DATA FUSION: ITS ADVANTAGE IN PUBLIC HEALTH

by

GEORGE MAGDY KHALIL

(Under the Direction of Mathew Lee Smith)

ABSTRACT

Maximizing the utility of surveys while not adding questions is of utmost importance to surveillance systems. Public health agencies need to keep the ever-decreasing number of participants from breaking off after an interview is started. A common reason a participant breaks off is due to the length of the survey. It is therefore important that organizations conducting surveillance investigate innovative techniques of combining data from multiple, less extensive surveys. Data fusion is one such technique that has been used to integrate databases to save time and money. Health insurance status is a good topic to use for the validation of data fusion because this variable is common to many data sources and has a body of literature documenting factors associated with being insured. Besides data availability, respondents are thought to be accurate in reporting health insurance status and type (Call et al., 2008a). The goal of this research was to create "statistical twins" based on health insurance status from two data sources. Matched respondents were considered "statistical twins" and used to test whether data fusion is an effective method of predicting a variable not originally asked in the survey, given the respondent's profile. Data from the Behavioral Risk Factor Surveillance System's (BRFSS's) survey and the National Health Interview Survey (NHIS) were matched by first harmonizing the variables from the two data sources. A propensity score was calculated, which was then used to

perform Mahalanobis and Nearest Neighbor matching across the two surveys. The efficiency of the match was then validated: 88.2% of the 297,734 BRFSS respondents reported being covered by a health insurance, while 83.0% of the 27, 921 NHIS respondents reported currently being insured. Propensity scores were left-modal for both the NHIS and the BRFSS. Quantile-Quantile (QQ) plots, which plot the quantiles of one data set against another data revealed that after the match, the empirical distributions were similar in the BRFSS and NHIS groups. Compared to the original BRFSS dataset, the 2-to1 Nearest Neighbor (NN) algorithm was the closest to the BRRFSS respondents (86.2% [86.0, 86.50] versus 88.2% [88.1, 88.3], respectively). This is quite good considering national estimates differ by a few percentage points from survey to survey. Our imputed estimates are not within the confidence interval of the BRFSS. However, being within the narrow BRFSS confidence interval may be too rigorous a standard because of the very large sample size of the BRFSS. Sensitivities and specificities reveal that 2-to-1 NN with replacement and Mahalanobis were more accurate than Nearest Neighbor methods with caliper, without replacement and 1-to-1 matching.

INDEX WORDS: Data Fusion; Data Integration, Matching, BRFSS, NHIS

DATA FUSION: ITS ADVANTAGE IN PUBLIC HEALTH

by

GEORGE MAGDY KHALIL

BA, Rutgers University, 2004

MPH, Rutgers University, 2008

A Dissertation Prospectus Submitted to the Graduate Faculty of The University of Georgia

DOCTOR OF PUBLIC HEALTH

ATHENS, GEORGIA

2015

DATA FUSION: ITS ADVANTAGE IN PUBLIC HEALTH

by

GEORGE MAGDY KHALIL

Major Professor:	Mathew Lee Smith
Committee:	Mark Ebell Ye Shen Derek Ford

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
December 2015

DEDICATION

To my wife Karen and three children Rahel, Jude, and Eliza who made so many sacrifices for this to be possible. To my parents who left all they had and knew to come to this land of opportunity. I am indebted to you all.

This work is also dedicated to the memory of my grandmother, Samira Rofaiel, who cared for me and taught me lessons that cannot be written on a page. This is in fulfillment of her orders to become a doctor.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Overview

Maximizing the utility of surveys while not adding questions is of utmost importance to surveillance systems. There is a point at which the length of a survey is detrimental to break-off[1] rates and, more importantly, the quality of the data. Public health agencies need to keep the ever-decreasing number of participants from breaking off after an interview is started due to the length of the survey. It is therefore important that organizations conducting epidemiologic surveillance investigate innovative techniques of combining data from multiple, less extensive surveys.

Data fusion is one such technique that has been used to integrate databases to save time and money. Data fusion is defined as the practice of using common variables to combine separate respondent-level databases into one (Van Der Puttan, Nok, & Gupta, 2002). Data fusion has been used extensively in economic and market research and has been previously validated (Van Der Puttan et al., 2002).

[1] The American Association for Public Opinion research uses the following criteria: a. Less than 50% of all applicable questions answered (with other than refusal or no answer) equals break-off, 50%-80% equals partial, and more that 80% equals complete, or b. Less than 50% of all applicable questions asked equals break-off, 50-80% equals partial, and more than 80% equals complete, or c. Less than 50% of all essential or crucial questions answered (with other than a refusal or no answer) equals a break-off, 50-99% equals partial, and 100% equals complete, or d. a combination of a, b, or c (American Association for Public Opinion Research, 2008)

Examples of surveys that have used data fusion are Nielsen's Scarborough fused with the MARS Healthcare module (Scarborough, 2015), and Kantar's MRI database fused with Nielsen's NetRatings database (GfK Mediamark Research & Intelligence LLC, 2014). Realizing the importance of each other's data while not wanting to increase the respondent burden, these companies merged their data to collectively save time and money.

These two examples integrated data after the collection of unrelated surveys. An alternative option with data fusion is an integrated survey design. An example of an integrated survey design is the Dutch Household Survey on Living Conditions (van der Laan & van Nunspeet, 2009). This survey asks questions about demographics and socio-economic variables, a screening question on living conditions, and an exhaustive survey on living conditions. The last questionnaire is then split into two sub-questionnaires, which are completed by subgroups of respondents that are then statistically matched to the first two survey respondents. This is done to decrease the burden placed on the respondent. Examples of media enterprises that have used data fusion are copious (van der Laan & van Nunspeet, 2009). However, the literature about how public health can leverage data fusion has been very limited, especially in the United States.

Data: The BRFSS and NHIS Surveys. The Behavioral Risk Factor Surveillance System (BRFSS) is the world's largest telephone health survey. Administered by the Centers for Disease Control and Prevention (CDC), BRFSS collects information about the health risk behaviors and preventive practices in the 50 U.S. states, the District of Columbia, Puerto Rico, the U.S. Virgin Islands, Guam, American Samoa, and Palau (Centers for Disease Control and Prevention, 2013a). The National Health Interview Survey (NHIS) is a survey administered by the National

Center for Health Statistics (NCHS) and collected by the U.S. Census Bureau through a contractual agreement with the NCHS (National Center for Health Statistics, 2015). The NHIS collects information from the civilian non-institutionalized population to monitor the health of the United States and is one of the principle data collection programs of the NHCS. NHIS has two major parts: Core and Supplemental. The Core questions collect basic demographic data and a variety of health-related questions. The Core consists of four sections: Household, Family, Sample Adult, and Sample Child. The Supplemental part is used for emergent health issues as they arise and may contain follow-up questions to the Core questions (National Center for Health Statistics, 2015).

Variables Used. Health insurance status is an appropriate subject to use to validate of our matching procedure because this variable is common to both data sources. Besides data availability, respondents are fairly accurate in reporting health insurance status and type (Call et al., 2008b). Thus, using this variable will assist us in determining the accuracy of the data match.

A review of the literature was conducted in order to determine which variables are known to be associated with health insurance status. Demographic variables and correlates of health insurance coverage were used to match the most similar respondents. Previous research has shown that low income and unemployment are the most common reasons most non-elderly adults in the U.S. (i.e., ages 18-64 years) do not have health insurance. About 28.6% of uninsured individuals indicate a loss of a job as the main reason for not being insured (The Kaiser Commission on Medicaid and the Uninsured, 2014). Another 32.8% of non-elderly Americans indicated that could not afford the premiums. These include workers who are self-

employed, work part-time, or those who work at small firms who do not offer health insurance (The Kaiser Commission on Medicaid and the Uninsured, 2014).

In addition to employment and income, race is another important factor associated with being uninsured. People of color are account for half of those uninsured, but make up 40% of the population. People of Hispanic origin are also at higher risk of being uninsured, accounting for 30% of the uninsured, but only 19% of the population as a whole (The Kaiser Commission on Medicaid and the Uninsured, 2014).

Additionally, health insurance status varies by region and state. The Southern and Western regions have higher proportions of uninsured individuals (19.1% and 17.9% uninsured, respectively versus 12.4% in the Northeast and 13.0% in the Midwest) (Smith & Medalia, 2014). In these regions, states that had the lowest healthcare insurance coverage were Arkansas, Florida, Georgia, Mississippi, Nevada, and Texas (Centers for Disease Control and Prevention, 2015). Other demographic characteristics associated with the likelihood of insurance coverage are age, marital status, and residence (rural vs. urban). Compared to other adult age groups between 18-64 years, adults 26-34 years of age had the lowest rates of health insurance among adults with rates of insurance increasing with age beyond that age group (Smith & Medalia, 2014). In 2014, married adults 18-64 years of age had lower rates of uninsurance than those who were divorced or never married (89.7% vs. 82.9 and 81.5 respectively) (Smith & Medalia, 2014). In regards to residence, rural areas have higher rates of uninsured individuals with 22.3% of non-elderly rural residents who are uninsured compared to 21.4% in urban areas. Educational attainment has also been shown to have an affect on the proportion of individuals with health insurance. Individuals with graduate or professional degrees have a higher proportion of being insured than lower levels

of educational attainment (94.9% with graduate or professional degrees compared with 92.0% of individuals with an undergraduate degree, 82.3% among high school graduates, and 69.4% with less than high school education) (Smith & Medalia, 2014).

Modifiable health behaviors have also shown to be associated with health insurance coverage. In a study where researchers studied the characteristics of influenza vaccine uptake, health insurance coverage was the highest predictor of vaccine uptake (Takayama, Wetmore, & Mokdad, 2012). Smoking has also been linked to healthcare coverage. According to a study by the Centers for Disease Control and Prevention, uninsured individuals have a higher prevalence of smoking than insured individuals (Jamal et al., 2015).

Specific Aims

The goal of this research was to create matched respondent pairs from two data sources using propensity score matching (PSM) in order to validate estimates from the matched data. To do this, we statistically matched the BRFSS to the NHIS. These matched respondents are considered "statistical twins" and were used to test whether data fusion is an effective method of predicting a variable not originally asked in the survey, given the respondent's demographic profile.

Brief Overview of Proposed Methods

This research sought to match the BRFSS and the NHIS datasets using propensity scores and matching algorithms. Examples of such algorithms include nearest neighbor, Mahalanobis and caliper. Statistical Analysis System (SAS) version 9.4 (SAS Institute Inc., 2014) or the R statistical software (R Core Team, 2015) were used for the analysis. Nearest neighbor algorithms take the absolute difference between propensity scores between two data respondents

are minimized (Austin, 2009). The caliper method is a nearest neighbor algorithm that discards values outside the suggested range defined by the caliper (0.20 of a standard error of the estimated propensity score is the most commonly used) (Austin, 2009). Mahalanobis Distance is a popular algorithm where subjects are ordered at random then the distance between respondents is calculated using the vector of the mean values and covariance matrix of the independent variables (Wicklin, 2012).

The dependent variable we used to validate our methods was health insurance coverage (coded as yes, currently insurance or no, not currently insured). A SAS macro developed by Marcelo Coca-Perraillon (2007) and an R package by Ho et al. (2007) were used to match respondents from the two data sources based on a set of covariates. These covariates include demographic variables such as: age, sex, race and ethnicity, marital status, education, employment, and income; health behavior variables such as: smoking, vaccine uptake, and screening uptake; and disease variables such as cancer, cardiovascular disease (stroke and myocardial infarction), and respiratory disease (asthma and chronic bronchitis).

Types of Data Integration Techniques from Marketing Research

Because the most common consumers of data integration are marketing firms, one of the most influential research authorities guiding integration methods is the Advertising Research Foundation (ARF). In their report, the ARF (2003) offers best practices when integrating data from multiple databases. The ARF provides users with considerations when assessing the quality of data integration. They define data integration as:

> "…a formal process to combine information from two or more separate data sources, making use of information in the databases for the purpose of accurately estimating certain values that are not available in any single data source." (p. 4)

The ARF states that data integration techniques include unweighted and weighted demographic profile matching, geo-demographic and audience-typology clustering, calibration, multi-base integration, and fusion. These techniques differ in method, yet each aims to merge two databases by using a set of characteristics common to both (ARF, 2003). In addition to the integrity of each data source, the matching methods are crucial to the quality of the combined data (ARF, 2003). The following is the ARF's definition of each of the five main types of data integration techniques that fall into two general categories: Group- and Individual-level Matching.

Group-Level Matching

Demographic profile matching. Demographic profile matching identifies matches by using one database to identify variables such as age, sex, and a few other demographic variables. Next, one survey is used to identify the demographic group that has the greatest propensity to be associated with the outcome. Finally, the other survey is analyzed with respect to the demographic group. (ARF, 2003).

Weighted profile matching is a slightly more complex type of demographic profile matching where one is not focused on identifying a single segment, but rather on all groups of a single demographic variable. Table 1.1 is an example from ARF's guidance (ARF, 2003). This example pertains to a media audience dataset using a consumer behavior dataset to calculate usage.

The incidence or rate of a demographic group from one database is used to weight another. Products of these weights are then summed up to come up with a grand total as an estimate of the total number of individuals with the outcome. Sissors (1971) uses a Euclidean distance measure (Equation 1.1) to calibrate a group's profile from and match one database to another. When *d* is near zero, the demographic profile from the first database is more appropriately matched to the second.

Table 1.1: ARF Example of Weighted Demographic Profiling

Age Groups	Audience by Age Group		Incidence of Usage by Age Group		Estimated Users in Audience
25-34	541,000	X	35.6%	=	193,000
35-49	951,000	X	30.0%	=	285,000
50-64	520,000	X	26.9%	=	140,000
65+	318,000	X	21.6%	=	69,000
Total				=	823,000

This method is not without a major flaw, however. Canon, Smith, and Williams (2008) point out that a major weakness of this method is that the method is no more than the strength of an underlying relationship with an outcome. Therefore, these relationships are not distinct with an outcome (Cannon, Smith, & Williams, 2008).

Equation 1.1: Euclidean Distance

$$d = \sqrt{\sum_{i}^{n}\left[p\left(T|D_i\right) - p\left(M|D_i\right)\right]^2}$$

where d=Euclidean distance, Di=one of n demographic categories, T=database one, M=database two.

Audience typology and geodemographic clustering. The ARF separates these types of cluster methods; however, Cannon, Smith and Williams (2008) maintain that these methods are not in fact alternative methods. This method is used to group clusters of individuals based upon their behavior characteristics in the case of audience typology clustering and based upon geographic identifiers in the case of geodemographic clustering. In television media analyses, a cluster may be a group of individuals who are heavy daytime soap opera viewers. It may also be a ZIP Code cluster of similar demographic characteristics. A combination of clusters is called a "typology" (ARF, 2003). Typologies are then used to calculate probabilities and tabulated for any variable in either of the surveys.

Audience typology is a similar typology cluster, however, it is based on a similarity of the audience in terms of behavior. This typology must have the ability to be replicated on a second dataset. Therefore, there must be common survey questions to both data sources. This type deals with the assumption that individuals in the same neighborhood may not have similar lifestyles or behaviors. An example of this methodology is Spectra™ by Nielsen (Nielsen, 2008).

Calibration. Calibration is a method of modifying one survey to conform to another (ARF, 2003). This method is usually done to a range of demographic cells so that the one survey is identical to the other in terms of a variable. An example of this is how Scarborough calibrates their radio listening levels based on Arbitron™ radio currency data (Scarborough, 2015).

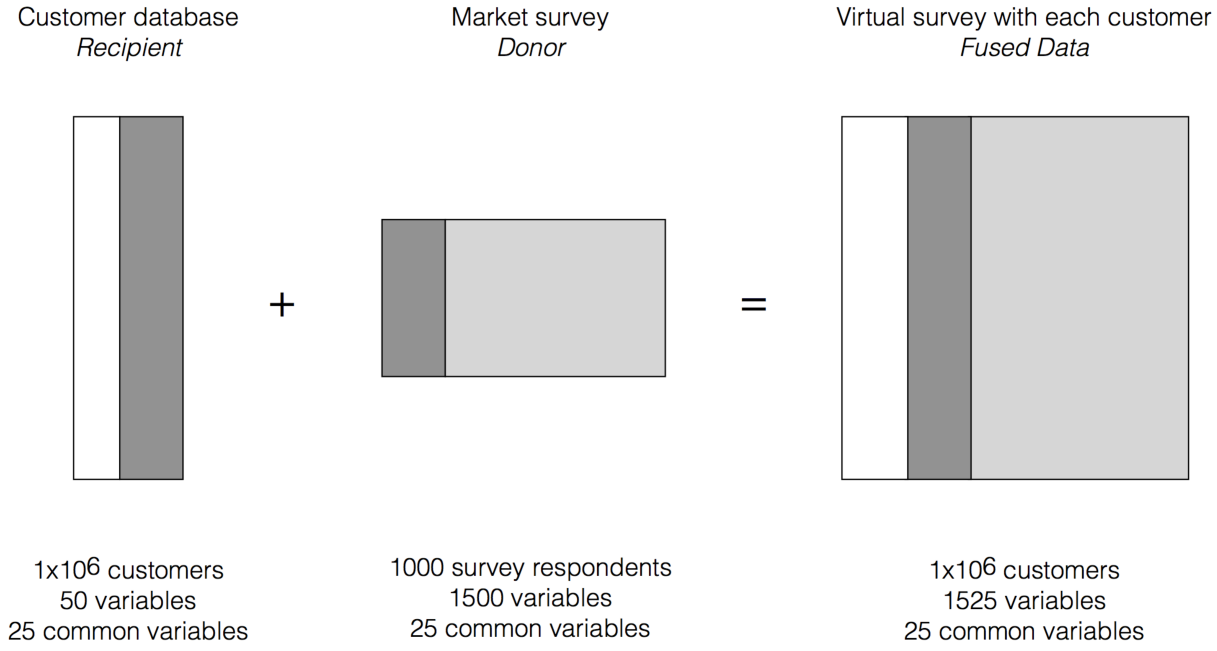
Scarborough modifies their data to conform to radio listening levels of their respondents based upon demographic and other variables of interest.

