

Use of Econometric Models to Estimate Disease-Specific Shares of Medical Expenditures\*

Justin G. Trogdon<sup>a</sup>, Eric A. Finkelstein<sup>a</sup>, Thomas J. Hoerger<sup>a</sup>

<sup>a</sup> RTI International

3040 Cornwallis Rd.

P.O. Box 12194

Research Triangle Park, NC 27709-2194, USA

E-mail addresses: jtrogdon@rti.org (JG Trogdon), finkelse@rti.org (EA Finkelstein), tjh@rti.org (TJ Hoerger).

\* This research was supported by Grant Number 1 P30 CD000138-01 from the Centers for Disease Control and Prevention to the RTI-UNC Center of Excellence in Health Promotion Economics.

Send all correspondence to:

Justin Trogdon  
RTI International  
3040 Cornwallis Road  
P.O. Box 12194  
Research Triangle Park, NC 27709 USA  
Voice: 00-1-919-541-6893  
Fax: 00-1-919-541-6683  
E-mail: jtrogdon@rti.org

**Objective.** To investigate the use of regression models to calculate disease-specific shares of medical expenditures.

**Data Sources/Study Setting.** Medical Expenditure Panel Survey (MEPS), 2000–2003

**Study Design.** The study illustrates the properties of regression models as used to estimate disease-specific shares of medical expenditures, provides an empirical example using linear and two-part regression models of prescription expenditures, and makes recommendations for alternative counterfactual simulations.

**Data Collection/Extraction Methods.** The study pooled MEPS annual files for 2000–2003. Condition files were used to define the self-reported presence of 10 medical conditions.

**Principal Findings.** Incremental effects of conditions on expenditures, expressed as a fraction of total expenditures, can only be interpreted as shares if (1) the true underlying expenditure generating function is additively separable or (2) the occurrence of conditions is mutually exclusive. When the expenditures for those with multiple conditions are greater than the sum of the expenditures associated with each condition (i.e., when the presence of one condition increases treatment costs for another condition), summing condition-specific shares leads to double-counting of expenditures and overestimation of the share of expenditures associated with the set of conditions.

**Conclusions.** Condition-specific shares generated from multiplicative models, such as ordinary least squares on log expenditures, nonlinear least squares, and generalized linear models, should not be summed. We provide an algorithm that addresses this issue and allows estimates based on these models to be interpreted as shares and summed across conditions.

**Keywords:** health expenditures, cost of illness, expenditure share, attributable fraction

## INTRODUCTION

In the health policy arena, there is often interest in estimating the financial contribution of particular conditions or sets of conditions to overall health care spending. Such cost-of-illness (COI) studies are widely used by organizations such as the National Institutes of Health and the World Health Organization to quantify burden, both prevalence and cost, and prioritize funding (Byford, Torgerson, and Raftery 2000; Bloom et al. 2001). COI estimates provide information describing present spending patterns and resource allocation, highlighting possible areas of cost savings through targeted prevention efforts or research into new treatments.

The usefulness of COI studies is widely debated. Economists often point out that the total cost estimates presented in COI studies do not account for the benefits gained from these expenditures and therefore do not directly imply the existence of inefficiencies in the market (Shiell, Gerard, and Donaldson 1987; Roux and Donaldson 2004). Moreover, a COI study alone should not drive a resource allocation decision because simply showing that a cost burden or need exists does not imply that the burden can be reduced using available technology. Despite these shortcomings, COI studies can be viewed as an informative first step in policy decisions (Corso, Grosse, and Finkelstein 2004) and/or as one of many inputs into a complicated decision-making process (Hodgson 1989; Tarricone 2006). Given the appeal of COI estimates to policy makers, it is important to provide the best estimates possible.

Econometric methods are being used increasingly to estimate medical expenditures associated with conditions while controlling for other conditions and determinants of expenditures (Akobundu et al. 2006). An extensive literature has developed alternatives to ordinary least squares (OLS) to account for the special properties of expenditure data (see Jones 2000, for a review; Manning, Basu, and Mullahy 2005; Cantoni and Ronchetti 2006). Related

literature provides specification tests to choose among the many available estimators (Manning and Mullahy 2001; Buntin and Zaslavsky 2004; Basu, Arondeka, and Rathouz 2006). The most commonly used and recommended models are multiplicative in levels of expenditures. These models include OLS on log (positive) expenditures, nonlinear least squares (Mullahy 1998), and generalized linear models (GLM).

Examples of COI studies using econometric methods include estimates of the cost of smoking (Miller, Ernst, and Collin 1999; Coller, Harrison, and McInnes 2002; Max et al. 2004), obesity (Finkelstein, Fiebelkorn, and Wang 2003, 2004), and injuries (Finkelstein et al. 2005). The costs associated with these conditions have been expressed as a share of overall expenditures. For example, smoking has been estimated to account for roughly 7% of medical expenditures (see Miller, Ernst, and Collin 1999, and papers cited therein).

