xalacom/Cosopt European Study Group. "Efficacy and safety of fixed combinations of latanoprost/timolol and dorzolamide/timolol in open-angle glaucoma and ocular hypertension." Eye. 2010 Jul; 24(7): 1234-42. Website: http://www.nature.com/eye/journal/v24/n7/fig_tab/eye2009307t3.html#t3-fn4

Measurement time	*IOP change from baseline to week 12 (mm□Hg) least square mean (SE)*		*95% CI*	*Between-group* **P-***value*
	Fixed-combination latanoprost/timolol (*N=*121)	***Fixed-combination dorzolamide/timolol (*N=*117)***		
Daytime	−9.7 (0.2)	−9.5 (0.2)	−0.8, 0.4	0.51
0800 hours	−9.8 (0.2)	−9.5 (0.3)	−0.96, 0.3	0.32
1200 hours	−9.8 (0.2)	−9.7 (0.3)	−0.8, 0.5	0.72
1600 hours	−9.6 (0.2)	−9.4 (0.3)	−0.9, 0.4	0.43

Study 2 Difference: Xalatan – Dorzolamide/Timolol: -9.725-(-9.525)= -0.200 mmHg

c. Study 3

Study 3 had 4 treatment groups and three time interval measurements. To approach the 8 week and 12 week intervals for studies 1 and 2, the shortest time interval of 1 year was selected. The Xalacom and Other Medications groups were selected for the head-to-head testing. The sample sizes of the Xalacom and 'Other Medications' groups at 1 year were 209 and 3029 patients, respectively. The change in IOP was -0.2 mmHg and -1.1 mmHg for Xalacom and the 'Other Medications' group. Each group had wide standard errors. The Xalacom group had a 0.9 mmHg ocular pressure difference.

Study 3 Actual Results5

[5] Study 3 Reference: Pfizer. "Non-Interventional Study in Patients With Ocular Hypertension And Open Angle Glaucoma Treated With Xalatan and Xalacom (XCHANGE). ClinicalTrials.gov Website. Data received November 11, 2009. Data last updated February 18, 2010. Website:

	All Subjects	Xalatan	Betablockers	Xalacom	Other Medications
Number of Total Participants Analyzed	20073	8735	1508	494	9336
Change in IOP: 1 year (n=10886,4134, 357, 209, 3029)	-1.5 ± 4.4	-0.7 ± 3.5	-1.0 ± 3.4	-0.2 ± 3.5	-1.1 ± 4.6

Study 3 Difference: Xalacom – Other: -0.2-(-1.1)= 0.9 mmHg

Results from the Analysis

Linear Regression Results

Study 2 Regression Results

	SS	df	MS
Model	1.73	3	0.58
Residual	14.38	196	0.07
Total	16.11	199	0.08

```
Number of obs =     200
F(  3,   196) =    7.84
```

http://www.clinicaltrials.gov/ct2/show/results/NCT01012245?term=xalatan&recr=Closed&rslt=With&type=Obsr&rank=2&sect=X6015#outcome1

```
Prob > F      =  0.0001
R-squared     =  0.1071
Adj R-squared =  0.0934
Root MSE      =  .27091
```

	Coef.	Std. Err.	t	P-value	95% Confidence Interval
d2	-0.17	0.04	-4.37	0.00	(-0.25,-0.09)
a2	0.00	0.00	0.72	0.47	(-0.00,0.00)
g2	-0.05	0.04	-1.38	0.17	(-0.13,0.02)
Constant	-9.58	0.13	-76.53	0.00	(-9.83,-9.34_

Study 3 Regression Results

	SS	df	MS
Model	151.58	3	50.53
Residual	33525.96	196	171.05
Total	33677.53	199	169.23

```
Number of obs =     200
F(  3,   196) =    0.30
Prob > F      =  0.8287
R-squared     =  0.0045
Adj R-squared = -0.0107
```

Root MSE = 13.079

	Coef.	Std. Err.	t	P-value	95% Confidence Interval
d2	-1.03	1.85	-0.55	0.58	(-4.68,2.62)
a2	-0.02	0.08	-0.20	0.84	(-0.17,0.13)
g2	1.46	1.90	0.77	0.44	(-2.28,5.21)
Constant	67.95	5.14	13.22	0.00	(57.82,78.09)

Study 2 Propensity Score Results

1. Nearest Neighbor Matching

Logistic regression — Number of obs = 200; LR chi2(2) = 3.30; Prob > chi2 = 0.1922