Multi-Base Integration. This method uses predictor variables to calculate estimates then seeks significant correlations between the two surveys. Thus, this method uses the common variables between the two surveys to build a predictive profile. This profile is then used to weight respondent cells similar to Weighted Profile Matching (ARF, 2003). The ARF points out that this method suffers from similar weaknesses as other techniques in that it heavily depends on the strength of the associations between the outcome and the variables you use for the match (p 14).

Individual-Level Matching.

Data Fusion Using Statistical Matching. In contrast to the previously discussed methods, data fusion integrates databases on the individual level. This method seeks to match a respondent from one survey to another based upon a group of variables (ARF, 2003; Van Der Puttan et al., 2002). The respondent pair is evaluated by a distance measure between a respondent from one survey to the other based upon pre-selected variables (ARF, 2003; Soong & de Montigny, 2001; Van Der Puttan et al., 2002). One survey, often called the 'recipient,' holds the bulk of the respondents along with a limited amount of variables. Another survey, often called the 'donor' has fewer respondents but contains variables not available in the recipient (Figure 1.1).

Figure 1.1: Data Fusion Example



Data fusion is not a new concept. In fact, it has been used for over for over 40 years in economic and marketing studies. Media researchers have used data fusion to integrate datasets since the 1980s (van Hattum & Hoijtink, 2008) and only recently in public health when the MARS health data were merged with Scarborough data (Scarborough, 2015). Statistical matching has also been used in medical literature as a method of matching in observational studies as a way to remove bias due to all observed covariates (Rosenbaum & Rubin, 1983, 1985).

One of the most popular uses of data fusion is integrating one media data source (e.g. television consumption) to another (e.g. purchasing behavior) with the aim of tailoring marketing strategies to specific groups of individuals with similar purchasing behaviors (Van Der Puttan et al., 2002). Van der Puttan et al. (2002) offer an example of Belgian National Readership, who has a media instrument and a separate product instrument each given to a group

of 10,000 respondents then fused together. They claim this reduced their cost and the required time for survey completion (Van Der Puttan et al., 2002). The accuracy of this study's match is

A paper by Kum and Masterson (2008), is another example of a data fusion application. The Levy Institute Measure of Economic Wellbeing (LIMEW) is an extended income measure calculated by merging multiple datasets via data fusion. This merged dataset includes the Current Population Survey's Annual Demographic Supplement for household for demographic and income data, the Survey of Consumer Finances for household wealth data, the American Time Use Survey for household production data, household income tax models, and administrative data for public consumption. Because no data source with such comprehensive wealth indicators exists, and because exact matching (on a specific person across the dataset) between the datasets is not possible because of confidentiality restrictions, data fusion is employed. Kum and Masterson evidence the accuracy of their fused match by calculating ratios of average and median net worth from the imputed data to the original data by each demographic variable. Ratios are all near 1 except for household income and family type where the ratios were about 1 to 0.8 some groups.

Constrained and Unconstrained Matching. Kum and Masterson (2008) give an overview of statistical matching used to fuse these datasets. They break down statistical matching into two categories: Constrained statistical matching (CSM) and unconstrained statistical matching (USM) (Kum & Masterson, 2008). Radner (1981) defines CSM as a match where every record from both datasets being matched are retained. Constrained matching also requires that the distance between the matched records be minimized and the weighted population totals between the datasets be equalized (Radner, 1981). This matching procedure is

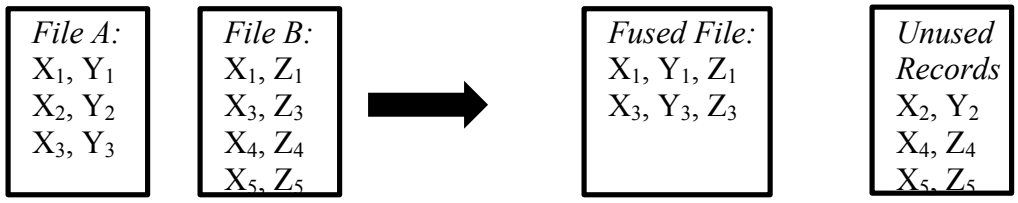
executed with replacement based upon rank, but not upon a distance measure such as Z values (Kum & Masterson, 2008). One main critique of CSM is that because of the rank matching employed, some matches are made based on greater distances than researcher would accept. Greater distances can lead to matched pairs that are too dissimilar and can affect the accuracy of the imputed data.

USM, on the other hand, uses a distance function to find the nearest neighbor. This is made based upon similarities in the donor and recipient records and is performed by imputing the nearest possible record among the closest records (Kum & Masterson, 2008). This means that this method may select a match from the donor or none at all. The authors argue that this method has a major flaw in that it can lead to differing empirical marginal or conditional distributions of the variables associated with the donor file (Y) given the variables common to both the recipient and donor files (Z) (Kum & Masterson, 2008).

Constrained Statistical Matching (CSM), on the other hand, requires that the weights and records be used. Kum and Masterson use this method to replicate the matched file from the donor file (2008, p. 5). CSM does this by minimizing the distance between the two matched records while equalizing the weighted population totals (Kum & Masterson, 2008). To do this, a weight-split procedure is employed during the matching (without replacement) so that the weights in the donor and recipient files are exhausted. CSM is sometimes called 'Imputation on Rank' since records are matched by rank and not an absolute distance measure (Kum & Masterson, 2008). Figure 1.2 below provides an example of how records are retained in CSM. When combined, files A and B have records that are unused after matching. These unused

records are used to match other records. In USM, the unused records are no longer retained (Moriarity & Scheuren, 2004).

Figure 1.2: Example of Constrained Statistical Match (CSM)*



*Adapted from (Moriarity & Scheuren, 2004)
X=matching or common variable; Y=variables in one dataset; Z=variables found in another.

Statistical Matching Algorithms

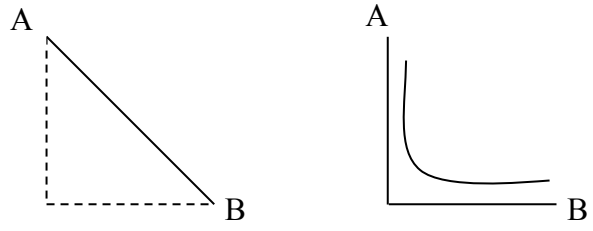
Overview of Matching Algorithms. There are many types of statistical matching algorithms used in data fusion including: regression techniques, discriminant analysis, nearest neighbor algorithms, network approaches, and more (Van Der Puttan et al., 2002). The algorithms we will focus on for this project are distance functions since they are the most commonly used algorithms for data fusion (Gilula, Mcculloch, & Rossi, 2006; Rogers, 1978; Van Der Puttan et al., 2002).

Distance Function Methods. These methods aim to calculate distance measures by using variables common to both the recipient and donor databases. Different techniques have been utilized to calculate these distance measures. Euclidean, City-block, and Mahalanobis (defined below) are some examples (van Hattum & Hoijtink, 2008). These distance measures are then used to match a donor to a recipient based on its nearest neighbor. These methods are

the most commonly used algorithms in data fusions (van Hattum & Hoijtink, 2008). One of the most cited criticisms of these algorithms is that the choice of distance score is subjective, which can affect the quality of the data fusion (van Hattum & Hoijtink, 2008). For this reason, the different algorithms need to be compared and validated.

Euclidean distance is the square root of the squared differences between corresponding elements of the rows or columns (D'Orazio, Di Zio, & Scanu, 2006). This is thought of as the hypotenuse, or the shortest distance between two points (see Figure 1.3). The City-block or Manhattan distance, on the other hand, is the calculated distance of x plus the distance of y. City-block distance is best visualized by a person walking around buildings in a city block (see figure 2). Some of the critics of Euclidean and City-block distance measures argue that they that they ignore scale and the correlation between variables (D'Orazio et al., 2006).

Figure 1.3: Comparison of Euclidean and City Block Distance



Mahalanobis distance is one the most popular distance measures used to sample match (D'Orazio et al., 2006). It is the distance between groups weighted by within-group dispersion (McCune & Grace, 2002). Mahalanobis distance is based on the mean and variance of predictor variables. One of the unique characteristics of this distance measure is that it incorporates the

covariance matrix of the variables to take into effect the covariance among the variables (see equation 1.2). This means that it transforms the data into standard uncorrelated data (Wicklin, 2012).

Equation 1.2: Mahalanobis Distance Equation

$D^2=(x-m)^T C^{-1}(x-m)$, where, D^2 denotes the Mahalanobis distance, x equals the vector of the data, m denotes the vector of the mean values of the independent variables, C^{-1} is the covariance matrix of the independent variables, and T indicates that the vector should be transposed. Mahalanobis Distance Scores (MDSs) are used to match based upon the smallest score between data sources.

As each subject from one data source is matched from the other, the pair is removed from the next set of matches. This continues until all units in a data source are exhausted. MDSs may also use a caliper where a common support region is defined. This method discards values outside the suggested range defined by the caliper (i.e., one-fourth of a standard error of the estimated propensity score) (Baser, 2006).

Distance functions can be performed with replacement, where subjects in the recipient data source are put back into the pool to be matched again. Stuart (2010) points out that matching with replacement may decrease bias because controls that look similar to numerous treated individuals can be used more than once. Stuart also points out that the order in which a subject is chosen does not matter when the subject is replaced into the pool (Stuart, 2010). However, with large sample sizes, these biases are minimized (D'Orazio et al., 2006; Van Der Puttan et al., 2002).

Outline

Manuscripts. In Chapters 2 and 3, the two required publishable journal articles with their four sections (Introduction, Methods, Results, Discussion) are presented. The first manuscript is a proof of concept paper for our selected data fusion methods. The second manuscript is a sensitivity analysis of our methods to evaluate the accuracy of our matched data.

Discussion: In the Discussion section, I discuss our key findings from both manuscripts and their implications to public health. Potential limitations and weaknesses of this study will also be discussed.

CHAPTER 2

DATA FUSION: ITS ADVANTAGE IN PUBLIC HEALTH [1]

[1] Khalil, G.M., Ford, D., Shen, Y, Ebell, M., Smith, M.L. To be submitted to *Journal of Public Health*.

Abstract

Aim: Maximizing the utility of surveys while not adding questions is of utmost importance to epidemiologic surveillance systems. Public health agencies need to keep the ever-decreasing percent of responding participants from breaking off after an interview is started due to the length of the survey. It is therefore important that organizations conducting surveillance activities investigate innovative techniques of combining data from multiple, less extensive surveys. This study aimed to match respondents from two surveys based on health insurance status and test whether data fusion is an accurate way to predict the value of a variable not originally asked in the survey, given the respondent's profile. **Methods**: Data from the Behavioral Risk Factor Surveillance System's (BRFSS's) survey and the National Health Interview Survey (NHIS) were matched by first harmonizing the variables from the two data sources. Next, we conducted a descriptive analysis on the possible matching variables and their association with health insurance. A propensity score was calculated for each respondent, which was then be used to perform Mahalanobis and Nearest Neighbor respondent matching across the two surveys. **Results:** About 88% of the 297,734 BRFSS respondents reported being covered by a health insurance, while 83.0% of the 27, 921 NHIS respondents reported currently being insurance. Propensity scores were left-modal for both the NHIS and the BRFSS. Quantile-Quantile (Q-Q) plots revealed that after the match, the empirical distributions were similar in the BRFSS and NHIS groups. Compared to the original BRFSS dataset, the 2-to1 Nearest Neighbor (NN) algorithm was the closest to the BRRFSS respondents (86.2% [86.0, 86.5] 2-to-1 NN vs. 88.2% [88.1, 88.3] for BRFSS). Replacement and a caliper of 0.20 had no effect on the algorithm's ability predict health insurance status. **Discussion:** Using propensity scores, the

algorithms' ability to accurately match one respondent from the BRFSS to an NHIS respondent to predict health insurance status was within two or three percentage points. This is quite good considering national estimates differ by a few percentage points from survey to survey. Sensitivities and specificities reveal that 2-to-1 NN with replacement and Mahalanobis were more accurate than Nearest Neighbor methods with caliper, without replacement and 1-to-1 matching.

Keywords: Data Fusion; Data Integration, Matching, BRFSS, NHIS

Introduction

Public Health Surveillance systems are continually looking for ways to maximize the use of their survey data without adding questions to the instrument. Although research has been inconclusive regarding the association between survey length and completing a survey (Rolstad, Adler, & Rydén, 2011), public health agencies put on emphasis on reducing respondent burden. The fear is that increasing the length of a survey will decrease survey completion rates and be detrimental to the quality of the data collected. Many agencies are investigating methods of gleaning information from the public without increasing respondent time burden. Combining data from multiple, less extensive survey is one of the ways they are looking to glean this information.

Data fusion is one of the techniques researchers are using to combine data from multiple data sources in hopes of saving time and money. Data fusion uses common variables to combine multiple databases on the respondent level (Van Der Puttan et al., 2002). Data fusion has been validated and used extensively in economic and market research (Van Der Puttan et al., 2002).

Examples of surveys that have used data fusion are copious. Examples of surveys that have used data fusion are Nielsen's Scarborough fused to the MARS Healthcare module (Scarborough, 2015) and Kantar's MRI database fused with Nielsen's NetRatings database (GfK Mediamark Research & Intelligence LLC, 2014). Not wanting to increase the time burden of a respondent on any one survey, they worked together to integrate their companies' data to collectively save time and money.

These unrelated marketing surveys fused their data after the collection. Another option is to use an integrated survey where the design of the survey (pre-collection) is based upon

integrating the data through data fusion. The Dutch Household Survey on Living Conditions (DHSLC) is one such example (van der Laan & van Nunspeet, 2009). The DHSLC has three survey instruments: one with demographics and socio-economic variables, another with screening questions on living conditions, and an comprehensive survey instrument on living conditions. The last more comprehensive instrument is split into two sub-questionnaires, completed by two respondent subgroups. The data from the two subgroups are then statistically matched to the data from the first two questionnaires. There are many examples of economic and media research agencies that have used data fusion (van der Laan & van Nunspeet, 2009). The literature on the use of data fusion in public health is very limited, especially in the United States.

Data. The BRFSS and NHIS Surveys. The Behavioral Risk Factor Surveillance System (BRFSS) is the world's largest telephone health survey. This survey is administered by the Centers for Disease Control and Prevention (CDC) and collects information about the health risk behaviors and preventive practices in the 50 U.S. states, the District of Columbia, Puerto Rico, the U.S. Virgin Islands, Guam, American Samoa, and Palau (Centers for Disease Control and Prevention, 2013a). The National Health Interview Survey (NHIS) is a survey administered by the National Center for Health Statistics (NCHS) and collected by the U.S. Census Bureau through a contractual agreement with the NCHS (National Center for Health Statistics, 2015). This survey collects information from the non-institutionalized civilian population in order to monitor the health of the population of the United States. NHIS is one of the principle data collection programs of the NHCS. NHIS's two major parts are categorized as Core and Supplemental. The questions in the Core section collect basic demographic data and a variety of

health-related questions. This section consists of four sections: Household, Family, Sample Adult, and Sample Child. The Supplemental section assesses emergent health issues as they arise as well as follow-up questions to the Core questions (National Center for Health Statistics, 2015).

Variables Used. The variable used to validate our matching procedure was health insurance status. It is an appropriate subject to use to because this variable is common to both data sources. Besides being available in both surveys, respondents are fairly accurate in reporting health insurance status and type (Call et al., 2008b). Therefore, the use of this variable will best determine the accuracy of the data match.

Because we used health insurance status to validate our matching procedure, we considered possible correlates of health insurance to use in matching respondents from the BRFSS and NHIS surveys. Previous research has shown that among non-elderly adults (i.e., ages 18-64 years) in the United States, the two most common reasons for not being insured is low income and unemployment. About a third (28.6%) of uninsured individuals state that the reason for not being insured is a loss of a job (The Kaiser Commission on Medicaid and the Uninsured, 2014). Another third (32.8%) of non-elderly Americans stated that they could not afford the premiums because they are self-employed, work part-time, or work at small firms who do not offer health insurance (The Kaiser Commission on Medicaid and the Uninsured, 2014).

Another factor associated with being uninsured is race. In fact, about half of the uninsured are people of color, yet they account for 40% of the population in the United States. People of Hispanic origin are also at a higher risk of being uninsured. They account for a third

(30%) of the uninsured, yet they account for 19% of the population (The Kaiser Commission on Medicaid and the Uninsured, 2014).

Health insurance status also varies by regionally. Regions in the Southern and West have higher proportions of uninsured individuals (19.1% and 17.9% uninsured, respectively versus 12.4% in the Northeast and 13.0% in the Midwest) (Smith & Medalia, 2014). The states that had the lowest rates of healthcare insurance coverage were Arkansas, Florida, Georgia, Mississippi, Nevada, and Texas (Centers for Disease Control and Prevention, 2015). Age, marital status, and residence (rural vs. urban) are also associated with the likelihood of insurance coverage. Adults 26-34 years of age had the lowest rates of health insurance among adults with rates of insurance increasing with age beyond that age group (Smith & Medalia, 2014). In 2014, married adults 18-64 years of age had higher proportion of health insurance coverage than those who were divorced or never married (89.7% vs. 82.9 and 81.5 respectively) (Smith & Medalia, 2014). In regards to residence, non-elderly rural residents have higher rates of uninsurance than their urban counterparts (22.3% compared to 21.4% in urban areas) (Smith & Medalia, 2014). The proportion of individuals with health insurance also have higher educational attainment (94.9% with graduate or professional degrees compared with 92.0% of individuals with an undergraduate degree, 82.3% among high school graduates, and 69.4% with less than high school education) (Smith & Medalia, 2014).