This study addresses the issues involved in using commonly applied regression models of medical expenditures to calculate shares of expenditures. When applied individually for a set of conditions, the separate condition-specific expenditure estimates, each expressed as a fraction of total expenditures, can add up to more than 100%. If the fractions are interpreted as shares of total expenditures, or slices of a total expenditure pie, this result is counterintuitive.

We show that the fractions can only be interpreted as shares if (1) the true underlying expenditure generating function is additively separable or (2) the occurrence of conditions is mutually exclusive (i.e., only one condition can be present per person in a given time period). The implication for estimating expenditure shares is that, when expenditures associated with the joint occurrence of conditions are greater than the sum of the condition-specific expenditures (e.g., if expenditure functions are multiplicative across conditions), expenditures calculated

separately for each condition double-count the contributions to expenditures of the joint occurrences. This result holds even if the condition prevalence rates are independent.

We present several recommendations for researchers interested in using econometric methods to estimate shares of expenditures for medical conditions or risk factors. We demonstrate our recommendations using numerical examples and an empirical example with data on prescription expenditures in the Medical Expenditure Panel Survey (MEPS). The recommendations are straightforward and can be implemented in a variety of circumstances.

### **ATTRIBUTABLE FRACTIONS: DEFINITIONS AND PROPERTIES**

We, as well as others in the literature (e.g., Miller, Ernst, and Collin 1999), borrow the concept of attributable fractions (AFs) from the epidemiologic literature. In the epidemiologic literature, AFs are used to measure the proportion of disease risk in a population that can be attributed to a risk factor or set of risk factors (Rockhill, Newman, and Weinberg 1998). Here the term refers to the proportion of medical expenditures that can be attributed to a condition or set of conditions.

Studies that use regression analysis to calculate the AF for a particular condition define AF using incremental effects. The numerator of the AF predicts expenditures for each person and subtracts from that predicted expenditures, setting the condition of interest to zero and leaving all other covariates and conditions as they are in the data. This incremental effect gives the amount of expenditures per person attributed to the condition. This is then summed over all the persons in the sample to show total condition-attributable expenditures and divided by total predicted expenditures for the entire sample to express condition-specific expenditures as a fraction of total expenditures. Other studies perform similar counterfactuals without explicitly calling it an AF;

for convenience, we employ the AF terminology for the rest of the paper to indicate condition-specific expenditures.

Predicted expenditures come from the analyst's empirical model for conditional expenditures. Choice of an empirical model for mean expenditures will directly affect the properties of the AF. The following examples describe the calculation of expenditure shares for a variety of empirical models and illustrate why the sum of condition-specific AFs can be greater than 100% in some models (for discussion of this property for epidemiologic formulations of AFs, see Rothman and Greenland 1998; Rowe, Powell, and Flanders 2004). For ease of exposition, assume expenditures are determined by two medical conditions and an intercept (to capture other expenditures when no conditions are present); no other covariates enter the models. All condition indicators are defined such that the presence of the condition leads to increased expenditures.

### *Calculating Attributable Fractions*

We begin with a numerical example of the double-counting problem in regression-based AFs (see the Standard AF columns in Table 1). There are four types of people in this model: those without any condition ("Neither condition"), those with one condition or the other ("Condition 1 only" and "Condition 2 only"), and those with both conditions ("Both conditions"). We assume that the probability of each condition is 50% and that the conditions are independent. There are 100 persons of each type.

[Insert Table 1 here.]

The second column lists predicted expenditures per person for each type of person. The key feature of this example is that the spending associated with both conditions is greater than the sum of the spending associated with each of the conditions; hence, the assumptions of

additive separability and no joint occurrence of conditions are violated. The third column shows the expenditures associated with each type of person. Expenditures for each type are found by multiplying the predicted expenditures per person by the number of people of each type observed in the sample, given in parentheses. Total predicted expenditures for the observed sample is given by summing the predicted expenditures for each type. In the example, total predicted expenditures are \$3,600 (i.e.,  $100 \times \$1 + 100 \times \$5 + 100 \times \$5 + 100 \times \$25$ ).

The fourth column shows the counterfactual where the first condition is removed. This effectively reclassifies any person who only had condition 1 as a person without either condition and any person who had both conditions as a person with only condition 2. Predicted expenditures in this counterfactual are \$1,200 (i.e.,  $200 \times \$1 + 200 \times \$5$ ). The attributable expenditures associated with condition 1 are \$2,400 ( $\$3,600 - \$1,200$ ), yielding an AF of 67%. If condition 2 is removed instead of condition 1, the AF for condition 2 is also 67%.

Clearly, the AFs in this example cannot be interpreted as shares of total expenditures due to each condition; the sum of these AFs is 133% of total expenditures. Furthermore, if both conditions were removed, spending would still be \$400. Adding this component of expenditures (11%) to the sum of the condition-specific AFs results in 144% of predicted total expenditures.

