MAX
CENTERS FOR MEDICARE & MEDICAID SERVICES

MEDICAID POLICY BRIEF

MATHEMATICA Policy Research

Brief 16 • January 2013

# Using the MAX-NHANES Merged Data to Evaluate the Association of Obesity and Medicaid Costs

Allison Hedley Dodd and Philip M. Gleason

*The Medicaid Analytic eXtract (MAX) data set is derived from the state reporting of Medicaid eligibility and claims data and is designed to enable research on Medicaid enrollment, service utilization, and expenditures per calendar year at the enrollee level. The National Health and Nutrition Examination Survey (NHANES), a nationally representative survey of the U.S. noninstitutionalized population, collects measured heights and weights, the most accurate way to calculate weight status. This brief uses the 1999–2004 merged MAX-NHANES data to evaluate the association of Medicaid costs and obesity. The merging of these data sets was highly anticipated because each is considered the gold standard within its field. The results of the analysis demonstrate the hazard of using a small national survey (NHANES) with a state-based data system (MAX) to perform cost analyses, particularly when the range of realistic costs is large. Researchers should be aware of how their research decisions will affect the sample size and thus their ability to detect significant findings. With thoughtful consideration, the MAX-NHANES data have the potential to provide valuable insights for researchers and policymakers.*

## Introduction

Obesity among adults has reached alarming levels in the United States. In 2009–2010, 35.9 percent of adults age 20 and older were obese (Flegal et al. 2012). Obesity has both individual- and societal-level costs. At the individual level, obesity is associated with an increased risk of hypertension, type 2 diabetes, stroke, sleep apnea, and certain cancers (NIH 2000). The association of obesity and increased morbidity results in increased medical spending, particularly inpatient, outpatient, and medication spending (Sturm 2002). At the societal level, obesity constitutes a serious economic burden. Recent research estimated that annual medical expenditures in states would be reduced by 6.7 to 10.7 percent in the absence of obesity (Trogdon et al. 2012). In addition, between 22 and 55 percent of state-level obesity costs are financed by Medicare and Medicaid (Trogdon et al. 2012). It is possible that the share of medical costs attributable to obesity is even higher than these estimates because they are based on data from the Medical Expenditure Panel Survey (MEPS), which uses self-reported medical expenditure figures (Trogdon 2012; Finkelstein et al. 2009; Finkelstein et al. 2003) and body mass index (BMI) values calculated from self-reported height and weight data, a method that is known to underestimate the prevalence of obesity (Gorber et al. 2007).

The National Health and Nutrition Examination Survey (NHANES) is a nationally representative survey of the United States that collects measured heights and weights. The Medicaid Analytical eXtract (MAX) data set is derived from the state reporting of Medicaid eligibility and claims data. Because MAX should include enrollment and claims information on all Medicaid recipients, information on Medicaid costs should be

### About This Series

The MAX Medicaid policy issue brief series highlights the essential role MAX data can play in analyzing the Medicaid program. MAX is a set of annual, person-level data files on Medicaid eligibility, service utilization, and payments that are derived from state reporting of Medicaid eligibility and claims data into the Medicaid Statistical Information System (MSIS). MAX is an enhanced, research-friendly version of MSIS that includes final adjudicated claims based on the date of service, and data that have undergone additional quality checks and corrections. CMS produces MAX specifically for research purposes. For more information about MAX, please visit: http://www.cms.gov/Research-Statistics-Data-and-Systems/Computer-Data-and-Systems/MedicaidDataSourcesGenInfo/MAXGeneralInformation.html.

available for any NHANES respondent who reported being on Medicaid and provided a social security number (SSN). Using the NHANES 1999–2004 and MAX 1999–2007 merged data file, we set out to assess the association of Medicaid costs and obesity status using Medicaid claims and measured heights and weights. Linking NHANES and MAX data provides an opportunity to assess the effects of obesity on Medicaid expenditures without relying on self-reported data on Medicaid expenditures or height and weight. Preventing and controlling obesity could be one way for financially burdened states to manage their health care costs.

## Methods

To analyze the association of obesity and Medicaid costs, we used a newly merged data set that consisted of 1999–2007 data from MAX and 1999–2004 data from NHANES. While both data sets are publicly available, the merged data are accessible only through the Research Data Center (RDC) at the National Center for Health Statistics (NCHS).