Modifiable health behaviors, such as disease prevention or high risk behaviors, have also shown to be associated with health insurance coverage. One example is increased uptake of influenza vaccinations among those with health insurance coverage. In a study by Takayama et al., health insurance coverage was the best predictor of influenza vaccine uptake (2012).

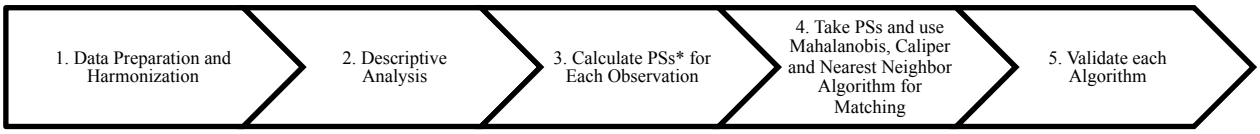
Smoking is another modifiable health behavior that is linked to health insurance coverage. Researchers at the Centers for Disease Control and Prevention revealed that smokers had a higher likelihood of being uninsured than non-smokers (Jamal et al., 2015).

Methods

An application for this study was submitted to the University of Georgia Institutional Review Board (IRB). The IRB reviewed our application (Protocol ID# STUDY00001602) and determined that the activities proposed is research not involving human subjects.

Figure 2.1 is an overview of the proposed methods for the research. First, we prepared and harmonized the data. Next, we conducted a descriptive analysis of the variables used for matching. A propensity score was then calculated and used to match respondents using Nearest Neighbor, Caliper, and Nearest Neighbor algorithms. Finally, each algorithm was validated by comparing the percentage with health insurance in the original data with the fused data.

Figure 2.1. Overview of Proposed Methods



*PS=propensity score

Data Preparation and Harmonization. Using the Household Serial Number (HHX), the Family Serial Number (FMX) and the Person Sequence Number (FPX), the adult and person NHIS files were merged. The NHIS files include documentation, including information about data collection, how variables were edited, survey sampling, as well as how to crosswalk the variables to the survey instrument. The files also include codebooks and programming statement files and American Standard Code for Information Interchange (ASCII) data files.

The Behavioral Risk Factor Surveillance System (BRFSS) 2012 data were downloaded from CDC's BRFSS Survey Data and Documentation public use website (Centers for Disease Control and Prevention, 2015b). The BRFSS data did not need to be merged to a parent dataset; therefore, it was ready upon download. A new variable was created to define the origin of the data (NHIS=0; BRFSS=1).

According to D'Orazio et al., (2006) the first step in a matching process is checking the consistency of the two surveys and harmonizing the variables if necessary (step 1 in Figure 2.1) (D'Orazio, Di Zio, & Scanu, 2006). This was done in two phases: 1) the harmonization of population and unit definitions; and 2) definition of variables. The variables in the NHIS were recoded to align with the BRFSS. Variables with similar responses were joined into a new recoded variable. Some categories such as race were collapsed into new categories when necessary. Continuous variables such as age were recoded into ordinal variables. Variable recoding to crosswalk the BRFSS and the NHIS were based upon empirical and logical decisions, not on initial analyses. Table 2.1 contains a few examples of variables that needed to be recoded. Since the BRFSS is an adult only survey, only respondents 18 or older were retained from the NHIS. Because we are examining health insurance, respondents 65 years and older

were excluded since the majority of individuals 65 years and older have Medicare (Smith & Medalia, 2014). The NHIS public use does not include the state the respondent is from, but rather the region, BRFSS respondents' state of residency were recoded into the region the state is in.

Descriptive Analyses. Health insurance status was used as a dependent variable for this data fusion exercise. A descriptive analysis of the variables within each dataset was executed. Demographics variables discovered to be correlated with health insurance status in the review of the literature and common to both the NHIS and the BRFSS datasets were examined in a descriptive analysis. Demographic variables such as age, sex, race, ethnicity, education, marital status, employment status, education, and region were investigated. Health behavior variables (e.g. smoking, vaccine and screening uptake) and disease variables (e.g., self-reported cancer, cardiovascular disease, or respiratory disease) were also explored as potential covariates.

Stepwise regression analyses were performed for each of the two data sources to choose variables independently explaining health insurance status. Significant variables were considered significant if the p-value is 0.20 or less since the 0.05 cutoff is thought to be too strict for stepwise regressions (Bendel & Afifi, 1977). Significant variables in both of the data sources were included in the propensity score calculation in the next section.

Propensity Score Estimation. Rubin, one of the first to use propensity score for matching, defines the propensity score as the "conditional probability of assignment to a particular treatment given a vector of observed covariates" (Rosenbaum & Rubin, 1983). Regression coefficients obtained through logistic regression are used to estimate the propensity scores. We calculated a propensity score for each observation in both datasets. The propensity

score, $e(x_i)$, is calculated by utilizing Equation 2.1; where α is the intercept, b is the covariate's regression coefficient, and x is the corresponding value of the variable.

Equation 2.1: Propensity Score Calculation.

$$e(\mathrm{x_i})=\frac{1}{1+e^{-(\hat{a}+S\,\hat{b}_k x_{ik})}}$$

In our case, the propensity score is the probability of having health insurance given the covariates. Propensity scores range between 0 and 1 since it is a probability. Variables from the data exploration stage described in the previous section were used in this equation as covariates. SAS 9.4 was used to calculate this regression. Propensity scores of the observations in each of the two datasets were examined. Balance statistics are important in matching since gives an indication of the ability of the matching procedure to closely match respondents. Balance statistics such as means, group standard deviations (where applicable), mean differences, and Quantile-Quantile (Q-Q) plot differences, which plots the quantiles of one data set against another data, were examined.

Data Fusion. Propensity scores were used in data integration algorithms. The units were matched using Mahalanobis, Nearest Neighbor and Caliper algorithms. We used Statistical Analysis Software (SAS) version 9.4 and R to execute these algorithms. A SAS macro written by Marcelo Coca-Perraillon was used to group observations that are similar together across the two data sources (Coca-Perraillon, 2007). The following algorithms were executed using SAS macros or the Matchit R package: 1-to-1 Nearest Neighbor; 2-to-1 Nearest Neighbor;

Mahalanobis; and Caliper. These SAS macros and R package automate the process of the matching algorithms. The output provided a BRFSS respondent identifier (control) and an NHIS respondent identifier (treatment) to match to based upon distance scores.

Validation. To examine the algorithms' ability to predict health insurance accurately, the healthcare insurance variable from the BRFSS was blinded. Next, we used the observation number from the matching output from the previous section to merge the selected NHIS respondents with their corresponding health insurance status to the selected BRFSS respondents. The original health insurance variable from the BRFSS was then merged and compared to the fused data by looking at the fused data percent with health insurance and how they compared to the original data.

Results

The results of the descriptive analysis of the potential matching variables by data sources are in Table 2.2. The variable frequencies are reported by data source with corresponding percentages in parentheses. Every variable observed significantly differed by data source with $p<.0001$. About 88.2% of the 297,734 BRFSS respondents reported being covered by a health insurance, while 83.0% of the 27, 921 NHIS respondents reported currently being insurance. The results of the descriptive analysis of the potential matching variables by health insurance status are in Table 2.3. The variable frequencies are reported by health insurance status with corresponding percentages in parentheses. Most notable from the results of the univariate analysis is the low percentage of Hispanics with health insurance with only 68.7% of Hispanic respondents from the BRFSS and 65.9% of Hispanic respondents from the NHIS reporting being insured.

Significant variables from the descriptive analysis were entered into a stepwise regression with $p<.20$ as the cutoff with health insurance status as the dependent variable. Variables found to meet this significance cutoff are found in Figure 2.2. Significant variables for the BRFSS included: region, sex, age group, race, Hispanic ethnicity, marital status, educational attainment, employment, reporting having received a flu vaccine, pneumonia vaccine, smoking status, having a checkup with a physician, general health, reporting chronic heart disease, COPD and cancer diagnoses. Significant variables for the NHIS included: region, age group, Hispanic ethnicity, marital status, education, employment, reporting having received a flu vaccine, general health, smoking status, having a checkup with a physician, and reporting a diabetes diagnosis, and reporting a cancer diagnosis. The results from the stepwise regression are presented Table 2.5 and Table 2.6. Hispanic ethnicity was again one of the most noteworthy variables, with Non-Hispanics more than twice more likely than Hispanics to have health insurance among the respondent from both surveys. Also noteworthy, respondents who had an influenza vaccine were two and a half times more likely to have health insurance.

The next step of the data matching process was to calculate propensity scores for each respondent using significant matching variables that were common to both data sources. The common variables found in Table 2.4 were used in a logistic regression to calculate propensity scores for each respondent. Figure 2.2 is a plot of the distribution of the propensity scores by source. The plot by source, the propensity scores were left-modal (positively skewed) for both the NHIS and the BRFSS. Appendix C has Q-Q plots of the matching variables used. Plots reveal that after the match, the empirical distributions are similar in the BRFSS and NHIS since they are not far from the reference line.

Using propensity scores, the algorithms' ability to match one respondent from the BRFSS to an NHIS respondent to predict health insurance status is found in Table 2.5. It shows that compared to the original BRFSS dataset, the 2-to-1 Nearest Neighbor (NN) algorithm without replacement was the closest to the BRRFSS respondents (86.3% [85.99, 86.50] vs.88.2% [88.11, 88.30], respectively). The caliper method is a nearest neighbor algorithm that discards values outside the suggested range defined by the caliper (0.20 of a standard error of the estimated propensity score is the most commonly used) (Austin, 2009). A caliper of 0.20 did not have an effect on the ability of the algorithm to predict health insurance status. NN with a 0.20 caliper were equal to NN without a caliper (1-to-1 NN with and without caliper=85.9% [85.49, 86.22]).

Table 2.6 lists the percent of individuals that were correctly classified, sensitivities and specificities for each matching algorithm. Although 2-to-1 Nearest Neighbor without replacement was the closest to the original BRSS in terms of percent with health insurance, the Mahalanobis algorithm had the highest percent of correctly classified individuals, sensitivity, and specificity (78.03%, 84.9%, and 38.5% respectively). The 2-to-1 Nearest Neighbor with replacement and 1-to-1 Nearest Neighbor with replacement algorithms had equivalent percent of correctly classified individuals, specificity, and specificity (76.0%, 83.5%, and 28.6% respectively).

Table 2.1: Sample NHIS and BRFSS Variable Recoding

BRFSS Variable, Categories	**NHIS Variable, Categories**	**Proposed New Recoded Variable Name and Variable Coding**
AGE (continuous)	AGE_P (continuous)	NAGE (ordinal), 1=18-24 2=25-44 3=45-64 4=65+

_PRACE 1=White 2=Black 3=Asian 4=Native Hawaiian/ Pacific Islander 5=American Indian/ Alaska Native 6=Other 7=Multi-Racial, no primary race 8=Multi-Racial, preferred race not asked. 77=Don't know/not sure 99=Refused	MRACBPI2, 1=White 2=Black 3=American Indian/Alaska Native 6=Chinese 7=Philipino 8=Asian Indian 16=Other Race 17=Multi-Racial, no primary race 97=Refused 99=Don't know/not sure	_PRACE2, 1=White 2=Black 3=Asian (includes NHIS values 6,7,8) 4=Native Hawaiian/Pacific Islander 5=American Indian/Alaska Native (includes NHIS category 16) 6=Other (includes NHIS category 17) 7=Multi-Racial (includes BRFSS categories 7,8 and NHIS category 17)
HISPANC2 1=Yes 2=No 7=Don't know/not sure 9=Refused	ORIGIN_I 1=Yes 2=No	HISPANC2 1=Yes 2=No
MARITAL 1=Married 2=Divorced 3=Widowed 4=Separated 5=Never married 6=Member of unmarried couple 9=Refused	R_MARITAL 1=Married, Spouse in Household Component 2=Married, Spouse not in Household Component 3=Married, Spouse in Household Component unknown 4=Widowed 5=Divorced 6=Separated 7=Never married 8=Living with partner 9=Unknown marital status	MARITAL 1=Married 2=Divorced 3=Widowed 4=Separated 5=Never married 6=Member of unmarried couple 9=Unknown/Refused

Table 2.2: Descriptive analysis comparing BRFSS and NHIS

Variable	BRFSS n (%)	NHIS n (%)
Health Insurance		
Currently Insured	262,611 (88.2%)	23,037 (82.5%)
P value		**P<.0001**
Demographic Variables		
Race		
White	241,524 (82.5%)	19,683 (71.1%)
Black	27,534 (9.4%)	4,110 (14.8%)
American Indian/Alaska Native	7,116 (2.4%)	348 (1.3%)
Asian	7,860 (2.7%)	731 (2.6%)
Other	7,412 (2.5%)	2,673 (9.7%)
Multi-Racial	1,146 (0.40%)	140 (0.51%)
P value		**<.0001**
Hispanic Ethnicity		
Yes	25,063 (8.5%)	5,273 (18.8%)
No	271,205 (91.2%)	22,780 (81.2%)
P value		**<.0001**
Age		
18-24	23,438 (7.8%)	3,353 (11.9%)
25-34	42,414 (14.2%)	6,431 (22.9%)
35-44	53,216 (17.8%)	5,947 (21.2%)
45-54	76,042 (25.4%)	6,117 (21.8%)
55-64	104,047 (34.8%)	6,205 (22.2%)
P value		**<.0001**
Sex		
Male	130,237 (43.5%)	12,916 (46.0%)
Female	168,920 (56.5%)	15,137 (54.0%)
P value		**<.0001**
Marital Status		
Married	167,728 (56.6%)	12,423 (44.4%)
Divorced	40,731 (13.7%)	3,825 (13.7%)
Widowed	10,514 (3.5%)	639 (2.3%)
Separated	7,482 (2.5%)	910 (3.2%)
Never Married	59,559 (20.1%)	8,088 (28.9%)
Member of an unmarried couple	10,204 (3.4%)	2,106 (17.1%)
P value		**<.0001**
Education		
Never	392 (0.13%)	100 (0.4%)
Elementary (1^{st}—8^{th})	5,744 (1.9%)	1,135 (4.1%)
Some High School (9^{th}—11^{th})	14,596 (4.9%)	2,492 (8.9%)
High School Graduate	7,889 (26.6%)	6,936 (24.8%)
Some College	83,310 (28.1%)	9,016 (32.3%)

College (4+ years)	113,745 (38.3%)	8,268 (29.6%)
P value		**<.0001**
Employment		
Employed for wages at job or business	200,330 (67.8%)	19,591 (75.7%)
Out of work > 1 yr.	9,048 (3.1%)	884 (3.4%)
Out of work < 1 yr.	8,881 (3.0%)	1,280 (4.9%)
Homemaker	18,496 (6.3%)	1,956 (7.6%)
Student	10,547 (3.6%)	911 (3.5%)
Retired	21,969 (7.4%)	942 (3.6%)
Unable to work	26,367 (8.9%)	307 (1.2%)
P value		**<.0001**
Region		
Northeast	54,648 (18.3%)	4298 (15.4%)
South	86,746 (29.1%)	9920 (35.5%)
Midwest	82,812 (27.8%)	5895 (21.1%)
West	73,528 (24.7%)	7808 (28.0%)
P value		**<.0001**
Behavior Variables		
Smoking Status		
Never Smoked	167,384 (58.7%)	17,295 (61.9%)
Former Smoker	65,345 (22.9%)	5,022 (18.0%)
Smokes on some days	15,272 (5.4%)	1,377 (4.9%)
Smokes everyday	37,171 (13.0%)	4,225 (15.1%)
P value		**<.0001**
Routine Checkup with healthcare provider		
Within 1 year	200,640 (67.9%)	22,184 (80.3%)
Within 2 years	41,592 (14.1%)	2,402 (8.7%)
Within 5 years	25,642 (8.7%)	1,597 (5.8%)
5+ years	24,250 (8.2%)	949 (3.4%)
Never	3,149 (1.1%)	491 (1.8%)
P value		**<.0001**
Flu Immunization (yes)	111,047 (39.4%)	9,773 (35.5%)
P value		**<.0001**
BMI Categories		
Obese (30.0+)	85421 (31.0%)	9255 (33.1%)
Overweight (25.0—29.9)	96,235 (34.9%)	9020 (32.3%)
Normal Weight (18.5—24.9)	89796 (32.6%)	9166 (32.8%)
Underweight (<18.5)	4259 (1.5%)	480 (1.7%)
P value		<.001
General Health		
Excellent	59,873 (20.2%)	8,273 (29.64%)
Very good	102,673 (34.6%)	9,011 (32.29%)
Good	87,296 (29.41%)	7,166 (25.68%)