The problem in interpreting AFs as shares is that both counterfactuals take credit for a large part of the reduction in expenditures for those that are observed to have both conditions. The double-counting does not rely on positive correlation in prevalence rates; the prevalence rates in Table 1 are independent. The conditions only need to occur jointly.

The following sections show how the underlying model (data generation process) affects the properties of regression-based AFs, with particular focus on model assumptions concerning

interaction and multiplicative effects. We assume that the empirical model matches the true underlying model and there is no misspecification.

### ***Linear Model***

Consider the simple linear model:

$$y = \alpha_0 + \alpha_1 d_1 + \alpha_2 d_2 + \varepsilon, \quad E[\varepsilon | d_1, d_2] = 0, \quad (1)$$

where  $y$  is expenditures and  $d_1$  and  $d_2$  are indicators for the presence of medical conditions of interest. Using the definition above, the OLS AF estimate for the first condition is the incremental effect of condition 1 summed over the sample ( $\alpha_1 p_1$ ) divided by predicted total expenditures ( $\alpha_0 + \alpha_1 p_1 + \alpha_2 p_2$ ), where  $p_1$  and  $p_2$  are the unconditional probabilities of having conditions 1 and 2, respectively, and are estimated by the fraction of the sample with each condition. The sum of the AFs is estimated by  $(\alpha_1 p_1 + \alpha_2 p_2) / (\alpha_0 + \alpha_1 p_1 + \alpha_2 p_2)$ .

There are two features of the linear model that must be satisfied for the AFs to represent shares that can be summed across conditions. First,  $\alpha_0$  must be non-negative. In our simple model, the intercept represents mean expenditures for persons without any condition, which are greater than or equal to zero. Adding a set of covariates,  $X$ , to the model with covariates  $\beta$  will only add  $X \times \beta$  to the denominator of the AFs. The intuition for the properties of the AFs remains the same; as long as mean predicted expenditures for those without any condition are non-negative, the sum of the AFs will be less than or equal to 100%.

Second, there must not be any interaction effects between conditions in the model. In other words, the marginal effect of each of the conditions on expenditures must not depend on other conditions. If this is not the case, double-counting occurs when condition-specific AFs are added, as the following example shows.

### ***Linear Model with Interactions***

Extending the linear model to include interaction effects adds  $\alpha_3 d_1 d_2$  to (Equation 1). In this model, the marginal effect of condition 1 depends on condition 2 ( $\alpha_1 + \alpha_3 d_2$ ) and vice versa. For example, with  $\alpha_1, \alpha_3 > 0$ , having diabetes increases expenditures ( $\alpha_1$ ) for everyone and increases expenditures by even more ( $\alpha_1 + \alpha_3$ ) for those with heart disease. The counterfactual that removes condition 1 alone will include  $\alpha_3$ , as will the counterfactual that removes condition 2 alone. The contribution to expenditures of the joint presence of the conditions will be attributed separately to both conditions in the condition-specific AFs. To appreciate this, the numerical example in Table 1 could be generated by this empirical model with  $\alpha_0 = 1$ ,  $\alpha_1 = \alpha_2 = 4$ , and  $\alpha_3 = 16$ . If economies of scope exist such that it is cheaper to treat two conditions together than separately (i.e.,  $\alpha_3 < 0$ ), the sum of the AFs will undercount expenditures. The recommendations below will still generate AFs that sum to 100% and can be interpreted as shares.

### ***Multiplicative Models with Exponential Means***

Medical expenditure data take on non-negative values, often contain many zeros, and are heavily skewed. An extensive literature has developed alternatives to OLS to account for the properties of expenditure data (see Jones 2000, for a review; Manning, Basu, and Mullahy 2005; Cantoni and Ronchetti 2006). In the majority of these models, positive expenditures have an exponential conditional mean,  $E(y | y > 0, X) = c \times \exp(X\beta)$ , where  $c$  is a scale factor. These models include OLS on log (positive) expenditures, nonlinear least squares (Mullahy 1998), and GLM with a log-link function. Because these models are multiplicative, the incremental effects (in levels) for each condition depend on other conditions. Therefore, summing condition-specific AFs will double-count attributable expenditures and overestimate the reduction in expenditures associated with the set of conditions.

Consider the following simple model of expenditures:

$$E(y | y > 0, d_1, d_2) = c \times \exp(\beta_0 + \beta_1 d_1 + \beta_2 d_2) = c \times \exp(\beta_0) \exp(\beta_1 d_1) \exp(\beta_2 d_2) \quad (2)$$

In a model with log-normally distributed expenditures,  $c = \exp[(1/2)v]$ , where  $v$  is the variance of the log-scale error term. In GLM with log-link and gamma-variance function,  $c$  is the shape parameter. The numerical example in Table 1 is consistent with this model with  $c = 1$ ,  $\beta_0 = 0$ , and  $\beta_1 = \beta_2 = \ln(5)$ .