MAX data are derived from the data that states submit quarterly to the Medicaid Statistical Information System (MSIS) regarding enrollee eligibility and claims paid in each quarter of the federal fiscal year for all of their Medicaid enrollees. Each Medicaid enrollee is classified as belonging to one of four basis-of-eligibility (BOE) groups: child, adult, disabled, and aged (Borck et al. 2012). MAX was designed to enable research on Medicaid enrollment, service utilization, and expenditures per calendar year at the enrollee level. It links claims data to eligibility records and creates summary variables, such as total fee-for-service (FFS) costs, so that analyses do not require the processing of individual claims.

Each year, NHANES selects a nationally representative sample of the noninstitutionalized U.S. population using a complex, stratified, multistage probability cluster sampling design (Flegal et al. 2012). In NHANES 1999–2004, low-income persons, persons age 12–19 and age 60 and older, African Americans, and Mexican Americans were oversampled. NHANES is an ongoing national survey that collects interview data at home and physical examination data at a mobile examination center (MEC). NHANES is considered the gold standard for measuring obesity in the U.S. because it measures participants' height and weight using standardized techniques and equipment and therefore, avoids the potential inaccuracies of self-reported height and weight information. NHANES data are released in two-year cycles, but the year of data collection is available as a restricted variable.

To link the files, the Centers for Disease Control and Prevention (CDC) determined which of the NHANES records were considered "linkage eligible." A record was considered linkage eligible if the respondent provided a date of birth and a SSN. Linkage eligibility was not related to whether the NHANES participant would meet the Medicaid eligibility requirements.[1] After verifying the NHANES SSNs with the Social Security Administration, MAX records were linked to the NHANES participant if the SSN, month and year of birth, and sex matched exactly (Simon et al. 2011).

## Analysis Approach

### Sample Identification

Our analysis population was nonpregnant, full-benefit, non-dual Medicaid FFS adult or disabled enrollees age 20 and older who had a measured height and weight in NHANES in the same year that they were enrolled in Medicaid. NHANES uses a complex weighting strategy to produce national-level estimates. While NHANES is a continuous-data survey, the CDC provides a strategy to adjust the weights in order to create estimates using the six-year data set, NHANES 1999–2004.

The first issue that we encountered in our analytic approach was how to handle the fact that an NHANES participant could match to multiple years of MAX data if the participant was enrolled in Medicaid for multiple years or in multiple states. We opted to limit the linkage between the MAX and NHANES data to one data point per NHANES participant for two reasons: (1) the NHANES weighting strategy is designed based on one data point per participant; and (2) participants who contributed multiple data points to the analysis would have undue influence on the association between costs and obesity. Therefore, we limited our analysis to NHANES participants who matched to MAX data from the same year so that the obesity status would apply to the year that the costs were incurred.

Table 1 illustrates how the size of the analysis population quickly diminished as the analysis approach was implemented. The NHANES 1999–2004 data set contains records for 31,126 participants. Of these, 25,750 participants were linkage eligible—that is, could potentially be linked to the MAX data set. Among the eligible, 11,312 participants linked to a MAX record from any year, but only 7,915 participants matched with a MAX record in the year in which their obesity status was measured in NHANES.

We limited our analysis to adults age 20 and older because the classification approach and short-term impacts on the cost of medical care differ for children and adults. Many of the

immediate health effects of childhood obesity are an increase in risk factors for chronic diseases, such as cardiovascular disease and diabetes, which are more likely to affect costs in the long-term (CDC 2012). The largest concern about childhood obesity is that obese children and adolescents are more likely to become obese adults. Obesity in adulthood is associated with a more immediate set of health problems potentially leading to near-term Medicaid expenditures such as an increased risk of heart disease, type 2 diabetes, stroke, several types of cancers, and osteoarthritis (CDC 2012). Because NHANES oversamples adolescents, this restriction cut our analysis population to 2,180 participants, almost a quarter of the original size. We limited our analysis to nonpregnant adults because the weight for pregnant women would be inflated due to pregnancy. We required a BMI calculated from measured height and weight to obtain participants' obesity status, which reduced the sample size to 1,666.