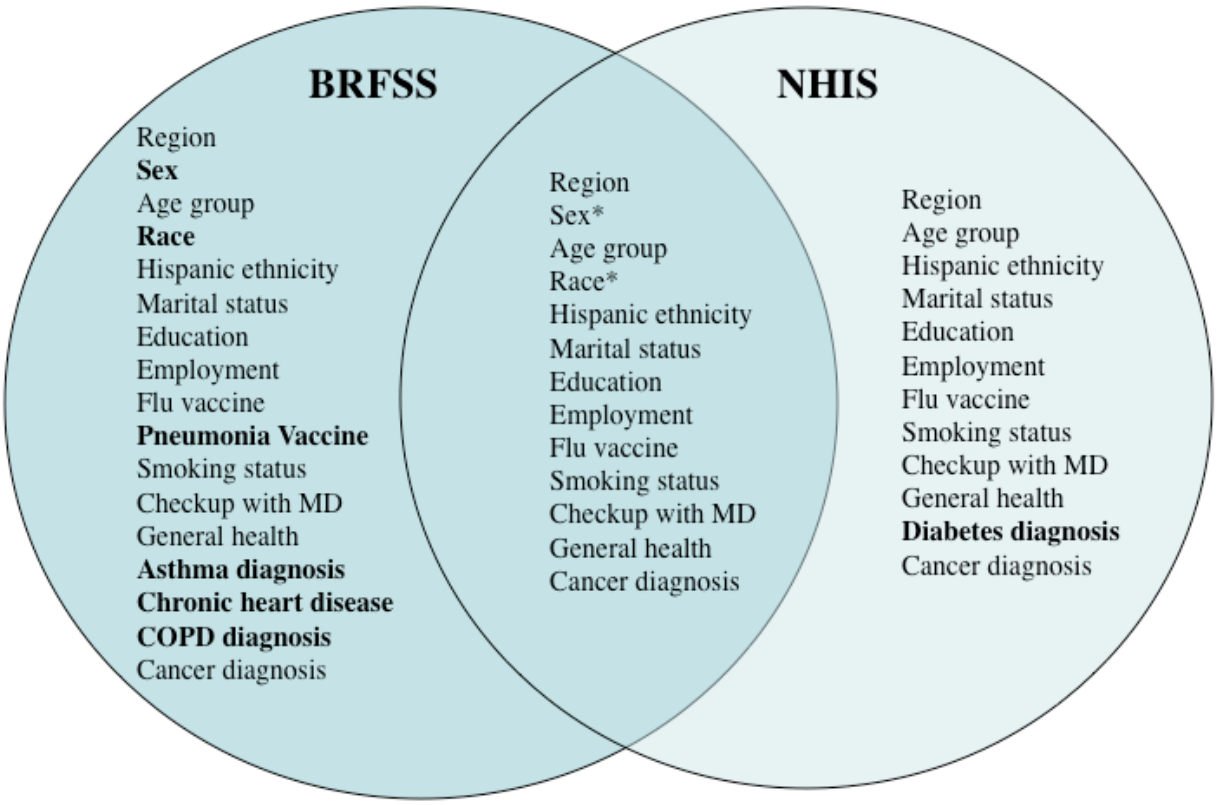
Fair	33,534 (11.3%)	2,662 (9.54%)
Poor	1,349 (14.54%)	797 (2.86%)
P value		**<.0001**
Disease Variables		
Asthma		
Now has asthma	28,740 (9.7%)	2241 (8.0%)
No longer has asthma	1,187 (4.0%)	1513 (5.4%)
Never had asthma	255,017 (86.3%)	24119 (86.5%)
P value		**<.0001**
Cardiovascular disease		
Myocardial infarction	9,564 (3.2%)	546 (2.0%)
Cerebrovascular accident	7,089 (2.4%)	518 (1.9%)
Chronic heart disease	9,376 (3.2%)	166 (0.59%)
P value		**<.0001**
COPD	18,417 (6.2%)	2,011 (7.2%)
P value		**<.0001**
Cancer (Not Skin)		
P value	16350 (5.5%)	1495 (5.4%)
Diabetes		.3003
Borderline	27350 (9.2%)	1998 (7.2%)
P value	4021 (1.4%)	366 (1.3%)
		<.0001

Table 2.3: Descriptive analysis of select variables comparing health insurance coverage between data sources

Variable	Currently Insured BRFSS	Currently Insured NHIS
Demographic Variables		
Race		
White	215,344 (89.5%)	16,508 (84.2%)
Black	23,020 (84.0%)	3,374 (82.8%)
American Indian/Alaska Native	6,067 (85.9%)	211 (60.6%)
Asian	7,013 (89.7%)	630 (87.0%)
Other	5,409 (73.5%)	1,895 (71.4%)
Multi-Racial	1,009 (88.66%)	89 (63.6%)
Hispanic Ethnicity		
Yes	17,112 (68.7%)	3,457 (65.9%)
No	243,087 (90.0%)	19,580 (86.4%)
Age		
18-24	18,755 (81.8%)	2,701 (81.6%)
25-34	34,462 (81.5%)	4,931 (77.0%)
35-44	46,091 (86.8%)	4,765 (80.4%)
45-54	67,736 (89.4%)	5,120 (84.0%)
55-64	95567 (92.2%)	5,520 (89.2%)
Sex		
Male	112,631 (87.0%)	10,346 (80.5%)
Female	149,980 (89.1%)	12691 (84.2%)
Marital Status		
Married	154,484 (92.3%)	10,717 (86.4%)
Divorced	34,402 (84.7%)	3,179 (83.4%)
Widowed	9,214 (88.0%)	533 (83.9%)
Separated	5,737 (76.9%)	659 (72.7%)
Never Married	48,587 (82.5%)	6,404 (79.9%)
Member of an unmarried couple	7,710 (75.9%)	1,500 (71.7%)
Education		
Never	199 (51.2%)	60 (60.6%)
Elementary (1^{st}—8^{th})	3,356 (58.7%)	615 (54.4%)
Some High School (9^{th}—11^{th})	10,194 (70.6%)	1,712 (69.1%)
High School Graduate	65,233 (83.3%)	5,258 (76.3%)
Some College	73,555 (88.7%)	7,582 (84.4%)
College (4+ years)	108,000 (95.1%)	7,735 (93.9%)
Employment		
Employed for wages	178,890 (89.6%)	16,360 (83.9%)
Out of work > 1 yr.	6,281 (67.0%)	567 (64.4%)
Out of work < 1 yr.	5,791 (65.8%)	763 (60.0%)
Homemaker	15,188 (82.4%)	1,447 (74.1%)

Student	8,932 (86.6%)	791 (87.8%)
Retired	20,719 (94.7%)	850 (90.9%)
Unable to work	23,972 (91.3%)	254 (83.3%)
Behavior Variables		
Smoking Status		
Never Smoked	150,124 (90.1%)	14,413 (83.8%)
Former Smoker	59,013 (90.6%)	4,330 (86.5%)
Smokes on some days	12,328 (81.1%)	1,051 (76.7%)
Smokes everyday	29,331 (79.2%)	3,128 (74.4%)
Routine Checkup with M.D.		
Within 1 year	185,952 (93.0%)	19,446 (88.0%)
Within 2 years	35,507 (85.8%)	1,702 (71.4%)
Within 5 years	19,721 (77.3%)	875 (55.3%)
5+ years	16,428 (68.2%)	396 (41.9%)
Never	1,972 (62.9%)	272 (56.2%)
Flu Immunization	104,652 (94.5%)	8,980 (92.2%)
BMI Categories		
Obese (30.0+)	75,493 (88.4%)	7,653 (82.7%)
Overweight (25.0—29.9)	85,585 (88.9%)	7,360 (81.6%)
Normal Weight (18.5—24.9)	79,221 (88.2%)	7,628 (83.2%)
Underweight (<18.5)	3,568 (83.8%)	398 (82.9%)
General Health		
Excellent	54,261 (90.6%)	7,008 (84.7%)
Very good	93,865 (91.4%)	7,558 (83.9%)
Good	74,757 (85.6%)	5,638 (78.7%)
Fair	27,551 (82.2%)	2,122 (79.7%)
Poor	11,487 (85.1%)	702 (88.1%)
Chronic Disease Indicators		
Asthma		
Now has asthma	25,777 (89.7%)	1,943 (86.7%)
No longer has asthma	10,430 (87.8)	1,285 (84.9%)
Never had asthma	224,656 (88.1%)	19,769 (82.0%)
Cardiovascular disease		
Myocardial infarction	8,457 (88.4%)	477 (87.4%)
Cerebrovascular accident	6,301 (88.9%)	463 (89.4%)
Chronic heart disease	8,492 (90.6%)	149 (89.8%)
COPD	16,158 (87.7%)	1,757 (87.4%)
Cancer (Not Skin)	15,022 (91.9%)	1,350 (90.3%)
Diabetes	24,648 (90.1%)	1,745 (87.3%)

Figure 2.2: Comparison of significant variables from stepwise regression by data source



*Variable was added despite failing to be significant for NHIS

Table 2.4: Results of significant variables from a stepwise regression for BRFSS respondents

Variable	Odds Ratio (CI) BRFSS
Region	0.92 [0.91, 0.93]
Sex	0.89 [0.87, 0.91]
Age group	1.10 [1.09, 1.11]
Race	0.97 [0.96, 0.98]
Hispanic	2.23 [2.16, 2.30]
General Health	0.86 [0.86, 0.87]
Marital Status	0.86 [0.86, 0.87]
Education	1.56 [1.55, 1.58]
Employment	1.03 [1.03, 1.04]
Flu Vaccine	0.44 [0.43, 0.45]
Pneumonia Vaccine	0.89 [0.87, 0.91]
Smoking status	1.20 [1.19, 1.21]
Checkup past year	0.68 [0.68, 0.69]
Asthma Diagnosis	0.94 [0.92, 0.95]
Chronic Heart Disease Diagnosis	0.91 [0.86, 0.97]
COPD Diagnosis	0.93 [0.89, 0.97]
Cancer Diagnosis	0.87 [0.83, 0.91]

Table 2.5: Results of significant variables from a stepwise regression for NHIS respondents.

Variable	Odds Ratio (CI) NHIS
Region	0.91 [0.89, 0.93]
Age group	1.06 [1.04, 1.08]
Hispanic	1.95 [1.84, 2.07]
General health	0.85 [0.83, 0.87]
Diabetes	0.78 [0.70, 0.88]
Marital status	0.90 [0.89, 0.92]
Education	1.58 [1.54, 1.61]
Employment	0.95 [0.94, 0.96]
Flu Vaccine	0.40 [0.38, 0.43]
Smoking status	1.19 [1.17, 1.22]
Checkup past year	0.71 [0.70, 0.72]
Cancer Diagnosis	0.76 [0.66, 0.88]

Figure 2.3: Distribution of Propensity Scores by Survey

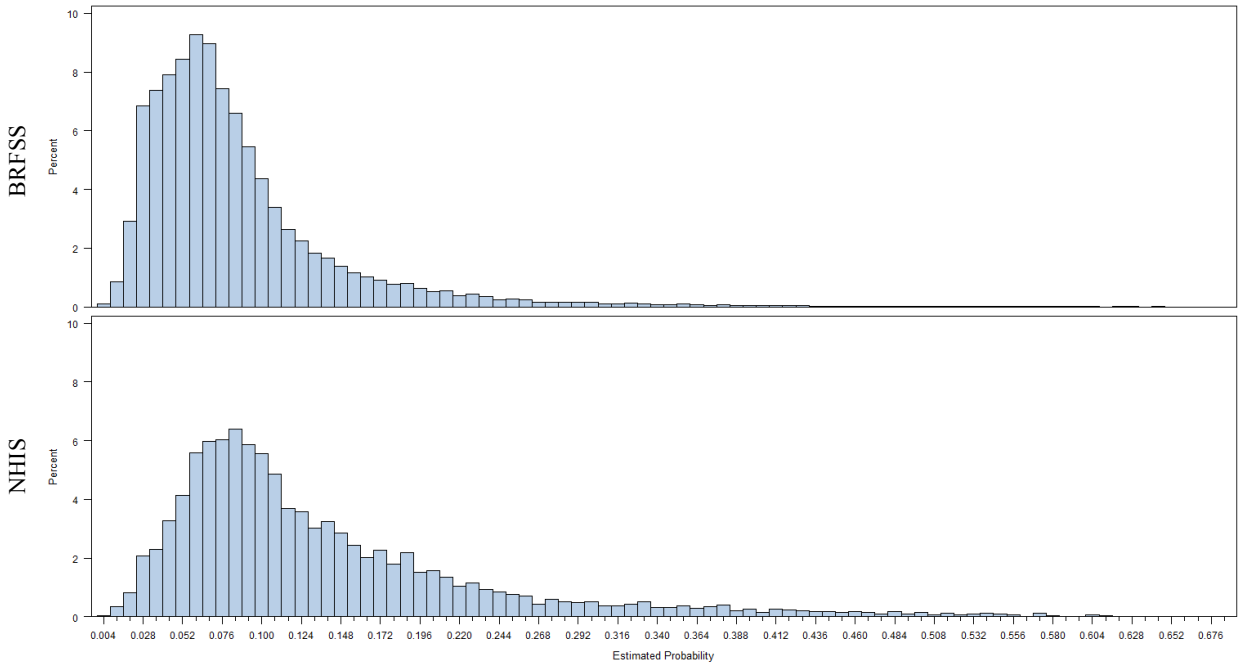


Table 2.6: Percent with Health Insurance by Matching Algorithm compared to Unmatched BRFSS respondents

Algorithm	**Percent with Health Insurance**	**Confidence Interval**
Unmatched BRFSS respondents	88.2%	[88.1, 88.3]
1-to-1 Nearest Neighbor No Replacement	85.9%	[85.5, 86.2]
1-to-1 Nearest Neighbor with Replacement	85.8%	[85.5, 86.2]
Nearest Neighbor with 0.20 Caliper	85.8%	[85.5, 86.2]
2-to-1 Nearest Neighbor No Replacement	86.2%	[86.0, 86.5]
2-to-1 Nearest Neighbor with Replacement	85.8%	[85.6, 86.1]
Mahalanobis	85.9%	[85.5, 86.3]

Table 2.7: Percent of Individuals that were Correctly Classified, Sensitivity and Specificity by Matching Algorithm

Algorithm	Percent Correctly Classified	Sensitivity	Specificity
1-to-1 NN without Replacement	75.9%	83.4%	28.3%
1-to-1 NN with Replacement	76.0%	83.5%	28.6%
2-to-1 NN without Replacement	75.6%	83.0%	27.0%
2-to-1 NN with Replacement	76.0%	83.5%	28.6%
1-to-1 0.20 Caliper without replacement	75.9%	83.4%	28.3%
Mahalanobis	78.0%	84.9%	38.5%

Discussion

Our research aimed to test the ability of data fusion methods to predict health insurance. Percentages of respondents with health insurance in our fused data were all similar to the original BRFSS data. We found that Stuart's assertion that replacement decreases bias because controls that look similar to numerous treated individuals can be used more than and that the order in which a subject is chosen does not matter when the subject is replaced into the pool, was not the case (Stuart, 2010). We believe this is because of the large sample size of our study.

A caliper of 0.20 had no effect due to the fact that the sample size was so large—the algorithm had a large pool of respondents from which to pull from for an accurate match. To investigate the effect of the sample size on caliper matching, we randomly chose 50,000 and 100,000 respondents from the BRFSS dataset. As the sample size decreased, more respondents were held back from the matched dataset since the caliper was not met for respondents' to be

matched. This is because this method discards values outside the suggested range defined by the caliper (0.20 of a standard error of the estimated propensity score).

Our study was subject to a number of limitations. One of the major limitations was the fact that the respondents were not selected via simple random sampling methods. Both the BRFSS and the NHIS oversample certain races and regions to have adequate sample for weighting certain races or rural areas (Centers for Disease Control and Prevention, 2013b; National Center for Health Statistics, 2015). Another limitation is that the two surveys used in our research had different survey administration modes. The BRFSS is a telephone only survey and the NHIS is an in-person survey. Future research needs to be conducted on the effect on weighting and survey modes on data fusion techniques for public health estimates.

The major finding of this research is that by fusing data from two data sources, we were able to calculate health insurance estimates within 2-3% of the original data. This is promising considering the fact that national estimates differ this much even when taking survey administration into consideration. In 2014, the national estimate of heath insurance for individuals less than 65 years of age was 86.7% for the National Health Interview Survey and 88.0% from Census Current Population Survey data (both in-person surveys collected by the Census Bureau). (CDC National Center for Health Statistics, 2014; Smith & Medalia, 2014). Especially when wanting to calculate estimates not otherwise found in any database, data fusion is a viable option.

CHAPTER 3

A SENSITIVITY ANALYSIS OF THE VARIABLES NEEDED FOR FUSING THE BEHAVIORAL RISK FACTOR SURVEILLANCE SYSTEM AND THE NATIONAL HEALTH INTERVIEW SURVEY FOR THE IMPUTATION OF HEALTH INSURANCE STATUS [2]

[2] Khalil, G.M., Ford, D., Shen, Y, Ebell, M., and Smith, M.L., To be submitted to *Journal of Information Fusion*.

Abstract:

Aim: Maximizing the utility of surveys while not adding questions is of utmost importance to surveillance systems. Public health agencies need to keep the ever-decreasing number of participants from breaking off after an interview is started due to the length of the survey. It is therefore important that organizations conducting surveillance activities investigate innovative techniques of combining data from multiple, less extensive surveys. This study aims to conduct a sensitivity analysis of the matching variables used in algorithms that match respondents from two surveys based on health insurance status. **Methods**: Data from the Behavioral Risk Factor Surveillance System's (BRFSS's) survey and the National Health Interview Survey (NHIS) were matched by first harmonizing the variables from the two data sources. Next, we conducted a descriptive analysis on the possible matching variables and their association with health insurance. A propensity score was calculated for each respondent, which was then be used to perform Mahalanobis, Caliper, and Nearest Neighbor respondent matching across the two surveys. A sensitivity analysis was then conducted on the variables used to integrate the two surveys. Chi-square tests performed on matching variables post-match to assess whether the propensity score model was appropriately specified for each algorithm **Results:** Health insurance status significantly differed by all variables in the analysis ($p<.0001$). The significant variables that were common to both were: region, age group, Hispanic ethnicity, marital status, education, employment, reporting having received a flu vaccine, general health, smoking status, having a checkup with a physician, and reporting ever being diagnosed with cancer. The 2-to-1 Nearest Neighbor algorithm with replacement matched more variables efficiently than did without replacement. Caliper did not make a difference in terms of matching than did 1-to-1-

Nearest Neighbor without replacement. The Mahalanobis algorithm was not able to efficiently match some of the variables that nearest neighbor algorithms were able to match. **Discussion:** The algorithms' ability to match by each variable varied. Chi-square tests conducted to examine whether the fused data differed by each matching variable, revealed that Nearest Neighbor without replacement was the most efficient at matching the two data sources.