Let  $p_1$  be the probability of condition 1 and  $p_2$  be the probability of condition 2 in the population; the conditions are assigned independently. With algebraic simplification, the population AF for the first condition can be shown to be

$$AF_1 = p_1 [\exp(\beta_1) - 1] / (1 + p_1 [\exp(\beta_1) - 1]). \quad (3)$$

This formulation has a number of interesting features. First, Equation 3 has a similar form to one formulation for AF given in the epidemiologic literature (Rockhill, Newman, and Weinberg 1998), replacing the relative risk (RR) of disease with relative expenditures (i.e.,  $\exp(\beta_1)$ , the ratio of expected expenditures with and without condition 1).

Second, the AF for condition 1 is not a function of the prevalence or the incremental effects of other conditions. Regardless of the presence of other conditions, expenditures will be reduced by the same percentage when the condition of interest is removed.

Finally, as long as  $\beta_1$  is greater than or equal to zero,  $AF_1$  will be between zero and one. However, when we sum the AFs for both conditions, it can be shown that the sum will be greater than one if  $p_1 p_2 [\exp(\beta_1) - 1][\exp(\beta_2) - 1] > 1$ . The sum of the AFs will be more likely to be greater than one the higher the joint probability of the two conditions ( $p_1 p_2$ ) and the larger the slope of log expenditures associated with the conditions ( $\beta_1, \beta_2$ ). More importantly, even if the

sum of the AFs is not greater than 100%, summing the condition-specific AFs will overestimate the combined share of expenditures.

The formulation in Equation 3 rests on three assumptions: homoscedasticity, lack of other covariates, and independence of prevalence rates. Homoscedasticity allows the constant term and the variance in log-normal models and the shape parameter in gamma models to be factored out of the numerator and denominator and canceled, yielding (Equation 3). Generally, the variance term cannot be factored out in log-normal models in the presence of heteroscedasticity.

The example used to generate formula (Equation 3) did not include other covariates ( $X$ ). In nonlinear models, AFs that average the predictions for each person over the observed distribution of  $X$  will depend on the levels of other covariates in the sample. If the AF is instead constructed using predictions for a prototypical person (i.e., using mean, median, or mode values for  $X$ ), the terms involving  $X$  could be factored out of both the numerator and denominator, yielding Equation 3.

The assumption of independent prevalence rates allows the denominator to be a simple function of  $p_1$  and  $p_2$ , which can be factored and simplified into Equation 3. Independent prevalence rates rule out causal relationships among conditions, such as diabetes leading to heart disease (Fox et al. 2004). Without independence, in general, AFs will be a function of other conditions. The problem of double-counting will still be present. Causal relationships among conditions lead to positive correlation of prevalence rates. The greater the positive correlation among the conditions, the higher the frequency of observed joint conditions and the greater the problem. If the conditions are perfectly negatively correlated, the conditions will be mutually exclusive and there will be no observed joint conditions and no double-counting in the condition-

specific AFs. The latter property provides intuition for the complete cross-classification recommendation in the Recommendations section.

### *Summary*

The AF for a condition represents the percentage reduction in expenditures that would happen in the absence of the condition, all else constant. Generally, the AF does not measure the share of total expenditures accounted for by a particular condition. It is only appropriate to interpret AFs as shares and sum the condition-specific AFs if (1) the effects of conditions on expenditures (in levels) are completely separable or (2) there is no overlap in the occurrence of conditions. Conditions are completely separable if the presence of a condition adds a fixed dollar amount to expenditures (e.g., treatment for a minor injury). This will occur if there are no interaction effects among diseases. When the expenditures for those with multiple conditions are greater than the sum of the expenditures associated with each condition, summing condition-specific AFs leads to double-counting of expenditures and overestimation of the share of expenditures associated with the set of conditions. Condition-specific AFs generated from multiplicative models, where the presence of a condition increases expenditures by a fixed percentage, should not be summed. Importantly, this implies that many models commonly recommended for expenditure data, such as OLS on logged expenditures and GLM, prohibit interpreting the AFs as shares of total expenditures and summing condition-specific AFs to gauge the total effect of the set of conditions.

## **EMPIRICAL EXAMPLE OF TREATING ATTRIBUTABLE FRACTIONS AS SHARES**

### *Data*

MEPS contains nationally representative data on the medical expenditures of the noninstitutionalized U.S. civilian population. We combined the annual files from 2000 to 2003

(N = 125,052). The Medical Care Consumer Price Index was used to inflate expenditures to 2005 dollars. Sample weights were used to generate predictions representative of the noninstitutionalized civilian population.