For the MAX data, many of our restrictions were implemented to ensure that costs among Medicaid participants were comparable theoretically. States are not required to report State Children's Health Insurance Program (SCHIP) data, so the data are not consistent in MAX. SCHIP-only participants are therefore typically removed from Medicaid analyses. We restricted our analysis to full-benefit participants because enrollees with restricted benefits would have lower costs, driven by the types of services they were entitled to receive.[2] These restrictions are typical for Medicaid analyses.

Because we were assessing the association of costs and obesity, we also limited our study population to nondual enrollees, which reduced our sample to 894 respondents. Dual enrollees are aged and disabled individuals who qualify for both Medicaid and Medicare; their costs are typically lower than those of nondual enrollees because some of their benefits—such as inpatient hospital stays, skilled nursing facilities, physician services, and prescription drugs—are covered by Medicare. In 2008, 92.7 percent of aged enrollees were dual eligible (Borck et al. 2012). Because the majority of aged enrollees are dual enrollees, removing duals from the analysis population left too few people in the aged category for analysis. We therefore restricted our analyses to the 856 nondual enrollees who were classified as either adult or disabled.

Finally, we restricted our analysis to the 455 Medicaid enrollees who were not enrolled in comprehensive managed care (CMC) at any time during the year. Comprehensive managed care includes health maintenance organizations (HMO), health insuring organizations (HIO), and the Program of All-Inclusive Care for the Elderly (PACE). Payments for comprehensive managed care are made using capitated payments, a set monthly fee that the state pays regardless of the service use of the enrollee. We assumed that the capitation rates for CMC were not determined based on the obesity status of the participant and therefore did not include them in our analysis.

After restricting the data set to our analysis criteria, we found nine NHANES participants who had two MAX records in the exam year. It is not uncommon for MAX enrollees to have multiple records in one year, either because they reside in more than one state or because there are data issues within a state. To resolve the problem of multiple records, we kept the record that had the highest cost in the analysis, assuming this record represented more of the enrollee's participation.

Because there is a large range of FFS costs among participants and our analysis sample was small, it became apparent in the analysis that the experience of a few people could drive the results. Therefore, further refinements were made to remove categories from categorical variables that had five or fewer participants, which reduced our sample from 446 to 375 participants.

**Table 1. Number of MAX-NHANES Data Records, by Analytical Step**

| Analysis Population | Number of Records |
|---|---|
| 1999–2004 NHANES participants | 31,126 |
| 1999–2004 linkage-eligible NHANES participants | 25,750 |
| Total records in which NHANES and MAX data matched in the same year | 7,915 |
| **Number of records after imposing each NHANES restriction** | |
| Age 20 or older | 2,180 |
| Not pregnant | 1,909 |
| Has BMI | 1,666 |
| **Number of records after imposing each MAX restriction** | |
| Not SCHIP only | 1,663 |
| Full-benefit | 1,473 |
| Not dual | 894 |
| Adult or disabled | 856 |
| Never enrolled in comprehensive managed care | 455 |
| One MAX record per NHANES participant | 446 |
| Not underweight | 433 |
| Not in "Other" race/ethnic group | 419 |
| Not in state with 5 or fewer records | 375 |

Source: MAX-NHANES 1999–2004 data.
Note: Data restrictions accumulate from the top of the table to the bottom. The number of records shown in a row includes all data restrictions shown on the row and above.

## Modeling Approach

Having defined the sample, we proceeded to model the effects of obesity status on total FFS costs using a multiple regression framework. The dependent variable in the model was the natural log of total FFS costs, and we used various alternative specifications of obesity status as the key independent variable in the model. We controlled for age, age squared, race/ethnic group, sex, education, smoking status, and state using variables from NHANES. See Table 2 for summary statistics for the key variables in the model.

We originally entered BOE group as a control variable, but because interactions with the "disabled" BOE were significant for numerous variables, we ran the models separately for the "adult" and "disabled" categories. The decision to use separate models for the "adult" and "disabled" populations resulted in even smaller analysis populations. If an analytical category could not be combined reasonably with another category and the category had five or fewer participants in either the disabled or the adult analysis population, the analytical category, and thereby the participants in that category, were removed from the analysis (as discussed below).