Keywords: Data Fusion; Data Integration, Matching, BRFSS, NHIS

Introduction

Maximizing the utility of surveys while not adding questions is of utmost importance to surveillance systems. There is a point at which the length of a survey is detrimental to break-off[3] rates and, more importantly, the quality of the data. Public health agencies need to keep the ever-decreasing number of participants from breaking off after an interview is started due to the length of the survey. It is therefore important that organizations conducting epidemiologic surveillance investigate innovative techniques of combining data from multiple, less extensive surveys.

Data fusion is one such technique that has been used to integrate databases to save time and money. Data fusion is defined as the practice of using common variables to combine separate respondent-level databases into one (Van Der Puttan et al., 2002). Data fusion has been used extensively in economic and market research and has been previously validated (Van Der Puttan et al., 2002).

Examples of surveys that have used data fusion are Nielsen's Scarborough fused with the MARS Healthcare module (Scarborough, 2015), and Kantar's MRI database fused with Nielsen's NetRatings database (GfK Mediamark Research & Intelligence LLC, 2014). Realizing the importance of each other's data while not wanting to increase the respondent burden, these companies merged their data to collectively save time and money.

---

[3] The American Association for Public Opinion research uses the following criteria: a. Less than 50% of all applicable questions answered (with other than refusal or no answer) equals break-off, 50%-80% equals partial, and more that 80% equals complete, or b. Less than 50% of all applicable questions asked equals break-off, 50-80% equals partial, and more than 80% equals complete, or c. Less than 50% of all essential or crucial questions answered (with other than a refusal or no answer) equals a break-off, 50-99% equals partial, and 100% equals complete, or d. a combination of a, b, or c (American Association for Public Opinion Research, 2008)

These two examples integrated data after the collection of unrelated surveys. An alternative option with data fusion is an integrated survey design. An example of an integrated survey design is the Dutch Household Survey on Living Conditions (van der Laan & van Nunspeet, 2009). This survey instrument includes demographics and socio-economic variables, screening questions on living conditions, and an exhaustive survey instrument on living conditions. The last questionnaire is then split into two sub-questionnaires, which are completed by subgroups of respondents that are then statistically matched to the first two survey respondents. This is done to decrease the burden placed on the respondent. Examples of media enterprises that have used data fusion are copious (van der Laan & van Nunspeet, 2009). However, the literature about how public health can leverage data fusion has been very limited, especially in the United States.

Data: The BRFSS and NHIS Surveys. The Behavioral Risk Factor Surveillance System (BRFSS) is the world's largest telephone health survey. Administered by the Centers for Disease Control and Prevention (CDC), BRFSS collects information about the health risk behaviors and preventive practices in the 50 U.S. states, the District of Columbia, Puerto Rico, the U.S. Virgin Islands, Guam, American Samoa, and Palau (Centers for Disease Control and Prevention, 2013a). The National Health Interview Survey (NHIS) is a survey administered by the National Center for Health Statistics (NCHS) and collected by the U.S. Census Bureau through a contractual agreement with the NCHS (National Center for Health Statistics, 2015). The NHIS collects information from the civilian non-institutionalized population to monitor the health of the United States and is one of the principle data collection programs of the NHCS. NHIS has two major parts: Core and Supplemental. The Core questions collect basic demographic data

and a variety of health-related questions. The Core consists of four sections: Household, Family, Sample Adult, and Sample Child. The Supplemental part is used for emergent health issues as they arise and may contain follow-up questions to the Core questions (National Center for Health Statistics, 2015).

Variables Used. Health insurance status is an appropriate subject to use to validate of our matching procedure because this variable is common to both data sources. Besides data availability, respondents are fairly accurate in reporting health insurance status and type (Call et al., 2008b). Thus, using this variable will assist us in determining the accuracy of the data match.

A literature review was conducted to determine variables that have been known to be associated with health insurance to assist us in choosing covariates. Demographic variables and correlates of health insurance coverage were used to match the most similar respondents. Previous research has shown that low income and unemployment are the most common reasons most non-elderly adults in the U.S. (i.e., ages 18-64 years) do not have health insurance. About 28.6% of uninsured individuals indicate a loss of a job as the main reason for not being insured (The Kaiser Commission on Medicaid and the Uninsured, 2014). Another 32.8% of non-elderly Americans indicated that could not afford the premiums. These include workers who are self-employed, work part-time, or those who work at small firms who do not offer health insurance (The Kaiser Commission on Medicaid and the Uninsured, 2014).

In addition to employment and income, race is another important factor associated with being uninsured. People of color are account for half of those uninsured, but make up 40% of the population. People of Hispanic origin are also at higher risk of being uninsured, accounting

for 30% of the uninsured, but only 19% of the population as a whole (The Kaiser Commission on Medicaid and the Uninsured, 2014).

Additionally, health insurance status varies by region and state. The Southern and Western regions have higher proportions of uninsured individuals (19.1% and 17.9% uninsured, respectively versus 12.4% in the Northeast and 13.0% in the Midwest) (Smith & Medalia, 2014). In these regions, states that had the lowest healthcare insurance coverage were Arkansas, Florida, Georgia, Mississippi, Nevada, and Texas (Centers for Disease Control and Prevention, 2015). Other demographic characteristics associated with the likelihood of insurance coverage are age, marital status, and residence (rural vs. urban). Compared to other adult age groups between 18-64 years, adults 26-34 years of age had the lowest rates of health insurance among adults with rates of insurance increasing with age beyond that age group (Smith & Medalia, 2014). In 2014, married adults 18-64 years of age had lower rates of uninsurance than those who were divorced or never married (89.7% vs. 82.9 and 81.5 respectively) (Smith & Medalia, 2014). In regards to residence, rural areas have higher rates of uninsured individuals with 22.3% of non-elderly rural residents who are uninsured compared to 21.4% in urban areas. Educational attainment has also been shown to have an affect on the proportion of individuals with health insurance. Individuals with graduate or professional degrees have a higher proportion of being insured than lower levels of educational attainment (94.9% with graduate or professional degrees compared with 92.0% of individuals with an undergraduate degree, 82.3% among high school graduates, and 69.4% with less than high school education) (Smith & Medalia, 2014).

Modifiable health behaviors have also shown to be associated with health insurance coverage. In a study where researchers studied the characteristics of influenza vaccine uptake,

health insurance coverage was the highest predictor of vaccine uptake (Takayama et al., 2012). Smoking has also been linked to healthcare coverage. According to a study by the Centers for Disease Control and Prevention, uninsured individuals have a higher prevalence of smoking than insured individuals (Jamal et al., 2015).

Methods

An application for this study was submitted to the University of Georgia Institutional Review Board (IRB). The IRB reviewed our application (Protocol ID# STUDY00001602) and determined that the activities proposed is research not involving human subjects. A copy of this documentation is located in Appendix B.

Data Preparation. Using the Household Serial Number (HHX), the Family Serial Number (FMX) and the Person Sequence Number (FPX), the adult and person NHIS files were merged. The NHIS files include documentation, including information on data collection, how variables were edited, survey sampling, as well as how to crosswalk the variables to the survey instrument. The files also include codebooks and programming statement files and American Standard Code for Information Interchange (ASCII) data files.

The Behavioral Risk Factor Surveillance System (BRFSS) 2012 data were downloaded from CDC's BRFSS Survey Data and Documentation public use website (Centers for Disease Control and Prevention, 2015b). The BRFSS data did not need to be merged to a parent dataset, therefore it was ready upon download. A new variable was created to define the origin of the data (NHIS=0; BRFSS=1).

Since the BRFSS is an adult only survey, only respondents 18 or older were retained from the NHIS. Because we used health insurance as our outcome, respondents 65 years and

older were excluded because the majority of individuals 65 years and older have Medicare (Smith & Medalia, 2014). The NHIS public use does not include the state the respondent is from, but rather the region; therefore BRFSS respondents' state of residency was recoded into the region the state is in.

Descriptive Analysis. Health insurance status was used as a dependent variable for this research. A descriptive analysis of the variables within each dataset was executed. Demographics variables discovered to be correlated with health insurance status in the review of the literature and common to both the NHIS and the BRFSS datasets were examined in a descriptive analysis. Demographic variables such as age, sex, race, ethnicity, education, marital status, employment status, education, and region were investigated. Health behavior variables include smoking, vaccine and screening uptake and disease variables include self-reported disease such as cancer, cardiovascular disease such as stroke and history of chronic disease, and respiratory disease such as asthma and chronic bronchitis were also explored as potential covariates.

Propensity Score Calculations. Rubin, one of the first to use propensity score for matching, defines the propensity score as the "conditional probability of assignment to a particular treatment given a vector of observed covariates" (Rosenbaum & Rubin, 1983). In our case, the propensity score is the probability of having health insurance given the covariates. Propensity scores were calculated and compared across the two data sources using variables associated with health insurance status. The propensity score is calculated by utilizing Equation 3.1. Variables from the data exploration stage described in the previous section were used in this equation. SAS 9.4 was used to calculate this regression. Propensity scores of the observations in

each of the two datasets were examined. Balance statistics such as means, group standard deviations (where applicable), mean differences, and Quantile-Quantile (Q-Q) plot differences (Appendix C) were examined.

Equation 3.1: Propensity Score Calculation

$$e(\mathrm{x_i})=\frac{1}{1+e^{-(\hat{a}+S\,\hat{b}_k x_{ik})}}$$

Matching. Propensity scores were used in data integration algorithms. The units were matched using Mahalanobis, Nearest Neighbor and Caliper algorithms. We used SAS version 9.4 and R to execute these algorithms. A SAS macro written by Marcelo Coca-Perraillon was used to group observations that are similar together across the two data sources (Coca-Perraillon, 2007). The following algorithms were executed using SAS macros or the Matchit R package: 1-to-1 Nearest Neighbor, 2-to-1 Nearest Neighbor, Mahalanobis, and Caliper. These SAS macros and R package automate the process of the matching algorithms. The output provided a BRFSS respondent identifier (control) and an NHIS respondent identifier (treatment) to match to based upon distance scores.

Sensitivity Analysis of Matching Variables. After the calculation of the propensity scores and matching the data by source, a sensitivity analysis was conducted to investigate which variables were needed to match data appropriately. While there is no standard to assess balance statistics, we chose to calculate a chi-square test recommended by Austin (Austin, 2009). This gives a good indication of how good the match was by each matching variable. Matching

variables were recoded into categorical variables (e.g. White Race—Yes/No). A Chi-Square test on the matching variables as categorical variables was carried out with the null being: NHIS and BRFSS respondents in the Fused Data significantly differed for variables included in the calculation of the propensity score at P<.05.

Results

The results of the descriptive analysis of the potential matching variables by health insurance status are in Table 3.1. The variable frequencies are reported by health insurance status with corresponding percentages in parentheses. Health insurance status significantly differed by all analyzed variables (p<.0001). Most notable from the results of the univariate analysis is the low percentage of Hispanics with health insurance; only 68.7% of Hispanic respondents from the BRFSS and 65.9% of Hispanic respondents from the NHIS reported being insured.

Significant variables from the descriptive analysis were entered into a stepwise regression with P<.20 as the cutoff with health insurance status as the dependent variable. Variables found to meet this significance cutoff are found in Table 3.2. The significant variables that were common to both were: region, age group, Hispanic ethnicity, marital status, education, employment, reporting having received a flu vaccine, general health, smoking status, having a checkup with a physician, and reporting ever being diagnosed with cancer.

The results from the stepwise regression are presented Table 3.3 for BRFSS and Table 3.4 for the NHIS. Hispanic ethnicity was again one of the most noteworthy variables, with Non-Hispanics about twice more likely than Hispanics to have health insurance among the respondent

in the NHIS and the BRFSS. Also noteworthy, respondents who had an influenza vaccine were two and a half times more likely to have health insurance in both the NHIS and the BRFSS.

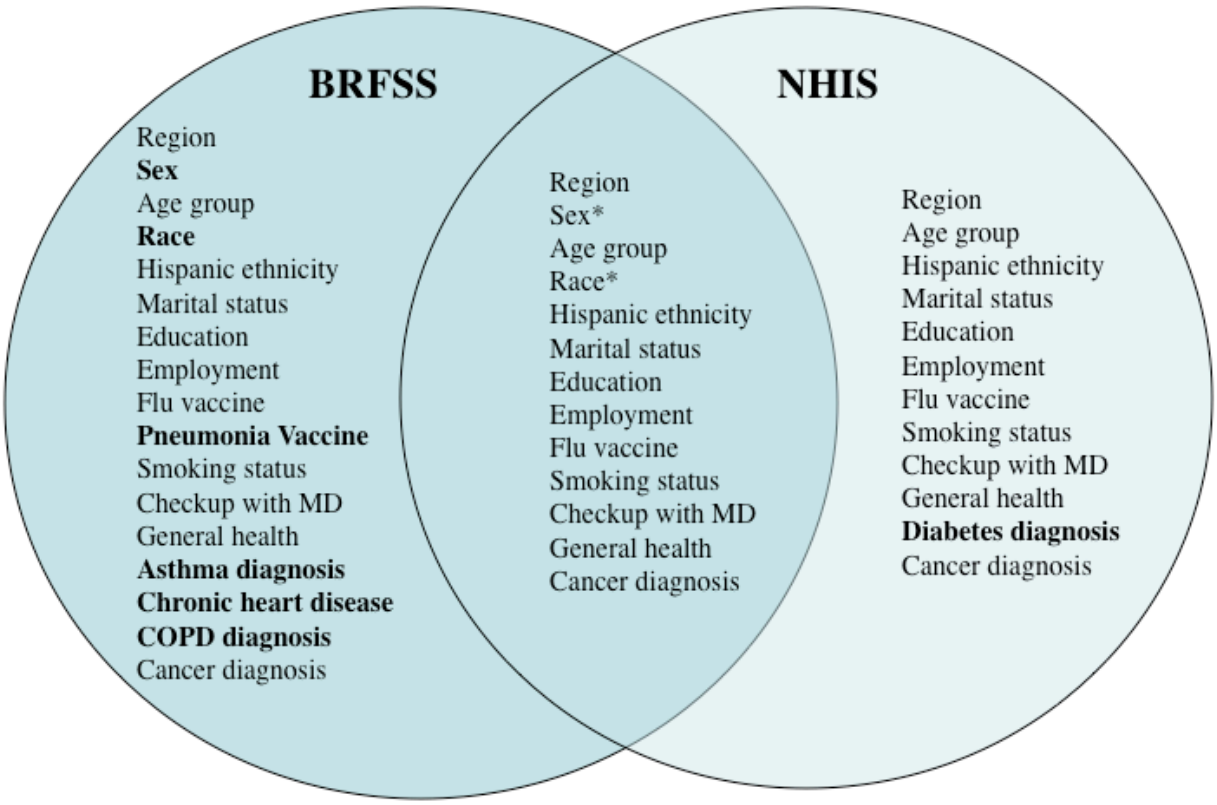
Chi-square tests performed on matching variables post-match to assess whether the propensity score model was appropriately specified for each algorithm are in Table 3.5. Variables that were significant in the logistic regression that calculated the propensity score are highlighted in gray. The algorithms' ability to match by each variable varied. The Mahalanobis algorithm was able to efficiently match Asian and Multi-Racial races, never and elementary school education, those out of work for longer than a year and students, those that never had a checkup, those diagnosed with Cancer and those with borderline diabetes. Caliper and 1-to-1 Nearest Neighbor without replacement algorithms did not differ at all in terms their ability to match each matching variable. These algorithms were able to accurately match Hispanic ethnicity, those in the oldest two age groups (45-54 and 55.64 years old), sex, married, separated and divorced individuals, retired individuals, the Northeast and Western regions, individuals that never smoked, those who had a routine checkup within that past 5 years, normal weight and over weight BMIs, all categories of general health besides those reporting poor health, those reporting a cardiovascular accident, or stroke, those who reported they ever had cancer, and borderline diabetics. The 2-to-1 Nearest Neighbor algorithms with and without replacement were similar in their ability to matching on matching variables. The 2-to-1 Nearest Neighbor algorithm with replacement matched more variables efficiently than did without replacement. None of the algorithms efficiently matched asthma.