Condition-specific data came from the MEPS Medical Conditions File, which uses self-reports and internally coded three-digit *International Classification of Diseases, Ninth Revision, Clinical Modifications* (ICD-9-CM) codes to identify conditions. We included 10 conditions expected to account for the majority of prescription expenditures: heart disease, hypertension, dyslipidemia, diabetes, asthma, depression, other mental health and substance abuse (MH/SA), HIV, arthritis, and skin disorders.

We focused on pharmaceutical expenditures, the expenditure category that most clearly exhibits the problem of treating AFs as shares and summing across conditions. The dependent variable of total pharmaceutical expenditures for each person was regressed against dummy variables for the presence of these medical conditions, controlling for age, age squared, gender, race (white non-Hispanic [omitted], black non-Hispanic, Hispanic, Asian, and other), geographic region (Midwest, South, West, and Northeast [omitted]), education (less than high school, high school, college [omitted], graduate degree, other degree, and missing), income level (below the federal poverty level [FPL], 100%–199% of FPL, 200%–399% of FPL [omitted], and 400% or greater than FPL), and insurance provider (Medicare, Medicaid, other public, private [omitted], and uninsured).

### ***Methods***

Pharmaceutical expenditures have a large fraction of observations at zero (39%) and are positively skewed. We therefore estimated a two-part model: a logit to predict the probability of positive expenditures and a second equation to model expenditures conditional on having

positive expenditures. The two-part model is nonlinear and multiplicative, implying interaction effects across conditions. Therefore, even though the analysis is somewhat more complicated because of the impact of the condition on the logit probability, the intuition for multiplicative models above remains the same in a two-part model. For the second part of the two-part model, we conducted the specification tests recommended by Manning and Mullahy (2001), the outcomes of which indicate heteroscedastic and kurtotic log-scale residuals. The kurtosis is statistically significant but mild, leading us to choose GLM to avoid retransformation of the predicted values. Box-Cox and Park specification tests indicate a log-link and a gamma variance function.

We also estimated standard OLS (without interaction effects) on levels of expenditures for comparison. The OLS model assumes that the impacts of the 10 conditions on expenditures are completely separable and additive. However, Hosmer-Lemeshow tests of model fit indicated that OLS does not fit the data as well as the two-part model.

## **RESULTS**

Most of the 10 conditions have high prevalence rates (Table 2, columns 2 and 3). Joint occurrence of conditions is common; among those with a condition, at least 45% have at least one other condition. The highest rate of joint conditions occurs for persons with diabetes, 82% of whom have at least one other condition. High rates of joint conditions indicate that double-counting could be a major problem when summing condition-specific AFs in a multiplicative model.

[Insert Table 2 here.]

Regression-based AFs combine prevalence and per-person expenditures (i.e., incremental effects). AFs are large if the prevalence is high, if the per-person expenditures are high, or if both

are true. For example, hypertension has relatively low prescription expenditures per person but a high prevalence rate, whereas diabetes has a lower prevalence rate but relatively high per-person prescription expenditures. All conditions are statistically significant at the 99% confidence level in OLS and both parts of the two-part model. Bootstrapped standard errors for per-person expenditures and AFs are available upon request.

OLS yields smaller AFs than the two-part model for each condition, and the rank order of the conditions based on AFs is the same for both specifications (Table 2, columns 7 and 8), although this would not necessarily hold in other applications. Under the strong assumption of linearity in the OLS model, the AFs can be interpreted as shares of total expenditures; the sum of AFs for these conditions is 74%. As noted, however, specification tests indicate OLS does not fit the data well. The AFs cannot be interpreted as shares in the two-part model; the sum of the AFs is 105%.

## **RECOMMENDATIONS**

For the reasons cited in the Introduction, policy makers are interested in understanding how medical expenditures are allocated across diseases. Given the above discussion, how can researchers use AF estimates to meet this goal without double-counting expenditures?

### ***Ordinary Least Squares***

As we have seen, under the (strong) assumption of additive separability in OLS, condition-specific AFs can be interpreted as shares of total expenditures and can be summed to estimate the combined share of all conditions of interest. However, OLS has a number of long-recognized, undesirable features when used to model medical expenditure data. Most importantly, if expenditures are not additively separable, OLS is a misspecification and will

result in biased estimates. Also, predicted expenditures can be negative. As shown above, this can lead to AFs not bounded by 100%. Finally, OLS can be sensitive to outliers.

### ***Sequential Attributable Fractions***

The second approach to avoid double-counting of expenditures and ensure that the sum of the AFs yields the total effect is to perform the counterfactuals sequentially and cumulatively, an approach recommended by Eide and Gefeller (1995) for epidemiologic AFs. For each condition counterfactual, subtract expenditures after removing the additional condition from predicted expenditures after removal of all previous conditions. Report the AFs as these reductions in expenditures relative to the initial predicted expenditures with observed conditions. The Sequential AF columns of Table 1 demonstrate this approach.