Because Medicaid costs are highly skewed, we used the natural log of the total FFS costs as the outcome variable. Dollar values from survey years 1999–2003 were transformed to 2004 dollars by multiplying the total dollar value from the survey year by the 2004 average consumer price index (CPI) divided by the average CPI of the survey year (U.S. Department of Labor 2012). Because the log of zero is undefined, and some participants had no FFS expenditures, we added one dollar to all costs; this transformation made it possible to treat the log value of no FFS expenditures as equal to zero.

The weight status of participants was classified using the body mass index variable provided in NHANES. Body mass index is calculated as weight in kilograms divided by height in meters squared and rounded to the nearest tenth. Following current recommendations, underweight was defined as a BMI of less than 18.5, normal weight as a BMI of 18.5 to 24.9, overweight as a BMI of 25.0 to 29.9, and obesity as a BMI of 30.0 or higher (CDC 2013; Flegal et al. 2012; "Clinical Guidelines" 1998). Obesity was further subdivided into class 1 (BMI of 30.0 to 34.9), class 2 (BMI of 35.0 to 39.9), and class 3 (BMI of 40 or higher). Class 3 obesity is also referred to as extreme obesity. Because there were too few respondents who were underweight, they were removed from the analysis.

For the analyses, we tested four characterizations of obesity status. The first model employed the traditional definition of weight status using three categories: normal weight, overweight, and obese. The second model used a dichotomous variable to indicate whether a participant was or was not obese. The third model used four categories to focus on levels of obesity, theorizing that higher BMIs among obese people might lead to increased medical costs: not obese, obese class 1, obese class 2, and obese class 3. The fourth model used all of the weight categories: normal weight, overweight, obese class 1, obese class 2, and obese class 3.

Given that obesity status differed significantly among adults by race/ethnic group during 1999–2004, we generated models with an interaction between race and obesity status for each characterization of obesity status (Ogden et al. 2006). However, because certain combinations of weight status and race/ethnic group had five or fewer participants, we did not include the interaction term in our analyses.

**Table 2. Characteristics of the MAX-NHANES Obesity-FFS Cost Analysis Population**

| | Adult | | Disabled | |
|---|---|---|---|---|
| | Total (n=213) | Percent | Total (n=162) | Percent |
| **Sex** | | | | |
| Male | 62 | 29.1 | 67 | 41.4 |
| Female | 151 | 70.9 | 95 | 58.6 |
| **Race/Ethnic Group** | | | | |
| Non-Hispanic White | 64 | 30.0 | 53 | 32.7 |
| Non-Hispanic Black | 86 | 40.4 | 57 | 35.2 |
| Hispanic | 63 | 29.6 | 52 | 32.1 |
| **Education** | | | | |
| Less than HS | 85 | 39.9 | 104 | 64.2 |
| HS Graduate | 69 | 32.4 | 33 | 20.4 |
| Some College or College Graduate | 59 | 27.7 | 25 | 15.4 |
| **Smoking Status** | | | | |
| Current Smoker | 81 | 38.0 | 64 | 39.5 |
| Former/Never Smoker | 132 | 62.0 | 98 | 60.5 |
| **Weight Status** | | | | |
| Normal Weight | 53 | 24.9 | 40 | 24.7 |
| Overweight | 59 | 27.7 | 48 | 29.6 |
| Obese | 101 | 47.4 | 74 | 45.7 |
| **Obesity Class** | | | | |
| Not Obese | 112 | 52.6 | 88 | 54.3 |
| Obese Class I | 59 | 27.7 | 32 | 19.8 |
| Obese Class II | 22 | 10.3 | 20 | 12.3 |
| Obese Class III | 20 | 9.4 | 22 | 13.6 |

Source: MAX-NHANES 1999–2004 data.

The age at exam is calculated in months and was therefore treated as a continuous variable. Respondents were classified as current or former/never smokers. Race/ethnic group was grouped in three categories: (1) non-Hispanic white; (2) non-Hispanic black; (3) Hispanic (defined as Mexican American or other Hispanic). There were too few participants with a race/ethnic group of "other", which includes multiracial, to include them in the analysis. Education was grouped in three categories: (1) less than a high school degree; (2) high school graduate; (3) some college or college graduate. The one respondent who did not report an education level was assigned the most common education level among the study population, "less than a high school degree."