Table 3.1: Descriptive analysis of select variables comparing health insurance coverage between data sources

Variable	Currently Insured BRFSS	Currently Insured NHIS
Demographic Variables		
Race		
White	215,344 (89.5%)	16,508 (84.2%)
Black	23,020 (84.0%)	3,374 (82.8%)
American Indian/Alaska Native	6,067 (85.9%)	211 (60.6%)
Asian	7,013 (89.7%)	630 (87.0%)
Other	5,409 (73.5%)	1,895 (71.4%)
Multi-Racial	1,009 (88.7%)	89 (63.6%)
Hispanic Ethnicity		
Yes	17,112 (68.7%)	3,457 (65.9%)
No	243,087 (90.0%)	19,580 (86.4%)
Age		
18-24	18,755 (81.8%)	2,701 (81.6%)
25-34	34,462 (81.5%)	4,931 (77.0%)
35-44	46,091 (86.8%)	4,765 (80.4%)
45-54	67,736 (89.4%)	5,120 (84.1%)
55-64	95567 (92.2%)	5,520 (89.2%)
Sex		
Male	112,631 (87.0%)	10,346 (80.5%)
Female	149,980 (89.1%)	12691 (84.2%)
Marital Status		
Married	154,484 (92.3%)	10,717 (86.4%)
Divorced	34,402 (84.7%)	3,179 (83.4%)
Widowed	9,214 (88.0%)	533 (83.9%)
Separated	5,737 (76.9%)	659 (72.7%)
Never Married	48,587 (82.5%)	6,404 (79.9%)
Member of an unmarried couple	7,710 (75.9%)	1,500 (71.7%)
Education		
Never	199 (51.2%)	60 (60.6%)
Elementary (1^{st}—8^{th})	3,356 (58.7%)	615 (54.4%)
Some High School (9^{th}—11^{th})	10,194 (70.6%)	1,712 (69.1%)
High School Graduate	65,233 (83.3%)	5,258 (76.3%)
Some College	73,555 (88.7%)	7,582 (84.4%)
College (4+ years)	108,000 (95.1%)	7,735 (93.9%)
Employment		
Employed for wages	178,890 (89.6%)	16,360 (83.9%)
Out of work > 1 yr.	6,281 (60.0%)	567 (64.4%)
Out of work < 1 yr.	5,791 (65.8%)	763 (60.0%)
Homemaker	15,188 (82.4%)	1,447 (74.1%)

Student	8,932 (86.6%)	791 (87.8%)
Retired	20,719 (94.7%)	850 (90.9%)
Unable to work	23,972 (91.3%)	254 (83.3%)
Behavior Variables		
Smoking Status		
Never Smoked	150,124 (90.1%)	14,413 (83.8%)
Former Smoker	59,013 (90.6%)	4,330 (86.5%)
Smokes on some days	12,328 (81.1%)	1,051 (76.7%)
Smokes everyday	29,331 (79.2%)	3,128 (74.4%)
Routine Checkup with M.D.		
Within 1 year	185,952 (93.0%)	19,446 (88.0%)
Within 2 years	35,507 (85.8%)	1,702 (71.4%)
Within 5 years	19,721 (77.3%)	875 (55.3%)
5+ years	16,428 (68.2%)	396 (42.0%)
Never	1,972 (62.9%)	272 (56.2%)
Flu Immunization	104,652 (94.5%)	8,980 (92.2%)
BMI Categories		
Obese (30.0+)	75,493 (88.4%)	7,653 (82.7%)
Overweight (25.0—29.9)	85,585 (88.9%)	7,360 (81.6%)
Normal Weight (18.5—24.9)	79,221 (88.2%)	7,628 (83.2%)
Underweight (<18.5)	3,568 (83.8%)	398 (82.9%)
General Health		
Excellent	54,261 (90.6%)	7,008 (84.7%)
Very good	93,865 (91.4%)	7,558 (83.9%)
Good	74,757 (85.6%)	5,638 (78.7%)
Fair	27,551 (82.2%)	2,122 (79.7%)
Poor	11,487 (85.1%)	702 (88.1%)
Chronic Disease Indicators		
Asthma		
Now has asthma	25,777 (89.7%)	1,943 (86.7%)
No longer has asthma	10,430 (87.8)	1,285 (84.9%)
Never had asthma	224,656 (88.1%)	19,769 (82.0%)
Cardiovascular disease		
Myocardial infarction	8,457 (88.4%)	477 (87.4%)
Cerebrovascular accident	6,301 (88.9%)	463 (89.4%)
Chronic heart disease	8,492 (90.6%)	149 (89.8%)
COPD	16,158 (87.7%)	1,757 (87.4%)
Cancer (Not Skin)	15,022 (91.9%)	1,350 (90.3%)
Diabetes	24,648 (90.1%)	1,745 (87.3%)

Figure 3.1: Comparison of significant variables from stepwise regression by data source



*Variable was added despite failing to be significant for NHIS

Table 3.2: Results of significant variables from a stepwise regression for BRFSS respondents using health insurance status as the dependent variable

Variable	**Odds Ratio (CI) BRFSS**
Region	0.92 [0.91, 0.93]
Sex	0.89 [0.87, 0.91]
Age group	1.10 [1.09, 1.11]
Race	0.97 [0.96, 0.98]
Hispanic	2.23 [2.16, 2.30]
General Health	0.86 [0.86, 0.87]
Marital Status	0.86 [0.86, 0.87]
Education	1.56 [1.55, 1.58]
Employment	1.03 [1.03, 1.04]
Flu Vaccine	0.44 [0.43, 0.45]
Pneumonia Vaccine	0.89 [0.87, 0.91]
Smoking status	1.20 [1.19, 1.21]
Checkup past year	0.68 [0.68, 0.69]
Asthma Diagnosis	0.94 [0.92, 0.95]
Chronic Heart Disease Diagnosis	0.91 [0.86, 0.97]
COPD Diagnosis	0.93 [0.89, 0.97]
Cancer Diagnosis	0.87 [0.83, 0.91]

Table 3.3: Results of significant variables from a stepwise regression for NHIS respondents using health insurance status as the dependent variable

Variable	**Odds Ratio (CI) NHIS**
Region	0.91 [0.89, 0.93]
Age group	1.06 [1.04, 1.08]
Hispanic	1.95 [1.84, 2.07]
General health	0.85 [0.83, 0.87]
Diabetes	0.78 [0.70, 0.88]
Marital status	0.90 [0.89, 0.92]
Education	1.58 [1.54, 1.61]
Employment	0.95 [0.94, 0.96]
Flu Vaccine	0.40 [0.38, 0.43]
Smoking status	1.19 [1.17, 1.22]
Checkup past year	0.71 [0.70, 0.72]
Cancer Diagnosis	0.76 [0.66, 0.88]

Table 3.4: Chi-square test performed on matching variables post-match

	1-to-1 NN NR	**1-to-1 NN R**	**2-to1 NN NR**	**2-to-1 NN R**	**Caliper**	**Mah.**
Demographic Variables						
Race						
White	0.0113	0.0037	<.0001	<.0001	0.0113	<.0001
Black	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
American Indian/Alaska Native	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
Asian	0.0003	0.0126	<.0001	0.0004	0.0003	**0.2714**
Other	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
Multi-Racial	<.0001	0.0002	<.0001	<.0001	<.0001	**0.407**
Hispanic Ethnicity						
Yes	**0.0554**	0.0329	<.0001	0.0026	**0.0554**	<.0001
Age						
18-24	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
25-34	0.0036	0.0016	<.0001	<.0001	0.0036	<.0001
35-44	<.0001	0.007	<.0001	0.0001	<.0001	<.0001
45-54	**0.8946**	**0.6345**	**0.354**	**0.5014**	**0.8946**	0.0044
55-64	**0.3601**	**0.5389**	0.0328	**0.3849**	0.3601	<.0001
Sex						
Female	**0.8145**	**0.5456**	**0.7643**	**0.3927**	**0.8145**	<.0001
Marital Status						
Married	**0.1436**	**0.6715**	**0.3716**	**0.5486**	**0.1436**	<.0001
Divorced	**0.3185**	**0.3601**	0.0278	**0.1955**	**0.3185**	<.0001
Widowed	0.0032	0.0164	<.0001	0.0007	0.0032	<.0001
Separated	**0.1135**	0.0431	0.0322	0.0042	**0.1135**	0.0024
Never Married	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001

Member of an unmarried couple	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
Education	**1-to-1 NN NR**	**1-to-1 NN R**	**2-to1 NN NR**	**2-to-1 NN R**	**Caliper**	**Mah.**
Never	0.0337	**0.0951**	0.0022	0.0183	0.0337	**0.469**
Elementary (1st—8th)	<.0001	<.0001	<.0001	<.0001	<.0001	**0.9626**
Some High School (9th—11th)	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
High School Graduate	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
Some College	<.0001	0.0012	<.0001	<.0001	<.0001	<.0001
College (4+ years)	0.0003	0.0014	<.0001	<.0001	0.0003	<.0001
Employment						
Employed for wages	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
Out of work > 1 yr.	<.0001	<.0001	<.0001	<.0001	<.0001	**0.0668**
Out of work < 1 yr.	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
Homemaker	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
Student	0.0006	0.0012	<.0001	<.0001	0.0006	**0.5878**
Retired	**0.2267**	**0.246**	0.0109	**0.1009**	**0.2267**	<.0001
Unable to work	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
Region						
Northeast	**0.4603**	**0.6976**	**0.2964**	**0.5827**	**0.4603**	<.0001
South	0.0022	0.004	<.0001	<.0001	0.0022	<.0001
Midwest	<.0001	0.0003	<.0001	<.0001	<.0001	<.0001
West	**0.2841**	**0.3719**	0.0108	**0.2067**	**0.2841**	<.0001
Behavior Variables						
Smoking Status						
Never Smoked	**0.0589**	0.049	0.0004	0.0054	**0.0589**	<.0001
Former Smoker	0.0138	0.0214	<.0001	0.0011	0.0138	<.0001
Smokes on some days	0.0002	0.001	<.0001	<.0001	0.0002	<.0001
Smokes everyday	0.0049	0.0225	<.0001	0.0013	0.0049	<.0001

Routine Checkup with M.D.	**1-to-1 NN NR**	**1-to-1 NN R**	**2-to1 NN NR**	**2-to-1 NN R**	**Caliper**	**Mah.**
Within 1 year	0.0012	0.0416	<.0001	0.0039	0.0012	<.0001
Within 2 years	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
Within 5 years	**0.8957**	**0.4982**	**0.2583**	**0.3381**	**0.8957**	0.0061
5+ years	0.0008	0.0042	<.0001	<.0001	0.0008	<.0001
Never	<.0001	<.0001	<.0001	<.0001	<.0001	**0.1596**
Flu Immunization	0.034	**0.1068**	0.0002	0.0226	0.034	<.0001
Pneumonia Immunization	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
BMI Categories						
Obese (30.0+)	0.0232	0.0498	0.0143	0.0055	**0.0232**	<.0001
Overweight (25.0—29.9)	**0.8049**	**0.6622**	**0.8825**	**0.5367**	**0.8049**	<.0001
Normal Weight (18.5—24.9)	**0.4736**	**0.5535**	**0.534**	**0.402**	**0.4736**	<.0001
Underweight (<18.5)	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
General Health						
Excellent	**0.1517**	**0.2574**	0.0013	**0.1093**	**0.1517**	<.0001
Very good	**0.1541**	**0.2342**	0.0029	**0.0925**	**0.1541**	0.0352
Good	**0.9918**	**0.7976**	0.896	**0.7169**	**0.9918**	<.0001
Fair	**0.1637**	**0.0733**	**0.0535**	0.0113	**0.1637**	<.0001
Poor	0.0033	0.0094	<.0001	0.0002	0.0033	<.0001
Chronic Disease Indicators						
Asthma						
Now has asthma	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
No longer has asthma	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
Never had asthma	0.0039	0.0148	<.0001	0.0006	0.0039	<.0001

Cardiovascular disease	**1-to-1 NN NR**	**1-to-1 NN R**	**2-to1 NN NR**	**2-to-1 NN R**	**Caliper**	**Mah.**
Myocardial infarction	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
Cerebrovascular accident	**0.0663**	**0.0662**	0.0087	0.0094	**0.0663**	<.0001
Chronic heart disease	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
COPD	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
Cancer (Not Skin)	**0.4027**	**0.4636**	**0.0987**	**0.30**	**0.4027**	**0.8992**
Diabetes						
Yes	0.0192	**0.0841**	0.0034	0.0146	0.0192	<.0001
No	0.0092	0.043	<.0001	0.0042	0.0092	<.0001
Borderline	**0.865**	**0.9662**	**0.1678**	**0.9522**	**0.865**	**0.4529**

NN=Nearest Neighbor
NR=No Replacement
R=Replacement
Mah.=Mahalanobis

Discussion

Because matching variables may be limited because of different goals of each survey, it is advantageous to include the least number of matching variables possible. The specific goal of this paper was to conduct a sensitivity analysis of the matching variables needed to impute accurately, health insurance status. Looking at balance statistics, we investigated which algorithms more efficiently matched the NHIS respondents to the BRFSS respondents.

We found that when fusing data, it is critical to examine variables most predictive of your variable of interest. Asthma was significant in the stepwise analysis, but was not efficiently matched in any of the algorithms. This may be because of the relatively low sample size for respondents with low educational attainment in the NHIS (n=60 for never and n=615 for elementary education only). While we do not believe we should have held back a significant variable from the propensity score calculation based upon sample sizes alone, sample size may have a far-reaching affect on matching (Soong, 2001) and may underestimate the probability of rare events (Baser, 2006). Researchers should beware of sample sizes prior to matching data. Sample sizes too large will overpower your study due to tight confidence intervals and small sample sizes will affect the efficiency of the match. Researchers suggest that when attempting to match two databases, the rule of thumb is that one data source should be no more than nine times as large as the other source you are attempting to match (Baser, 2006).

Researchers should also beware that variables that were not significant in the original propensity score calculation may not be matched well in the fused data. However, the choice of these matching variables should be based upon theory and not purely statistical (D'Orazio et al.,

2006). Therefore, these variables should not be the only variables you are matching on but should be included in the model.

The Mahalanobis algorithm failed to match many of the categories of variables that the Nearest Neighbor Methods were able to match based upon our chi-square tests. Matching with a 0.20 caliper, which discards values outside the suggested range defined by the caliper, was very similar to nearest neighbor algorithms in its efficiency. The caliper did not seem to affect variable matching efficiency. Nearest Neighbor with a 1-to-1 ratio without replacement; however, replacement with the 2-to-1 Nearest Neighbor fared much better than without replacement.

When deciding between matching algorithms, one should consider the ability of the algorithm to efficiently match on categories of the matching variables used for the propensity score calculation. Failing to do so may result in unbalanced fused data that may bias estimates (Baser, 2006). We have shown that for this specific case, some algorithms fare better than others in balancing matching variables. If none of the algorithms efficiently matched on variables used for the calculation of the propensity score, researchers should consider re-specifying the model since the model may not be correctly specified.

CHAPTER 4

DISCUSSION

Summary of Key Findings

This study investigated the feasibility of data fusion techniques to integrate data from two separate, independently collected surveys. We aimed to first match the BRFSS and the NHIS datasets using propensity scores and matching algorithms. The first section investigated what has been done in research pertaining to data fusion. The types of data integration techniques and their strengths and weakness were also investigated. The second section was a manuscript ready for publication in a public health journal. The goal of this first paper was to implement data fusion using Nearest Neighbor (with and without replacement and using 2-to-1 and 1-to-1 ratios of the control to treatment), Caliper, and Mahalanobis algorithms and investigate balance statistics after matching. The third section was another manuscript to be submitted to an information fusion journal. The goal of this manuscript was to conduct a sensitivity analysis of the matching variables and investigate each algorithm's ability to match on these variables.

For the first manuscript, our research aimed to test the ability of data fusion methods to predict health insurance. Percentages of respondents with health insurance in our fused data were all similar (within 2-3 percentage points) to the original BRFSS data. The usual approach is to see if the match is within the 95% confidence interval. However, with a very large sample, these confidence intervals were quite narrow, so a match that is clinically relevant and useful may fail this overly rigorous statistical test.

Matching with replacement in the Nearest Neighbor algorithms increased the algorithm's ability to accurately predict health insurance status, decreased false positives and false negatives. The 2-to-1 Nearest Neighbor algorithms did not differ from the 1-to-1 algorithms when replacement was used. However, there was a slight increase in sensitivities and specificities with 1-to-1 Nearest Neighbor without replacement. We believe this was the case because of our large sample size. When not using replacement, the algorithm had an ample pool of respondents from which to match. When replacing the respondent back into the pool, that respondent was used multiple times and the ratio of BRFSS to NHIS respondents did not matter.

A caliper of 0.20 did not have an effect because the sample size was so large; the algorithm again had a large pool of respondents from which to pull from for an accurate match. To investigate the effect of the sample size on caliper matching, we randomly chose 50,000 and 100,000 respondents from the BRFSS dataset. As the sample size decreased, more respondents were held back from the matched dataset since the caliper was not met for respondents' to be matched. This is because this method discards values outside the suggested range defined by the caliper (0.20 of a standard error of the estimated propensity score).

For the second manuscript, we found that, when fusing data, it is critical to examine variables most predictive of your variable of interest. It became evident that race and sex did not need to be included in the propensity score calculation, since the algorithms consistently did not match respondents on these variables. However, asthma was significant in the stepwise analysis, but was not efficiently matched in any of the algorithms.

Significance of The Work

In the review of the literature it became evident that data fusion is under-utilized in the field of public health. Since public health agencies need to keep the ever-decreasing number of participants from breaking off after an interview is started, it is important that organizations conducting surveillance activities investigate innovative techniques of combining data from multiple, less extensive surveys.

The analyses conducted in this study suggest that agencies should proceed with caution when wanting to utilize data fusion for public health prevalence. For surveillance programs with large sample sizes and tight confidence limits, data fusion may not be appropriate. Data fusion may be more appropriate for small sample studies that want to integrate their data with another small study or for estimates unavailable to any surveillance agency.

This study is also significant because it offers public health agencies looking to conduct data fusion, techniques to test for the efficiency of matching their data before matching an unobserved variable in their data. This is because matching depends largely on the structure of the data (Baser, 2006). To test for data fusion's efficacy to match their data, agencies need to conduct the following analyses using an observed variable in their data and another source:

1. Review the literature to know the context of the health issue and what is correlated to it as this will assist in basing your decision upon theory and not purely upon statistical models.
2. Prepare data from both sources by harmonizing and re-categorizing variables.

3. Conduct a univariate analysis of the prospective matching variables to investigate whether any variable is associated with the variable you are trying to impute.
4. Conduct three logistic regressions:
    a. One with the variable you are imputing as the dependent variable and the matching variables from the univariate analysis as independent variables in Data Source 1.
    b. Another logistic regression with the variable you are imputing as the dependent variable and the matching variables from the univariate analysis as independent variables in Data Source 1.
    c. Use the variables that were significant in both sources as independent variables and the data source as the dependent variable and conduct a descending stepwise regression to calculate propensity scores.
5. Use the propensity scores and match using a variety of matching algorithms, such as Nearest Neighbor, Caliper, and Mahalanobis.
6. Assess matching quality by testing whether the match was able to balance matching variables.

Limitations of The Work

Our study was subject to a number of limitations. One of the major limitations was the fact that the respondents were not selected via simple random sampling methods. Both the BRFSS and the NHIS oversample certain races and regions to have adequate sample for weighting certain races or rural areas (Centers for Disease Control and Prevention, 2013b; National Center for Health Statistics, 2015). Another limitation is that the two surveys used in

our research had different survey administration modes. The BRFSS is a telephone only survey instrument and the NHIS is an in-person survey instrument. This may lead to biased estimates because of a possible survey mode effect.