Log likelihood = -136.9802 — Pseudo R2 = 0.0119

	Coef.	Std. Err.	z	P-value	95% Confidence

					Interval
a2	-0.01	0.01	-1.12	0.26	(-0.04,0.01)
g2	0.42	0.29	1.47	0.14	(-0.14,0.99)
Constant	0.82	0.91	0.90	0.37	(-0.97,2.61)

ATT estimation with Nearest Neighbor Matching method (random draw version)
Analytical standard errors

n. treat	n. contr.	ATT	Std. Err.	t
100	49	-0.20	0.06	-3.22

2. Kernel Matching

ATT estimation with the Kernel Matching method

n. treat	n. contr.	ATT	Std. Err.	t
100	95	-0.17	--	--

3, Radius Matching

ATT estimation with the Radius Matching method

n. treat	n. contr.	ATT	Std. Err.	t
100	95	-0.17	0.04	-4.13

4. Stratification Matching

ATT estimation with the Stratification method

n. treat	n. contr.	ATT	Std. Err.	t
100	95	-0.18	0.04	-4.19

Study 3 Propensity Score Results

1. Nearest Neighbor Matching

```
Logistic regression                               Number of obs   =        200
                                                  LR chi2(2)      =       0.25
                                                  Prob > chi2     =     0.8834
Log likelihood = -138.50549                       Pseudo R2       =     0.0009
```

	Coef.	Std. Err.	z	P-value	95%

					Confidence Interval
a3	-0.00	0.01	-0.41	0.68	(-0.03,0.02)
g3	0.10	0.29	0.35	0.73	(-0.47,0.67)
Constant	0.27	0.77	0.35	0.73	(-1.24,1.78)

ATT estimation with Nearest Neighbor Matching method (random draw version)
Analytical standard errors

n. treat	n. contr.	ATT	Std. Err.	t
100	58	1.36	2.34	0.58

2. Kernel Matching

ATT estimation with the Kernel Matching method

n. treat	n. contr.	ATT	Std. Err.	t
100	100	-0.94	--	--

3, Radius Matching

ATT estimation with the Radius Matching method
Analytical standard errors

n. treat	n. contr.	ATT	Std. Err.	t
100	100	-0.95	1.85	-0.52

4. Stratification Matching

ATT estimation with the Stratification method

n. treat	n. contr.	ATT	Std. Err.	t
100	99	-1.21	1.84	-0.66

Standard Error Calculations

Gold Standard (GS):
Xalacom SE: SExc=0.17
Xalatan SE: SExt=0.17
Pooled SE=√(0.17^2+0.17^2)=0.24 [for n=289 for the two treatment arms]

Simple Subtraction Method (SS):
Study 2 Standard Errors
Xalatan SE: SExt=0.2
Dorzolamide SE: Sedz=0.2

Pooled Study 2 SE=√(0.2^2+0.2^2)=0.28 [for n=270 for the two treatment arms]

Study 3 Standard Errors
Xalacom SE: SExc=3.4/[√(494)]=0.15,
Other SE : SEo=4.6/[√(9334)]=0.21,
Pooled Study 3 SE=(0.15^2+0.21^2)^0.5=0.26 [for n=9,770 for the two treatment arms]

Overall Study 2-Study 3 Pooled SE=(0.26^2+0.28^2)^0.5=0.38 [for n=10,040 for the combined studies 2 and 3]

Linear Regression Method (LR):
Study 3 SE: Ses3=1.85
Study 2 SE: Ses2=0.039
Pooled SE=√(0.039^2+1.85^2)=1.85 [for n=400 for the combined studies 2 and 3]

Propensity Score Method (PS): Nearest Neighbor
Study 2 SE: 0.061 (for n=149)
Study 3 SE: 2.344 (for n=158)
Overall Study 2-Study 3 Pooled SE=(0.061^2+2.344^2)^0.5=2.344
[for n=307 for the combined studies 2 and 3]

Propensity Score Method (PS): Kernel Matching
Study 2 SE: N/A (for n=195)
Study 3 SE: N/A (for n=200)
Overall Study 2-Study 3 Pooled SE: N/A

Propensity Score Method (PS): Radius Matching
Study 2 SE: 0.04 (for n=195)
Study 3 SE: 1.85 (for n=200)
Overall Study 2-Study 3 Pooled SE=(0.04^2+1.85^2)^0.5=1.85
[for n=395 for the combined studies 2 and 3]