The main driver in total FFS costs is the benefit package provided by the state. Because Medicaid is a state-driven program, we decided that we could not create a model without controlling for NHANES participants' state of residence. However, the state in which the NHANES participant was examined is such a restricted variable that NCHS was hesitant to share the variable, even at the RDC. We arrived at a compromise by asking NCHS to provide a state variable with a value of 1 through 51 for the 50 states plus the District of Columbia (hereafter referred to collectively as "states"), but NCHS did not have to identify which state belonged to each of the 51 values. With the state variable, we could control for but not interpret the data by state, because it was not clear which value corresponded to which state. The disadvantage of including "state" in the model is that controlling for the 51 possible states reduces the power of the analysis model. However, "state" was so strongly significant in every model, we could not justify leaving it out of the analysis model. We removed the states with five or fewer participants (10 of the 24 states in the disabled model, 11 of the 26 states in the adult model) from the analysis population, which improved the fit of the models slightly.

For the analysis, we used six-year weights (1999–2004) for participants who had received an exam (MEC weight). We adjusted the MEC weights using the WTADJUST procedure in STATA to account for the fact that not all NHANES respondents were eligible to be linked. We adjusted the weights using definitions of race/ethnic group (non-Hispanic black, Mexican American, and other) and age group (0–19 years, 20–44 years, 45–64 years, and 65 years and older) as recommended by NCHS. The adjusted weights were calculated for the entire NHANES population, not just the analysis subpopulation.[3]

The models were run using SUDAAN 10.0.1. The significance of variables was evaluated using the Satterthwaite-adjusted chi-square test, a moderately conservative test of significance. To assess the results of the adult model, the product of the regression coefficient and sample value were added together for each variable and the exponential function was used to transform the sum back to total FFS 2004 dollars for multiple hypothetical scenarios. Because one dollar was added to each value before the log transformation, one dollar was removed from the estimated total FFS annual costs for each value. After ranking the state coefficients from lowest to highest, we used the percentile function in Microsoft Office Excel 2007 to calculate the 20th, 50th, and 80th percentile values for the state coefficients.

## Results

None of the regression models showed a significant association between obesity and total FFS costs at the 0.05 level. The adult models fit the data better than the disabled models. The $R^2$ values were 0.28 for each of the disabled models, with the p-values of the obesity measure ranging from 0.69 to 0.99.[4] For adults, the $R^2$ values ranged from 0.41 to 0.43, with the p-values of the obesity measure ranging from 0.07 to 0.15, closer to the significance threshold of 0.05.

Obesity as a dichotomous variable (the second model) among the adult population was the characterization of obesity that was closest to being significant at the 0.05 level (p=0.07). To put the model into context, in Table 3 we show the estimated expected annual total FFS costs in 2004 dollars for the adult population based on hypothetical respondents, controlling for other variables in the model. For each scenario, the hypothetical values for the variables were plugged into the regression model to calculate the log annual total Medicaid FFS cost. The log cost was then transformed back to a dollar value by exponentiating the log value and subtracting the dollar that was added prior to the log conversion. The annual total FFS cost is lower for the non-obese person than the obese person in each hypothetical scenario.

In the first hypothetical example, the value of each variable was based on the sample distribution of the adult population. For example, because 70.9 percent of the adult study population was female, a value of 0.709 was multiplied by the coefficient for female and a value of 0.291 was multiplied by the coefficient for male. Based on the regression model, the annual total FFS cost in 2004 dollars for the average non-obese person is $95.02, while for the obese person, the cost is $225.72. Table 3 also shows the projected costs for two hypothetical respondents: (1) a 30-year old white female non-smoker with no high school degree and (2) a 40-year old black male smoker with a high school degree. The projected costs are shown for each hypothetical respondent based on which state he or she resides: (1) State A, the state with the regression coefficient in the 20th percentile of the state coefficients (controlling for characteristics

of the state's sample members), (2) State B, the state with the median regression coefficient, and (3) State C, the state with the regression coefficient in the 80th percentile. For each state, the annual total FFS costs are lower for the hypothetical male than the hypothetical female. For each hypothetical person, the state made a substantial difference in the projected annual FFS costs. Although the estimated obesity-cost relationship was not statistically significant at the 0.05 level, in every scenario, the estimated annual total FFS costs for the obese person were more than double the estimated costs for the non-obese person.