The order in which the conditions are removed is important for the sequential AFs. In our example, the conditions are completely symmetric, but if a condition is first to be removed, the AF is 67%; if the condition is second, the AF is 22%. The sequential AF assigns more of the expenditures associated with joint occurrence of the conditions to the condition that is removed earlier in the ordering. The standard approach is equivalent to choosing each condition to be the first removed (Rowe, Powell, and Flanders 2004). This attributes most of the joint expenditures to each condition, provides the largest possible AFs, and leads to double-counting of these joint expenditures when the condition-specific AFs are summed.

This is especially troubling because, for most studies, there will be a large number of potential orderings for removal of the conditions; the number of orderings is  $J!$ , where  $J$  is the total number of conditions. In the MEPS example, there are more than 3 million possible orderings for the 10 conditions. Eide and Gefeller (1995) suggest constructing upper and lower bounds on the AF by performing the calculation for each condition as if it were the first and last

condition removed. Removing a condition first yields the largest possible AF estimate, and removing a condition last yields the smallest possible AF. This design allows the researcher to present the range of expenditures attributable to each condition under different assumptions about how to divide expenditures associated with joint conditions.

This approach is demonstrated using the MEPS data in Table 2 (column 9). The lower-bound AFs are calculated by predicting expenditures with and without the condition, assuming all other conditions have been removed in the sample. The lower-bound AFs are much smaller than the standard, upper-bound, AFs. On average across the conditions, the AFs fall by more than 50%. For example, the AF for diabetes decreases from 14% to 5% when the other conditions in the model are removed first.

The total impact of the entire set of conditions can be calculated by removing all of the conditions at the same time in the counterfactual, leaving all else constant (Rowe, Powell, and Flanders 2004). The resulting AF accounts for the separate condition effects and the joint effects and will provide a better estimate of the total impact of the conditions of interest than simply summing the condition-specific (i.e., nonsequential) AFs.

The last two rows of Table 2 show an example of the combined effect calculations using the MEPS data. The results for the sum of the condition-specific AFs and the combined AF are the same for OLS. According to the two-part model, in the absence of all 10 conditions, expenditures would be lower by 70%. As expected, the combined AF is in between the sum of the upper- and lower-bound AFs.

The combined AF still should be interpreted carefully. In particular, the combined AF generally cannot be interpreted as a share of expenditures for the same reasons that condition-

specific AFs cannot: there could be interactions between the conditions of interest and other conditions used as covariates in the model.

### ***Complete Cross-Classification***

Another approach is to treat each condition and combination of conditions observed in the data as its own separate entity when calculating counterfactuals (i.e., complete cross-classification). When this is done, the AF regains the distributive property, ensuring that the AF for each unique combination of conditions is a share of the total spending and that these can be summed to get the total effect of the entire set of conditions (for a discussion of this property in epidemiologic AFs, see Rockhill, Newman, and Weinberg 1998). This is illustrated in the “Complete Classification” section of Table 1, where “Condition 1 only,” “Condition 2 only,” and “Both conditions” are treated as three separate conditions. Not only does this procedure avoid double-counting of expenditures, it makes explicit the share of expenditures associated with joint occurrence of conditions. However, the number of unique combinations of conditions that need to be considered is  $2^J$ . While technically feasible, interpreting the results for this many combinations is daunting.

One possible solution to this problem is to divide the expenditures attributable to the combinations of conditions back to the constituent conditions. In our numerical example, this would mean dividing the \$2,400 of expenditures associated with the joint occurrence of both conditions between conditions 1 and 2. One might consider splitting the joint expenditures evenly among constituent conditions or dividing into shares based on relative prevalence, relative impact (i.e., parameters), or both.

We recommend a procedure to allocate the joint expenditures that satisfies the following principles. First, under the assumption of independent prevalence rates, the division should not

be based on the relative prevalence rates of the conditions. The problem of double-counting is a problem of *observed* joint conditions. In order to observe joint occurrence, all conditions must be present and therefore contribute equally to the observation. Second, the condition with the greater impact on expenditures should get the higher share of the joint expenditures. Third, a condition with no associated excess expenditures should receive zero share. Fourth, the shares must add up to the total expenditures associated with the joint conditions.

The following formula can be used in exponential mean models to allocate the share(s) of expenditures associated with  $K$  joint conditions to condition  $k$ :

$$s_k = [\exp(\beta_k) - 1] / \sum_{k=1}^K [\exp(\beta_k) - 1]. \quad (4)$$

Subtracting 1 in the numerator ensures that if the condition has no discernible impact on expenditures (i.e.,  $\beta_k = 0$ ), it will receive zero share of the joint expenditures. The denominator ensures that the sum of the shares across all of the joint conditions is 1. In the example in Table 1,  $\beta_1 = \beta_2 = \ln(5)$  and  $s_1 = s_2 = 0.5$ .