Future Directions

The results of this research were encouraging for possible use in public health, especially when there are no estimates available for certain topics. Some matters that should be considered in more detail in future research are the effect of sample size on data fusion and possible inferiority or superiority analysis. Finally more research on whether or not confidence intervals are most appropriate test of significance for overpowered studies such as ours would be useful.

We tested the affect of sample size on caliper and found that the algorithm was more exploited as the sample sized decreased. More research also needs to be conducted on whether our large sample size had an effect on the other algorithms' ability to efficiently match respondents. Research looking at how sample size effects matching when replacing a respondent back into the matching pool would be useful to the body of public health research.

Similar to a clinical trial non-inferiority analysis, research into whether our imputed estimate of health insurance status is inferior to current public health estimates should be conducted. Non-inferiority analyses in clinical trials test whether the intervention is unacceptably worse in the treatment than in the control (Schumi & Wittes, 2011). Such analyses may be more appropriate since in our case, we are attempting to ascertain whether our estimates are less accurate than national estimates that are currently available.

Our model uses traditional equivalence confidence interval to test whether the imputed estimate are within 95% of the mean. However, traditional significance tests may not be ideal in

data fusion since it calculates whether the estimates are equivalent. With a chosen non-inferiority margin, public health agencies can ascertain whether imputed estimates meet their threshold of acceptance, not a threshold of equivalence.

It may be useful to use our approach but calculate a new model based upon another dependent variable rather than health insurance status. For example, we can test whether visits to a medical profession can be imputed using our model. However, it is our belief that our model is not a one-size-fits-all solution and that it is only applicable for health insurance status. For this reason, researchers should take time to investigate the dependent variable they are hoping to impute and its association with variables used to match. Including variables in a model should be based on theoretical reasons because it is known to be associated in other research, not merely because they are significant or not in the model.

Additionally, future research also needs to be conducted on whether matching two different data sources based upon variables found to be significant in our analysis would be effective in imputing health insurance status. This would alleviate the time needed to retest for significance for the new data sources. This would also suggest that our methods are valid.

Future research needs to be conducted on the effect of weighting and survey modes on data fusion techniques for public health estimates. This may have caused unobservable bias in our estimates. One way researchers can investigate the impact of survey weights on imputed estimates through data fusion can be to bring in matched respondents' survey weights from the BRFSS and comparing weighted estimates pre- and post match.

Conclusion

Our study found that estimates of health insurance imputed through matching were close to the original respondent's unweighted health insurance status. By fusing data from two public health surveillance data, we were able to calculate health insurance estimates within 2-3% of the original data. This is promising considering the fact that national estimates differ from survey to survey even when taking survey administration into consideration. For example, in 2014, the national estimate of heath insurance for individuals less than 65 years of age was 86.7% for the National Health Interview Survey and 88.0% from Census Current Population Survey data (both in-person surveys collected by the Census Bureau). (CDC National Center for Health Statistics, 2014; Smith & Medalia, 2014). Especially when there is a need to calculate estimates not otherwise found in any database, data fusion is a viable option.

Because data structure differs from survey to survey and topic to topic, agencies wanting to conduct data fusion should check balance statistics for their data using an observable variable from both data sources. Unbalanced fused data may suggest that the researcher needs to re-specify the logistic regression calculating the propensity score. Failing to do so may result in unbalanced fused data that may bias estimates (Baser, 2006). We have shown that for our research, some algorithms fare better than others in balancing matching variables.

REFERENCES

American Association for Public Opinion Research. (2008). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. Retrieved from https://www.aapor.org/AAPORKentico/AAPOR_Main/media/MainSiteFiles/Standard_Definitions_07_08_Final.pdf

ARF. (2003). *ARF Guidlines for Data Integration*. New York. Retrieved from http://s3.amazonaws.com/thearf-org-aux-assets/downloads/research/DataIntegrationGuidelines.pdf

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*, *28*(25), 3083–107. doi:10.1002/sim.3697

Baser, O. (2006). Too Much Ado about Propensity Score Models? Comparing Methods of Propensity Score Matching. *Value in Health*, *9*(6), 377–385. doi:10.1111/j.1524-4733.2006.00130.x

Bendel, R. B., & Afifi, A. A. (1977). Comparison of Stopping Rules in Forward "Stepwise" Regression. *Journal of the American Statistical Association*, *72*(357), 46–53. doi:10.1080/01621459.1977.10479905

Call, K. T., Davidson, G., Davern, M., Brown, E. R., Kincheloe, J., & Nelson, J. G. (2008a). Accuracy in self-reported health insurance coverage among medicaid enrollees. *Inquiry*, *45*(4), 438–456. doi:10.5034/inquiryjrnl_45.04.438

Call, K. T., Davidson, G., Davern, M., Brown, E. R., Kincheloe, J., & Nelson, J. G. (2008b). Accuracy in self-reported health insurance coverage among medicaid enrollees. *Inquiry*, *45*(4), 438–456. doi:10.5034/inquiryjrnl_45.04.438

Cannon, H. M., Smith, J. A., & Williams, D. L. (2008). A Data-Overlay Approach to Synthesizing Single-Source Data. *Journal of Advertising*, *36*, 7–18. doi:10.2753/JOA0091-3367360401

CDC National Center for Health Statistics. (2014). Health Insurance Coverage. *FastStats*. Retrieved December 8, 2015, from http://www.cdc.gov/nchs/fastats/health-insurance.htm

Centers for Disease Control and Prevention. (2013a). Behavioral Risk Factor Surveillance System. *About the Behavioral Risk Factor Surveillance System (BRFSS)*. Retrieved from http://www.cdc.gov/brfss/about/about_brfss.htm

Centers for Disease Control and Prevention. (2013b). Behavioral Risk Factor Surveillance System Summary Data Quality Report. Retrieved from http://www.cdc.gov/brfss/annual_data/2012/pdf/SummaryDataQualityReport2012_20130712.pdf

Centers for Disease Control and Prevention. (2015). BRFSS Prevalence and Trends Data. *BRFSS Data By Location*. Retrieved January 1, 2015, from http://www.cdc.gov/brfss/brfssprevalence/index.html

Coca-Perraillon, M. (2007). *Local and Global Optimal Propensity Score Matching* (No. 185). Cary, NC. Retrieved from http://www2.sas.com/proceedings/forum2007/185-2007.pdf

D'Orazio, M., Di Zio, M., & Scanu, M. (2006). *Statistical Matching: Theory and Practice* (First Edit.). Hoboken: Wiley.

GfK Mediamark Research & Intelligence LLC. (2014). Nielsen TV/GfK MRI Data Fusion. Retrieved from http://www.gfkmri.com/Products/NielsenTVDataFusion.aspx

Gilula, Z., Mcculloch, R. E., & Rossi, P. E. (2006). A Direct Approach to Data Fusion. *Journal of Marketing Research*, *43*(1), 73–83. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=549263

Ho, D., Imai, K., King, G., & Stuart, E. (2007). Matchit: Nonparametric Preprocessing for Parametric Causal Inference. *Journal of Statistical Software*. Retrieved from http://gking.harvard.edu/matchit/

Jamal, A., Homa, D., O'Connor, E., Babb, S., Caraballo, R., Singh, T., … King, B. (2015). *Current Cigarette Smoking Among Adults — United States, 2005–2014*. *Morbidity and Mortality Weekly Report (MMWR)*. Atlanta, GA. Retrieved from http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6444a2.htm?s_cid=mm6444a2_w

Kum, H., & Masterson, T. (2008). Statistical Matching Using Propensity Scores: Theory and Application to the Levy Institute Measure of Economic Well-Being. *SSRN Electronic Journal*, (535). doi:10.2139/ssrn.1136363

McCune, B., & Grace, J. B. (2002). Distance Measures. In *Analysis of Ecological Communities* (pp. 45–57). Gleneden Beach, Oregon: MjM Software Design.

Moriarity, C., & Scheuren, F. (2004). Regression-based statistical matching: recent developments. *ASA Proceedings of the Joint Statistical Meetings, American Statistical Association*, 4050–4057.

National Center for Health Statistics. (2015). NHIS - About the National Health Interview Survey. *About the National Health Interview Survey*. Retrieved October 16, 2015, from http://www.cdc.gov/nchs/nhis/about_nhis.htm

Nielsen. (2008). *Spectra: Powerful Insights for Consumer Targeting*. Retrieved from http://www.nielsen.com/content/dam/nielsen/en_us/documents/pdf/Fact Sheets/Nielsen Spectra Overview - Powerful Insights for Consumer Targeting.pdf

R Core Team. (2015). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org/

Radner, D. (1981). An Example of the Use of Statistical Matching in the Estimation and Analysis of the Size Distribution of Income. *Review of Income and Wealth*, *27*(3), 211–242.

Rogers, W. L. (1978). An Evaluation of Statistical Matching. *Journal of Marketing Research*, *15*(1), 103–112.

Rolstad, S., Adler, J., & Rydén, A. (2011). Response burden and questionnaire length: is shorter better? A review and meta-analysis. *Value in Health : The Journal of the International Society for Pharmacoeconomics and Outcomes Research*, *14*(8), 1101–8. doi:10.1016/j.jval.2011.06.003

Rosenbaum, P. R., & Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, *70*(1), 41–55.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. *The American Statistician*, *39*(1), 33–38. doi:10.1080/00031305.1985.10479383

SAS Institute Inc. (2014). SAS. Cary, NC.

Scarborough. (2015). Nielsen Local. *LOCAL MARKET CONSUMER INSIGHTS*. Retrieved from www.scarborough.com

Schumi, J., & Wittes, J. T. (2011). Through the looking glass: understanding non-inferiority. *Trials*, *12*(1), 106. doi:10.1186/1745-6215-12-106

Sissors, J. Z. (1971). Matching Media with Markets. *Journal of Advertising Research*, *11*(5), 39–43.

Smith, J. C., & Medalia, C. (2014). *Health Insurance in the United States : Coverage in the United States : 2013*. Washington D.C. Retrieved from http://www.census.gov/content/dam/Census/library/publications/2014/demo/p60-250.pdf

Soong, R. (2001). Data Fusion in Latin America. In *47th Annual Conference of the Advertising Research Foundation*. New York. Retrieved from http://www.zonalatina.com/Zldata166.htm

Soong, R., & de Montigny, M. (2001). The anatomy of data fusion. *Proceedings of the 2001 Worldwide Readership Research Symposium. Venice*, (1984), 1–23. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:THE+ANATOMY+OF+DATA+FUSION#0

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science : A Review Journal of the Institute of Mathematical Statistics*, *25*(1), 1–21. doi:10.1214/09-STS313

Takayama, M., Wetmore, C. M., & Mokdad, A. H. (2012). Characteristics associated with the uptake of influenza vaccination among adults in the United States. *Preventive Medicine*, *54*(5), 358–362. doi:10.1016/j.ypmed.2012.03.008

The Kaiser Commission on Medicaid and the Uninsured. (2014). *Key Facts about the Uninsured Population*. Menlo Park. Retrieved from http://kff.org/uninsured/fact-sheet/key-facts-about-the-uninsured-population/

van der Laan, P., & van Nunspeet, W. (2009). *Modernising Household Surveys in the Netherlands: Design, Efficiency Gains and Perspectives* (No. Discussion paper (09044)). The Hague/Heerlen. Retrieved from http://www.cbs.nl/NR/rdonlyres/0D30D23B-FE40-4570-B41A-E9B2CADF01DB/0/200944x10pub.pdf

Van Der Puttan, P., Nok, J. N., & Gupta, A. (2002). *Data Fusion through Statistical Matching* (No. 185) (Vol. Paper 185). Retrieved from http://ebusiness.mit.edu/research/papers/185_gupta_data_fusion.pdf

van Hattum, P., & Hoijtink, H. (2008). The proof of the pudding is in the eating Data fusion: An application in marketing. *Journal of Database Marketing and Customer Strategy Management*, *15*, 267–284. doi:10.1057/dbm.2008.24

Wicklin, R. (2012). What is Mahalanobis distance? *The DO Loop*. Retrieved August 29, 2015, from http://blogs.sas.com/content/iml/2012/02/15/what-is-mahalanobis-distance.html

APPENDICES

Appendix A: Code

```
/*************************************************
        PSmatching.sas   adapted from

        Paper 185-2007  SAS Global Forum 2007
        Local and Global Optimal Propensity Score Matching
        Marcelo Coca-Perraillon
        Health Care Policy Department, Harvard Medical School, Boston, MA

        ------------------------------
Treatment and Control observations must be in separate datasets such that
  Control data includes: idC =  subject_id, pscoreC = propensity score
  Treatment data includes: idT, pscoreT
  id must be numeric

        method = NN (nearest neighbor), caliper, or radius

        caliper value = max for matching

        replacement = yes/no  whether controls can be matched to more than one case

        out = output data set name

        example call:

         %PSMatching(datatreatment= T, datacontrol= C, method= NN,
          numberofcontrols= 1, caliper=, replacement= no, out= matches);

Output format:
        Id              Matched
    Selected    PScore      To      PScore
Obs    Control   Control   TreatID    Treat
```

```
1     18628    0.39192    16143    0.39192
2     18505    0.23029    16158    0.23002
3     15589    0.29260    16112    0.29260
```

All other variables discarded. Reformat for merge on subject_id with original data:

```
 data pairs;
   set  matches;
          subject_id = IdSelectedControl; pscore = PScoreControl; pair = _N_;
          output;
          subject_id = MatchedToTreatID; pscore = PScoreTreat; pair = _N_;
          output;
          keep subject_id pscore pair;


**********************************************/

%macro PSMatching(datatreatment=, datacontrol=, method=, numberofcontrols=, caliper=,
replacement=, out=);

/* Create copies of the treated units if N > 1 */;
data _Treatment0(drop= i);
 set &datatreatment;
 do i= 1 to &numberofcontrols;
 RandomNumber= ranuni(12345);
output;
end;
run;
/* Randomly sort both datasets */
proc sort data= _Treatment0 out= _Treatment(drop= RandomNumber);
by RandomNumber;
run;
data _Control0;
set &datacontrol;
RandomNumber= ranuni(45678);
run;
proc sort data= _Control0 out= _Control(drop= RandomNumber);
```

```
by RandomNumber;
run;

data Matched(keep = IdSelectedControl PScoreControl MatchedToTreatID PScoreTreat);
 length pscoreC 8;
 length idC 8;
/* Load Control dataset into the hash object */
 if _N_= 1 then do;
declare hash h(dataset: "_Control", ordered: 'no');
declare hiter iter('h');
h.defineKey('idC');
h.defineData('pscoreC', 'idC');
h.defineDone();
call missing(idC, pscoreC);
end;
/* Open the treatment */
set _Treatment;
%if %upcase(&method) ~= RADIUS %then %do;
retain BestDistance 99;
%end;
/* Iterate over the hash */
rc= iter.first();
if (rc=0) then BestDistance= 99;
do while (rc = 0);
/* Caliper */
%if %upcase(&method) = CALIPER %then %do;
if (pscoreT - &caliper) <= pscoreC <= (pscoreT + &caliper) then do;
ScoreDistance = abs(pscoreT - pscoreC);
if ScoreDistance < BestDistance then do;
BestDistance = ScoreDistance;
IdSelectedControl = idC;
PScoreControl =  pscoreC;
MatchedToTreatID = idT;
PScoreTreat = pscoreT;
end;
end;
```

```
%end;
/* NN */
%if %upcase(&method) = NN %then %do;
ScoreDistance = abs(pscoreT - pscoreC);
if ScoreDistance < BestDistance then do;
BestDistance = ScoreDistance;
IdSelectedControl = idC;
PScoreControl =  pscoreC;
MatchedToTreatID = idT;
PScoreTreat = pscoreT;
end;
%end;
%if %upcase(&method) = NN or %upcase(&method) = CALIPER %then %do;
rc = iter.next();
/* Output the best control and remove it */
if (rc ~= 0) and BestDistance ~=99 then do;
output;
%if %upcase(&replacement) = NO %then %do;
rc1 = h.remove(key: IdSelectedControl);
%end;
end;
%end;
/* Radius */
%if %upcase(&method) = RADIUS %then %do;
if (pscoreT - &caliper) <= pscoreC <= (pscoreT + &caliper) then do;
IdSelectedControl = idC;
PScoreControl =  pscoreC;
MatchedToTreatID = idT;
PScoreTreat = pscoreT;
output;
end;
rc = iter.next();
%end;
end;
run;
/* Delete temporary tables.  Quote for debugging */
```

```
proc datasets;
delete _:(gennum=all);
run;
data &out;
  set Matched;
run;
%mend PSMatching;
```

Appendix B: IRB Protocol and Letter of Determination

1) Protocol Title

Importance, Implementation, and Validation of a Data Matching Method using the Behavioral Risk Factor Surveillance System and the Medical Expenditure Panel Survey.

This protocol is for a Dissertation project by George Khalil, MPH, Doctor of Public Health (DrPH) candidate in the UGA College of Public Health. Development of research questions and project planning will be advised by George Khalil's Doctoral Committee which include: Dr. Matthew Lee Smith (Committee Chair), Department of Health Promotion and Behavior, Dr. Ye Shen, Department of Epidemiology and Biostatistics, and Dr. Mark Ebell, Department of Epidemiology and Biostatistics. There is one external committee member: Dr. Derek Ford, Centers for Disease Control and Prevention (CDC).