**Table 3. Estimated Annual FFS Costs for Adults by Obesity Status in 2004 dollars**

| | Not Obese | Obese |
|---|---|---|
| Average Person in Data Set[a] | $95.02 | $225.72 |
| 30 year-old white female non-smoker with no high school degree who resided in: | | |
| State A[b] | 32.80 | 78.81 |
| State B[b] | 247.45 | 585.65 |
| State C[b] | 568.22 | 1,343.06 |
| 40 year-old black male smoker with a high school degree who resided in: | | |
| State A[b] | 3.97 | 10.73 |
| State B[b] | 35.52 | 85.24 |
| State C[b] | 82.67 | 196.58 |

Source: MAX-NHANES 1999-2004 data.
Note: Regression controlled for sex, age, age-squared, education level, smoking status, weight class, and race/ethnic group. Estimated annual log costs were estimated by multiplying the coefficient times the value for each variable, summing the values to get the total natural log costs and exponentiating the log costs. The only difference between the non-obese and obese models is the value of the obesity variable (1 or 0). Because one dollar was added to the total costs (to prevent a value of zero from being entered in the natural logarithm model) before the log transformation was made, a dollar was subtracted after the value had been transformed back into the dollar value.
[a] In the average person model, the value for each variable was the percentage distribution of the study population. For example, because 70.9 percent of the adult study population was female, a value of 0.709 was multiplied by the coefficient for female and a value of 0.291 was multiplied by the coefficient for male.
[b] Of the 15 states in the model, after controlling for characteristics of the state's sample members, average costs were estimated for State A (20th percentile of the state coefficients), State B (50th percentile), and State C (80th percentile).

## Discussion

Both NHANES and MAX are considered the gold standard for data within their fields of research. MAX does not rely on self-reported health care costs or Medicaid status. NHANES does not rely on self-reported height and weight data to determine obesity status. Therefore, it was with great anticipation that we began our research using the newly merged MAX-NHANES data set. The results of our research were disappointing and demonstrate the potential hazards of using a national data set (NHANES) with a state-based data system (MAX). The combination of a small sample size, wide variation in costs among Medicaid enrollees, and the necessity of controlling for state variation yielded an unstable model with imprecisely estimated relationships, leading to results that most researchers would be hesitant to use.

The quality of NHANES's body measurement data comes at the cost of a relatively small sample size. Because participants are both interviewed and given a physical exam, the annual sample size is small compared to that of surveys that carry out interviews only. The expected annual sample size for NHANES is approximately 5,000 individuals, compared to (for example) 75,000 to 100,000 persons for the National Health Interview Survey (NHIS) (NCHS 2012a, 2012b). While the NHANES sample size is adequate for many purposes, it can be limiting for analyses that focus on subgroups of the original sample or rely on matching to another data set.

One of the strengths of the MAX data is the plethora of data points that enables analysis of large numbers of Medicaid enrollees, often across multiple years of data. However, this strength is lost when the data are merged with NHANES. This is because the cross-sectional design of the NHANES sample implies that, in a design like ours that required contemporaneous matching of the measurement of obesity (based on NHANES) and Medicaid costs (based on MAX), the merged MAX-NHANES data had to be limited to no more than one analysis point per NHANES participant.

Our use of the merged MAX-NHANES data to do a cost analysis also affected the size of our data set because this analysis involved restricting the population (to nondual, full-benefit enrollees not enrolled in CMC) in order to make the costs consistent across the analysis population. If we had been looking at total costs and not looking only at costs that could be affected by obesity status, we might have been able to leave in the capitation payments associated with the CMC data. While restrictions of this type are typical of MAX analyses, they had the effect of cutting our data set to nearly a quarter of its size. In MAX analyses, where there are millions of records, a reduction of this magnitude can still yield a sizable data set for analysis. But its effect on the MAX-NHANES analysis population was to leave fewer than 500 records. Differences in the costs of the disabled and adult populations made it necessary to run the regression models separately for each population, which produced even smaller sample sizes for each analysis.