Propensity Score Method (PS): Stratification Matching
Study 2 SE: 0.04 (for n=195)
Study 3 SE: 1.84 (for n=199)
Overall Study 2-Study 3 Pooled SE=(0.04^2+1.84^2)^0.5=1.84
[for n=394 for the combined studies 2 and 3]

T-Test Calculations:

Gold Standard versus Simple Subtraction

```
. ttesti 289 0.97 4.08 10040 -1.1 28.06, une

Two-sample t test with unequal variances
```

	Obs	Mean	Std. Error	Std. Dev.	95% Confidence Interval
x	289	0.97	0.24	4.08	(0.50,1.44)
y	10,040	-1.1	0.28	28.06	(-1.65,-0.55)
combined	10,329	-1.04	0.27	27.68	(-1.58,-0.51)
diff		2.07	0.37		(1.35,2.79)

```
    diff = mean(x) - mean(y)                                      t =   5.6126
Ho: diff = 0                     Satterthwaite's degrees of freedom =  1524.99

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

Gold Standard versus Linear Regression

```
. ttesti 289 0.97 4.08 400 0.86 37, une
```

Two-sample t test with unequal variances

	Obs	Mean	Std. Error	Std. Dev.	95% Confidence Interval
x	289	0.97	0.24	4.00	(0.50,1.44)
y	307	0.86	1.85	37	(-2.78,4.50)
combined	689	0.91	1.08	28.30	(-1.21,3.02)
diff		0.11	1.87		(-3.56,3.78)

```
    diff = mean(x) - mean(y)                                      t =   0.0590
Ho: diff = 0                     Satterthwaite's degrees of freedom =  412.381

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.5235         Pr(|T| > |t|) = 0.9530          Pr(T > t) = 0.4765
```

Gold Standard versus Propensity Score (Nearest Neighbor)

```
. ttesti 289 0.97 4.08 307 -1.56 41.07, une

Two-sample t test with unequal variances
```

	Obs	Mean	Std. Error	Std. Dev.	95% Confidence Interval
x	289	0.97	0.24	4.08	(0.50,1.44)
y	307	-1.56	2.34	41.07	(-6.17,3.05)
combined	596	-0.33	1.21	29.62	(-2.72,2.05)
diff		2.53	2.36		(-2.11,7.17)

```
    diff = mean(x) - mean(y)                                      t =   1.0737
Ho: diff = 0                     Satterthwaite's degrees of freedom =  312.413

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.8581         Pr(|T| > |t|) = 0.2838          Pr(T > t) = 0.1419
```

Gold Standard versus Propensity Score (Radius Matching)

```
. ttesti 289 0.97 4.08 395 0.77 36.77, une
```

Two-sample t test with unequal variances

	Obs	Mean	Std. Error	Std. Dev.	95% Confidence Interval
x	289	0.97	0.24	4.08	(0.50,1.44)
y	395	0.77	1.85	36.77	(-2.87,4.41)
combined	684	0.85	1.07	28.05	(-1.25,2.96)

diff		0.2	1.87		(-3.47,3.87)

```
 diff = mean(x) - mean(y)                                      t =   0.1072
Ho: diff = 0                     Satterthwaite's degrees of freedom =  407.214

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.5427         Pr(|T| > |t|) = 0.9147          Pr(T > t) = 0.4573
```

Gold Standard versus Propensity Score (Kernel Matching)

```
Not available because unable to obtain standard errors for kernel matching.
```

Gold Standard versus Propensity Score (Stratification Matching)

```
. ttesti 289 0.97 4.08 394 1.03 36.52, une

Two-sample t test with unequal variances
```

	Obs	Mean	Std. Error	Std. Dev.	95% Confidence

					Interval
x	289	0.97	0.24	4.08	(0.50, 1.44)
y	394	1.03	1.84	36.52	(-2.59,4.65)
combined	683	1.00	1.07	27.85	(-1.09,3.10)
diff		-0.06	1.86		(-3.71,3.59)

```
    diff = mean(x) - mean(y)                                      t =  -0.0323
Ho: diff = 0                     Satterthwaite's degrees of freedom =  406.328

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.4871         Pr(|T| > |t|) = 0.9742          Pr(T > t) = 0.5129
```

STATA Code

Data Simulation

```
. set mem 1g

. set obs 100

. set seed 2000

. gen YXT2=-9.7+0.2*invnorm(uniform())