2) Research Design and Methods

This is a quantitative secondary data analysis of de-identified publicly available data. Below is the goal:

Goal: *To create "statistical twins" from two data sources using PSM in order to validate estimates from the matched data. To do this, we will statistically match BRFSS respondents to another data source such as the National Center for Health Statistic's (NCHS's) Medical Expenditure Survey (MEPS). These matched respondents will be considered "statistical twins". We will be using propensity score matching (PSM) to statistically match and link two data sources already available on public use websites.*

The Behavioral Risk Factor Surveillance System and the Medical Expenditure Panel Survey have collected survey responses using approved protocols as required by their respective Institutional Review Board and the Office of Management and Budget (OMB). This project is a secondary data analysis using their de-identified, publicly available data.

3) Study Timelines

Analyses will be conducted by January 2015. These analyses will be used for two publishable journal articles as required by the DrPH program.

4) Procedures Involved

Please see Data Analysis Section below for Procedures. The entirety of the procedures is data analysis since this protocol is for a secondary data analysis.

5) Data and Specimen Banking

Not applicable

6) Data Analysis

***Matching Procedure**: We will first seek to match the BRFSS and the MEPS datasets using propensity score matching. Statistical Analysis System (SAS) version 9.3 will be used for the entirety of this research. The dependent variable we will be using to validate our methods is health insurance coverage and another access to healthcare question such as routine checkup delay. A SAS macro developed by Marcelo Coca-Perraillon will be used to calculate a probability estimates based on a set of covariates. These covariates include demographic variables such as: age, sex, race/ethnicity, marital status, education, region, employment, and income. We will also see what predicted the dependent variables pre-match and build the match around significant variables. The SAS macro will also assist us in matching respondents from each of the data sources based on the PSM score using two matching techniques (caliper and nearest neighbor matching). A sensitivity analysis will be performed with the number and type of variables used in each type of match for a separate paper of this project.*

***Matching Validation:** We will use a split sample validation to test the accuracy of the match. Our planned procedure is as follows (see attached figure). 1. The 2012 BRFSS dataset and the 2011-2012 dataset are randomly split into two halves (Part A and B). 2. The split samples are matched together using procedures described above. 3. The real results from Part A are compared with the matched results from Part B.*

The University of Georgia

Phone 706-542-3199

Fax 706-542-3660

Office of the Vice President for Research
Human Subjects Office

NOT HUMAN RESEARCH DETERMINATION

November 26, 2014

Dear Matthew Smith:

The University of Georgia Institutional Review Board (IRB) reviewed the following protocol on 11/26/2014:

Type of Review:	Initial Study
Title of Study:	Importance, Implementation, and Validation of a Data Matching Method using the Behavioral Risk Factor Surveillance System and the Medical Expenditure Panel Survey
Investigator:	Matthew Smith
IRB ID:	STUDY00001602
Funding:	None
Grant ID:	None

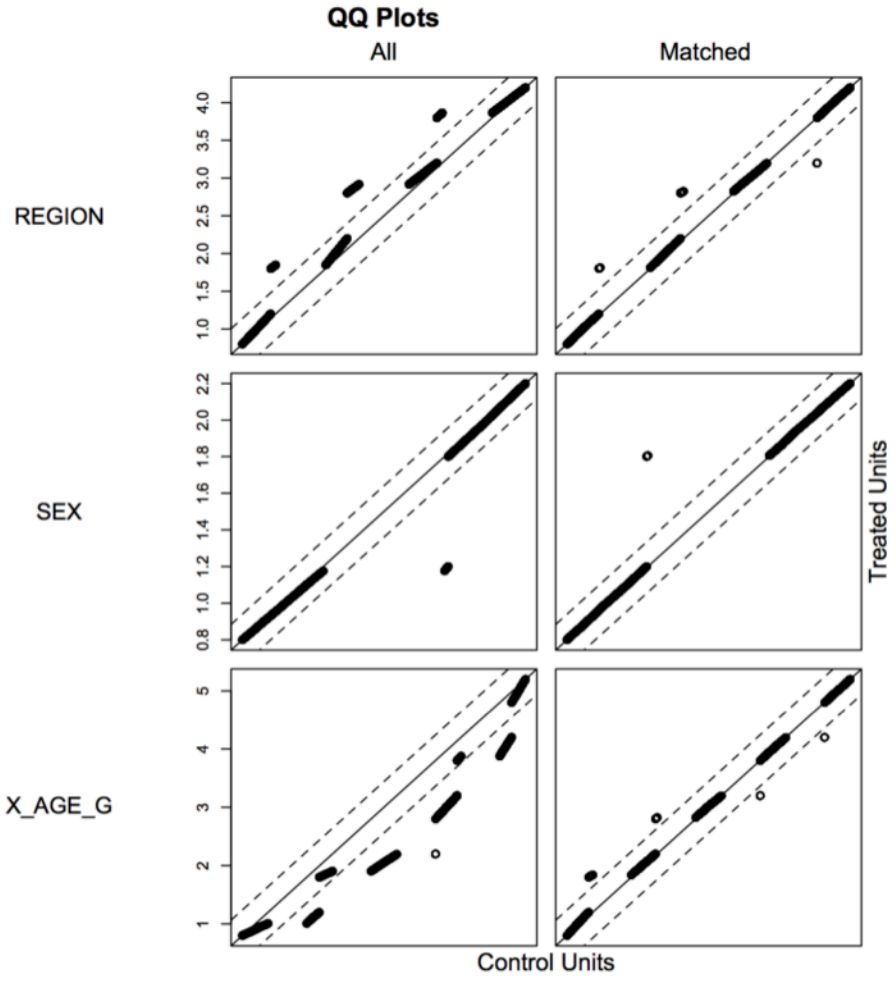
The IRB determined that the proposed activity is not research involving human subjects as defined by DHHS and FDA regulations because it is limited to the analysis of publicly available data that are not individually identifiable.

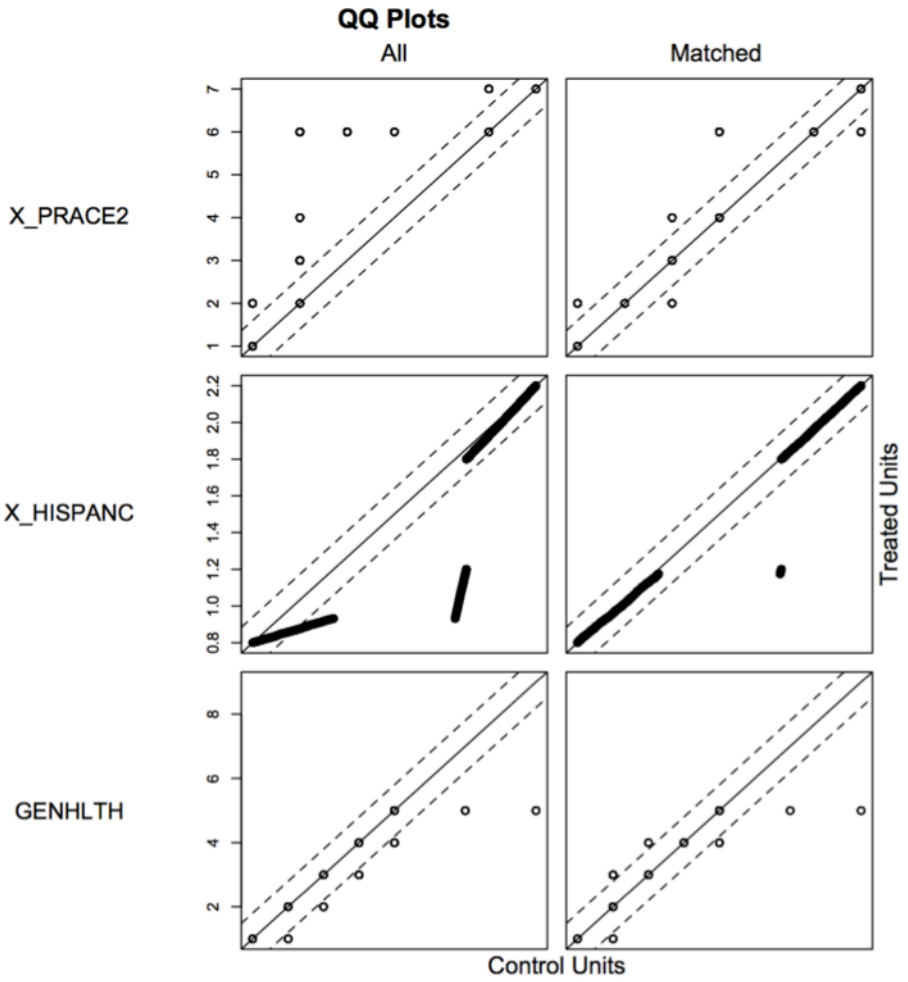
University of Georgia (UGA) IRB review and approval is not required. This determination applies only to the activities described in the IRB submission and does not apply should any changes be made. If changes are made and there are questions about whether these activities are research involving human subjects, please submit a new request to the IRB for a determination.

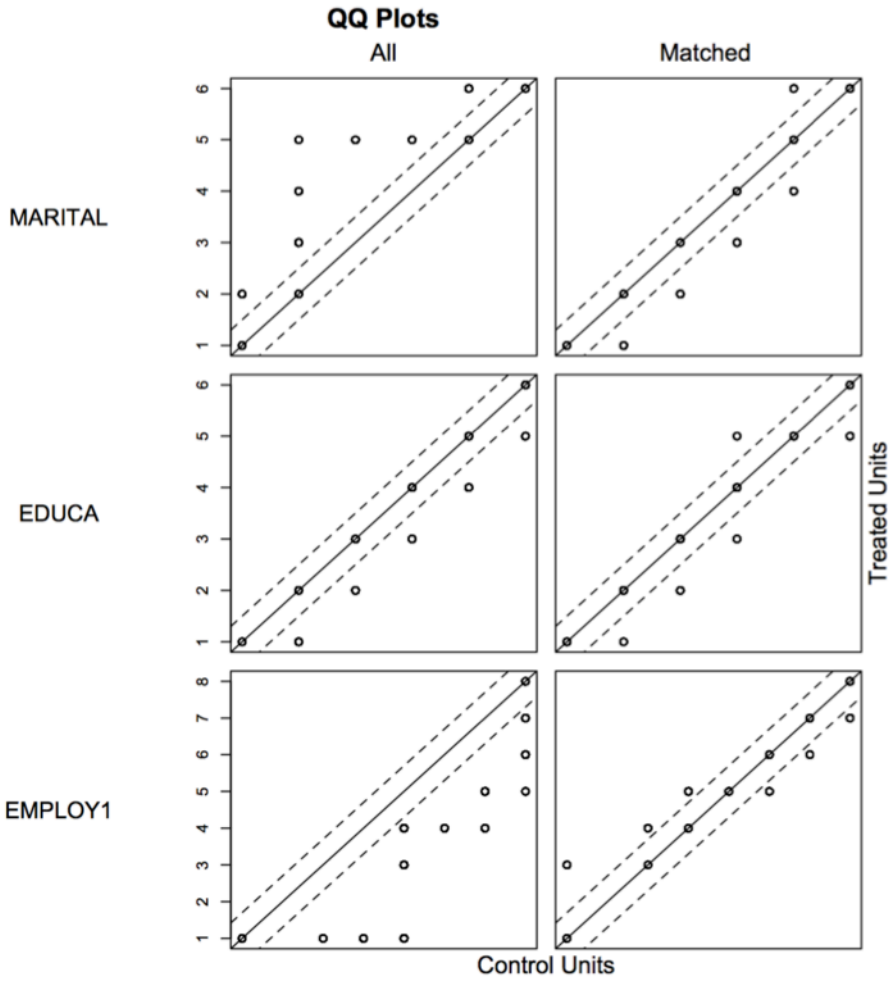
Sincerely,

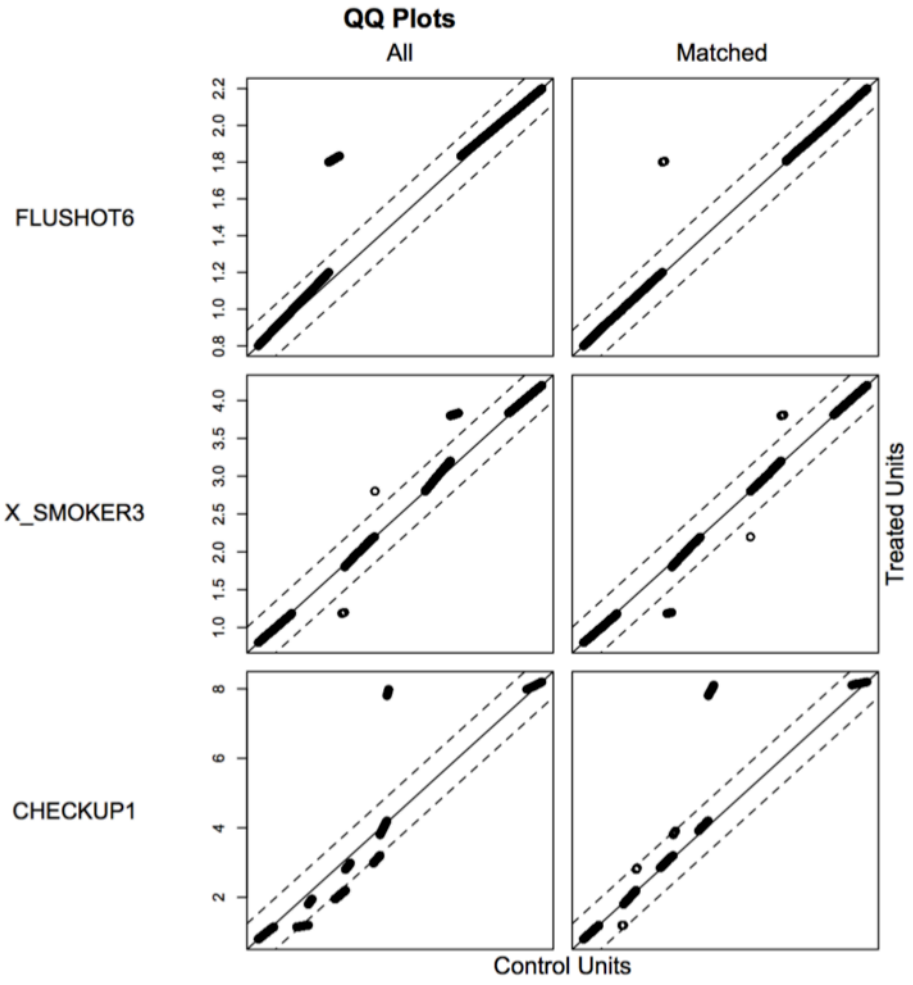
Larry Nackerud, PhD
University of Georgia
Institutional Review Board Chairperson

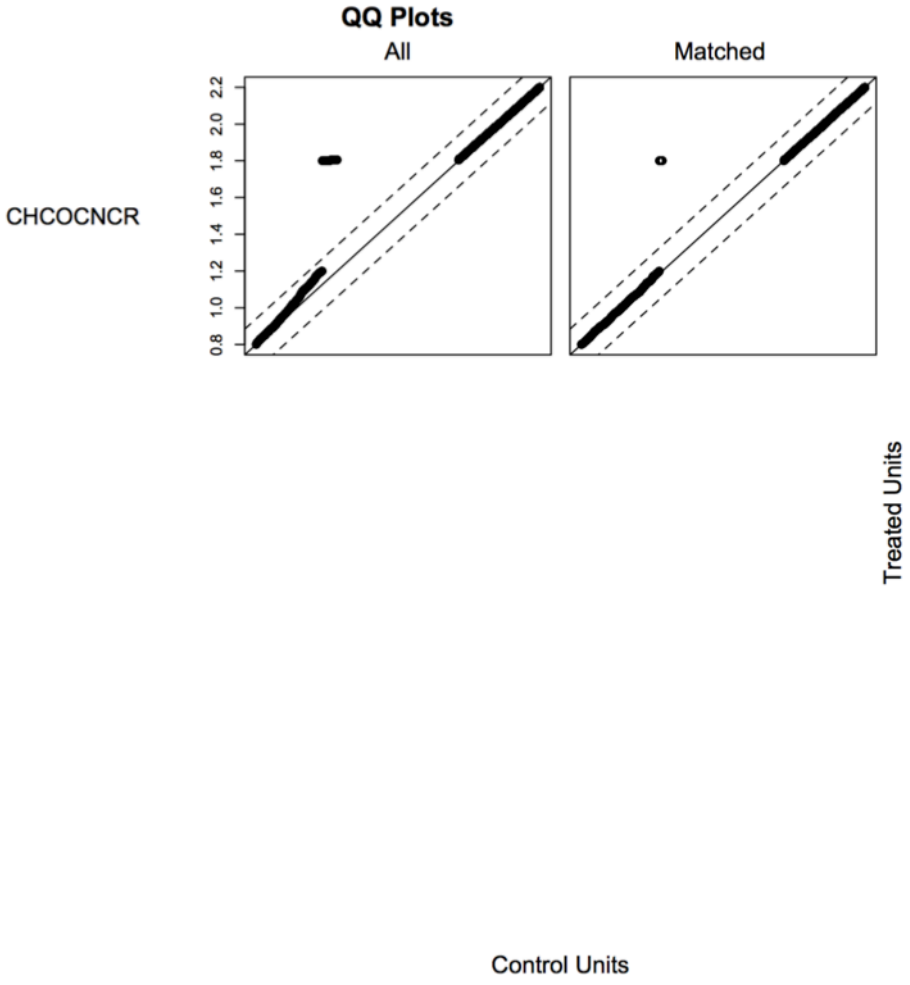
Appendix C: Q-Q Plots of Matching Variables




QQ Plots
All
Matched
X_PRACE2
X_HISPANC
GENHLTH
Treated Units
Control Units


QQ Plots
All
Matched
MARITAL
EDUCA
EMPLOY1
Treated Units
Control Units


QQ Plots
All
Matched
FLUSHOT6
X_SMOKER3
CHECKUP1
Treated Units
Control Units


QQ Plots
All
Matched
CHCOCNCR
0.8
1.0
1.2
1.4
1.6
1.8
2.0
2.2
Treated Units
Control Units