The fit of the models was better and the obesity measures were closer to statistical significance among the adult population than the disabled population. This is likely due to innate differences

in the different study populations' morbidity and the related medical costs. In the disabled analysis population, only 19 of the 196 participants (9.7 percent) had no FFS costs. The annual FFS costs among the remaining disabled participants ranged from $33 to $173,870. Among the adult analysis population, the percentage of participants with no FFS costs was higher (25.2 percent) and the range of total FFS costs among the remaining participants was smaller ($8 to $42,728). The combination of a small sample and wide range of potential annual total FFS costs for Medicaid recipients made it possible for one or two respondents to drive the analysis results (data not shown). To reduce the impact of a few participants, the analysis categories were redefined and participants were removed who would have fallen into a category with five or fewer people. This further reduced our sample size, which made it difficult to capture an impact of obesity on total FFS costs.

The wide range of potential costs resulted in an estimated annual FFS cost of the average participant that seems low. However, the estimated total FFS cost value is affected by the independent variables, particularly the respondent's state, so greatly that the typically reported value of the average participant may not be a useful measure. The variation in state costs makes it difficult to assess the cost of the average participant. In every model, the state in which the NHANES participant was surveyed was significant. Because the Medicaid benefit packages can differ by state, it would be theoretically incorrect to run an analysis without controlling for the state of the resident. However, given the small sample size for our analysis, controlling for all 51 states contributed to the unstable model. Originally, the adult analysis population included NHANES participants from only 26 states. After removing participants from states with five or fewer respondents, there were 15 states in the adult population. The lack of representation for some states is not surprising given that NHANES visits 15 counties per year and may not visit every state over a six-year period (NCHS 2012a).

The willingness of federal agencies to merge data sets is an exciting prospect for researchers. However, the merged data sets may not be a viable resource for all types of research. It is difficult to use the MAX-NHANES data to draw conclusions about the national Medicaid population. Researchers should be aware of how their research decisions will affect the sample size and thereby, the ability to detect significant findings. For example, an analysis that focuses on children would yield a larger starting sample than an analysis on adults. A study that focuses on enrollment or does not look at cost association may be able to leave CMC enrollees in the analysis. An analysis that focuses on a subset of states with similar programs may be able to control for program characteristics, which might reduce the effect of state. In addition, future analyses using the MAX-NHANES data may choose to focus on verifying measurements in one data set against the self-reported data in another, which would be useful for future research projects that use either MAX or NHANES data. With thoughtful consideration, the MAX-NHANES data have the potential to provide valuable insights for researchers and policymakers.

## References


Borck, Rosemary, Allison Hedley Dodd, Ashley Zlatinov, Shinu Verghese, Rosalie Malsberger, and Cara Petroski. "The Medicaid Analytic eXtract 2008 Chartbook." Washington, DC: Centers for Medicare & Medicaid Services, February 2012.

Centers for Disease Control and Prevention (CDC). "Childhood Obesity Facts." U.S. Department of Health and Human Services. Available at [http://www.cdc.gov/healthyyouth/obesity/facts.htm]. Accessed November 14, 2012.

Centers for Disease Control and Prevention (CDC). "Defining Overweight and Obesity." U.S. Department of Health and Human Services. Available at [http://www.cdc.gov/obesity/adult/defining.html]. Accessed January 9, 2013.

"Clinical Guidelines on the Identification, Evaluation, and Treatment of Overweight and Obesity in Adults: Executive Summary: Expert Panel on the Identification, Evaluation, and Treatment of Overweight in Adults." *American Journal of Clinical Nutrition*, vol. 68, no. 4, 1998, pp. 899–917.

Finkelstein, E. A., I. C. Fiebelkorn, and G. Wang. "National Medical Spending Attributable to Overweight and Obesity: How Much, and Who's Paying?" *Health Affairs*, January–June 2003, Supplement Web Exclusives, W3, 219–26.

Finkelstein, E. A., J. G. Trogdon, J. W. Cohen, and W. Dietz. "Annual Medical Spending Attributable to Obesity: Payer- and Service-Specific Estimates." *Health Affairs*, vol. 28, no. 5, 2009, pp. w822–w831.