. gen YDT2=-9.5+0.3*invnorm(uniform())
```

```
. gen YO3=66.9+13.3*invnorm(uniform())

. gen YXC3=66.5+12.7*invnorm(uniform())

. gen A2XT=65.8+11.3*invnorm(uniform())

. gen A2D=66.6+10.0*invnorm(uniform())

. gen A3XC=64.9+13.4*invnorm(uniform())

. gen A3O=66.9+13.3*invnorm(uniform())
```

Nearest Neighbor

```
. pscore  dummy3 age3 gender3, pscore(mypscore) blockid(myblock)
comsup numblo(5)
level(0.005) logit

. attnd y3  dummy3 age3 gender3, comsup bootreps(100) dots logit

. pscore dummy2 age2 gender2, pscore(mypscore2) blockid(myblock2)
comsup numblo(5) level(0.005) logit

. attnd y2  dummy2 age2 gender2, comsup bootreps(100) dots logit
```

Kernel Matching

```
. attk y3 d3 a3 g3, comsup bootreps(100) dots logit
. attk y2 d3 a3 g3, comsup bootreps(100) dots logit
```

Radius Matching

```
. attr y2 d3 a3 g3, comsup bootreps(100) dots logit
. attr y2 d3 a3 g3, comsup bootreps(100) dots logit
```

Stratification

```
.pscore d3 a3 g3, pscore(mypscore) blockid(myblock) comsup numblo(5)
level(0.005) logit

. atts y3 d3 a3 g3, pscore(mypscore) blockid(myblock) bootstrap


.pscore d2 a2 g2, pscore(mypscore1) blockid(myblock1) comsup
numblo(5) level(0.005) logit

. atts y2 d2 a2 g2, pscore(mypscore1) blockid(myblock1) bootstrap
```

Regression

```
. regress y3  dummy3 age3 gender3
```

```
. regress y2  dummy2 age2 gender2
T-tests
. ttesti 289 0.97 0.24 10040 -1.1 0.174, une

. ttesti 289 0.97 0.24 200 0.86 1.85, une

. ttesti 289 0.97 0.24 307 -1.56 2.34, une
```

Concluding Remarks

This dissertation tackled subjects central to comparative effectiveness research today. While the topics are different, each paper is connected to the other. There is a common emphasis on measurement, comparisons between multiple sources, and variation. Each paper is seeking to aggregate data from multiple sources to increase understanding of a relevant clinical or policy issue. The hospice care paper compares length-of-stay data between hospital referral regions, while the statistics paper uses different statistical methods to compare glaucoma treatment levels between papers. Similarly, the HPV paper aggregates cost-effectiveness data between different published papers. Overall, each one of the papers is looking at variation in the measurement of data of from multiple sources: the hospice care paper in terms of what are the determinants of hospice stay variation between hospital-referral regions, the statistics paper for the variation in estimates between different statistical methods, and the HPV paper in the amount of variation between female-only HPV vaccination compared to male and female HPV vaccination.

The hospice care paper points to the potential importance of reimbursement for decisions on end-of-life care. Hospice reimbursement is substantially lower than inpatient care reimbursement, but even a small increase in hospice reimbursement appears to have a significant effect on how long a patient stays in hospice. Further work that clarifies how hospice reimbursement is affected by family decision-making and the strength of the physician-patient relationship is important. The statistics paper provides compelling initial evidence that quantitative techniques might possibly be used to simulate randomized control trial results. Next steps are to evaluate the different techniques for additional data sets, and to evaluate the techniques for simulations with larger sample sizes. Finally, the HPV paper points to only two papers in the male HPV literature that show a significant effect of male and female HPV vaccination. In addition to the lack of significance of the other papers, male and female HPV vaccination is cost-effective at a $100,000/QALY threshold, while female-only HPV vaccination is cost-effective at a $50,000/QALY cutoff. Important future work includes re-testing the accuracy of these results, and using different ICER estimation techniques such as Bayesian analysis.

These three essays in comparative effectiveness research each make a unique and original contribution to CER. They contribute to the fundamental CER questions - *what works best, and under what circumstances?* As CER is in the national spotlight with the passage of healthcare reform, the way that healthcare is assessed and measured in the United States will increasingly draw from the field. With their application of quantitative methods to relevant policy topics, it is hoped that these papers will continue moving the field forward in its goal of improving healthcare delivery, optimizing health outcomes, and increasing healthcare access.