The last column of Table 2 demonstrates the results of the procedure using the estimates from the MEPS data. We calculate the shares in Equation 4 using the coefficients from the second part of the two-part model for every observed unique combination of conditions (390 of the 1,024 possible combinations of conditions were observed in the data). The complete cross-classification AFs are in between the lower and upper estimates. Most importantly, the AFs add up to equal the combined effect of 70%. Double-counting in the sum has been removed.

One limitation of our weighting scheme is that it does not account for causality among conditions. If condition 1 (e.g., diabetes) causes condition 2 (e.g., heart disease) but is relatively cheap to treat compared with condition 2, our weighting scheme would still assign most of the expenditures associated with the joint conditions to condition 2. Complete classification can still

be used to estimate the expenditures associated with joint conditions and could be combined with alternative weighting schemes that incorporate the causal role of conditions that are also risk factors for other conditions.

## **DISCUSSION**

In general, the predicted reduction in expenditures that would occur in the absence of a condition, expressed as a fraction of total expenditures (AF), is not a share of total expenditures associated with each condition. When the presence of one condition affects the spending associated with other conditions, condition-specific AFs include expenditures associated with the joint occurrence of that condition and other conditions. If these AFs are summed, a portion of the expenditures associated with the joint occurrence of conditions will be double-counted, and the sum will not give the appropriate combined share of expenditures attributable to the set of conditions. Commonly used empirical models in health economics imply these types of nonconstant marginal effects, including OLS on log expenditures, nonlinear least squares, and GLM. Therefore, researchers must be careful when interpreting AFs, especially with multiple conditions of interest. AFs indicate the extent to which medical expenditures would be lower in the absence of particular conditions, all else constant.

For researchers interested in dividing existing expenditures into mutually exclusive categories of conditions, we recommend the following uses of AFs. First, report the upper and lower bounds for the AF for each condition. The upper bound can be calculated by holding all other conditions of interest fixed, as observed in the sample, when calculating the AF. The lower bound can be calculated by first removing all other conditions of interest in the sample when calculating the AF. The lower-bound AF does not attribute any of the expenditures associated with joint conditions to the condition, while the upper-bound AF attributes the largest share of

expenditures associated with joint conditions to the condition. Second, report the AF calculated by removing all conditions of interest at once. This is preferable to simply summing the condition-specific AFs.

Third, calculate an AF for each unique combination of conditions. Then use a weighting scheme such as Equation 4 to assign the attributable expenditures associated with joint conditions to the constituent conditions. The resulting AFs for each condition of interest will include both the direct impact on expenditures and the condition's portion of expenditures when observed with other conditions, but will avoid double counting. These AFs can be interpreted as shares of total expenditures and can be summed to get the total impact of the set of conditions.

## REFERENCES

- Akobundu, E., J. Ju, L. Blatt, and C. D. Mullins. 2006. "Cost-of-Illness Studies: A Review of Current Methods." *Pharmacoeconomics* 24 (9): 869-90.
- Basu, A., B. V. Arondeka, and P. J. Rathouz. 2006. "Scale of Interest Versus Scale of Estimation: Comparing Alternative Estimators for the Incremental Costs of a Comorbidity." *Health Economics* 15: 1091-107.
- Bloom, B. S., D. J. Bruno, D. Y. Maman, and R. Jayadevappa. 2001. "Usefulness of US Cost-of-Illness Studies in Healthcare Decision Making." *Pharmacoeconomics* 19 (2): 207-13.
- Buntin, M. B., and A. M. Zaslavsky. 2004. "Too Much Ado about Two-Part Models and Transformation? Comparing Methods of Modeling Medicare Expenditures." *Journal of Health Economics* 23: 525-42.
- Byford, S., D. J. Torgerson, and J. Raftery. 2000. "Economic Note: Cost of Illness Studies." *British Medical Journal* 320 (7245): 1335.
- Cantoni, E., and E. Ronchetti. 2006. "A Robust Approach for Skewed and Heavy-Tailed Outcomes in the Analysis of Health Care Expenditures." *Journal of Health Economics* 25 (2): 198-213.
- Coller, M., G. W. Harrison, and M. M. McInnes. 2002. "Evaluating the Tobacco Settlement Damage Awards: Too Much or Not Enough?" *American Journal of Public Health* 92 (6): 984-9.
- Corso, P., S. Grosse, and E. Finkelstein. 2004. "The Skinny on COI Analysis." *Obesity Research* 12 (7): 1189.
- Eide, G. E., and O. Gefeller. 1995. "Sequential and Average Attributable Fractions as Aids in the Selection of Preventive Strategies." *Journal of Clinical Epidemiology* 48: 645-55.