Flegal, K. M., M. D. Carroll, C. L. Ogden, and L. R. Curtin. "Prevalence and Trends in Obesity Among U.S. Adults, 1999–2010." *JAMA: The Journal of the American Medical Association*, vol. 307, no. 5, 2012, pp. 491–497.

Gorber S. C., M. Tremblay, D. Moher, and B. Gorber. "A Comparison of Direct vs. Self-Report Measures for Assessing Height, Weight and Body Mass Index: A Systematic Review." *Obesity Reviews*, vol. 8, no. 4, 2007, pp. 307-326.

National Center for Health Statistics (NCHS). "National Health and Nutrition Examination Survey 2011–2012, Overview." U.S. Department of Health and Human Services, Centers for Disease Control and Prevention. Available at [http://www.cdc.gov/nchs/data/nhanes/nhanes_11_12/2011-12_overview_brochure.pdf]. Accessed October 22, 2012a.

National Center for Health Statistics (NCHS). "National Health Interview Survey (NHIS) Brochure." U.S. Department of Health and Human Services, Centers for Disease Control and Prevention. Available at [http://www.cdc.gov/nchs/data/nhis/brochure2010January.pdf]. Accessed October 22, 2012b.

National Institutes of Health (NIH). "The Practical Guide: Identification, Evaluation, and Treatment of Overweight and Obesity in Adults." October 2000. NIH Publication Number 00-4084.

Ogden, C. L., M. D. Carroll, L. R. Curtin, M. A. McDowell, C. J. Tabak, and K. M. Flegal. "Prevalence of Overweight and Obesity in the United States, 1999–2004." *JAMA: The Journal of the American Medical Association*, vol. 295, no. 13, 2006, pp. 1549–1555.

Simon, A. E., A. K. Driscoll, C. Golden, R. Tandon, C. R. Duran, E. A. Miller, K. C. Schoendorf, and J. D. Parker. "Documentation and Analytic Guidelines for NCHS Surveys Linked to Medicaid Analytic eXtract (MAX) Files." Hyattsville, MD: National Center for Health Statistics, 2011. Available at [http://www.cdc.gov/nchs/data/datalinkage/documentation_and_analytic_guidelines_nchs_survey_max_linked_data.pdf]. Accessed December 12, 2012.

Sturm, R. "The Effects of Obesity, Smoking, and Drinking on Medical Problems and Costs." *Health Affairs*, March-April 2002, vol. 21, no. 2, pp. 245–253.

Trogdon, J. G., E. A. Finkelstein, C. W. Feagan, and J. W. Cohen. "State- and Payer-Specific Estimates of Annual Medical Expenditures Attributable to Obesity." *Obesity*, vol. 20, no. 1, 2012, pp. 214–220.

U.S. Department of Labor, Bureau of Labor Statistics. "Consumer and Produce Price Indices." Available at [http://www.cdrpc.org/CPI_PPI.html]. Accessed October 15, 2012.

## Endnotes

[1] Not all MAX records contain SSNs because many states are unable to collect SSNs for all enrollees. Most of the records with missing SSNs are for children, people receiving only family planning services, or aliens eligible for emergency services only (Simon et al. 2011).

[2] Restricted-benefit enrollees include (1) aliens eligible for emergency services only, (2) duals receiving coverage for Medicare premiums and cost sharing only, and (3) people receiving only family planning services (Borck et al. 2012).

[3] The entire NHANES population includes participants who did not complete a MEC exam. Therefore, their MEC weight is zero. WTADJUST does not include participants with a weight of zero in its calculations. The adjusted weight for these participants was set to zero.

[4] An $R^2$ value of 0.28 indicates that the variables in the model explained 28 percent of the variation in the log annual total FFS costs. We classify a p-value of 0.05 or below as statistically significant, which means that the probability of finding a relationship between the variables due to chance is less than or equal to five percent.

**For further information on this issue brief series, visit our website at www.mathematica-mpr.com**

Princeton, NJ • Ann Arbor, MI • Cambridge, MA • Chicago, IL • Oakland, CA • Washington, DC