- Finkelstein, E. A., H. Chen, T. R. Miller, P. S. Corso, and J. A. Stevens. 2005. "A Comparison of the Case-Control and Case-Crossover Designs for Estimating Medical Costs of Non-Fatal Fall-Related Injuries among Older Americans." *Medical Care* 43 (11): 1087-91.
- Finkelstein, E. A., I. C. Fiebelkorn, and G. Wang. 2003. "National Medical Expenditures Attributable to Overweight and Obesity: How Much and Who's Paying?" *Health Affairs* (Web Exclusive): W3-219-W3-226.
- Finkelstein, E. A., I. C. Fiebelkorn, and G. Wang. 2004. "State-Level Estimates of Annual Medical Expenditures Attributable to Obesity." *Obesity Research* 12 (1): 18-24.
- Fox, C. S., S. C. Coady, P. D. Sorlie, D. Levy, J. B. Meigs, R. B. D'Agostino Sr, P. W. F. Wilson, and P. J. Savage. 2004. "Trends in Cardiovascular Complications of Diabetes." *Journal of the American Medical Association* 292: 2495-9.
- Hodgson, T. A. 1989. "Cost of Illness Studies: No Aid to Decision Making? Comments on the Second Opinion by Shiell, et al. (Health Policy, 8 (1987) 317-323)." *Health Policy* 11 (1): 57-60.
- Jones, A. 2000. "Health Econometrics." In *Handbook of Health Economics*, edited by A. Culyer and J. Newhouse, pp. 265-344. Amsterdam: Elsevier.
- Manning, W. G., and J. Mullahy. 2001. "Estimating Log Models: To Transform or Not to Transform?" *Journal of Health Economics* 20: 461-94.
- Manning, W. G., A. Basu, and J. Mullahy. 2005. "Generalized Modeling Approaches to Risk Adjustment of Skewed Outcomes Data." *Journal of Health Economics* 24 (3): 465-88.
- Max, W., D. P. Rice, H.-Y. Sung, X. Zhang, and L. S. Miller. 2004. "The Economic Burden of Smoking in California." *Tobacco Control* 13: 264-7.

- Miller, L. S., C. Ernst, and F. Collin. 1999. "Smoking-Attributable Medical Care Costs in the USA." *Social Science & Medicine* 48: 375-91.
- Mullahy, J. 1998. "Much Ado about Two: Reconsidering Retransformation and the Two-Part Model in Health Econometrics." *Journal of Health Economics* 17: 247-81.
- Rockhill, B., B. Newman, and C. Weinberg. 1998. "Use and Misuse of Population Attributable Fractions." *American Journal of Public Health* 88 (1): 15-9.
- Rothman, K. J., and S. Greenland. 1998. "Causation and Causal Inference." In *Modern Epidemiology*, 2nd ed., edited by K. J. Rothman and S. Greenland, pp. 7-29. Philadelphia PA: Lippincott-Raven Publishers.
- Roux, L., and C. Donaldson. 2004. "Economics and Obesity: Costing the Problem or Evaluating Solutions?" *Obesity Research* 12 (2): 173-9.
- Rowe, A. K., K. E. Powell, and W. D. Flanders. 2004. "Why Population Attributable Fractions Can Sum to More than One." *American Journal of Preventive Medicine* 26 (3): 243-9.
- Shiell, A., K. Gerard, and C. Donaldson. 1987. "Cost of Illness Studies: An Aid to Decision Making?" *Health Policy* 8: 317-23.
- Tarricone, R. 2006. "Cost-of-Illness Analysis: What Room in Health Economics?" *Health Policy* 77: 51-63.

**Table 1. Calculating Attributable Fractions (AF)**

(1) Person Types	(2) Predicted Expenditures Per Person	(3) Standard AF* Expenditures (Number of Persons)			(6) Sequential AF† Expenditures (Number of Persons)		(8) Complete Cross-Classification‡ Expenditures (Number of Persons)		
		(4) Observed	(5) Remove Condition 1	(5) Remove Condition 2	(6) Remove Condition 1	(7) Remove Condition 2	(8) Remove Condition 1	(9) Remove Condition 2	(10) Remove Both
Neither condition	\$1	\$100 (100)	\$200 (200)	\$200 (200)	\$200 (200)	\$400 (400)	\$200 (200)	\$200 (200)	\$200 (200)
Condition 1 only	\$5	\$500 (100)	\$0 (0)	\$1,000 (200)	\$0 (0)	\$0 (0)	\$0 (0)	\$500 (100)	\$500 (100)
Condition 2 only	\$5	\$500 (100)	\$1,000 (200)	\$0 (0)	\$1,000 (200)	\$0 (0)	\$500 (100)	\$0 (0)	\$500 (100)
Both conditions	\$25	\$2,500 (100)	\$0 (0)	\$0 (0)	\$0 (0)	\$0 (0)	\$2,500 (100)	\$2,500 (100)	\$0 (0)
Total expenditures <sup>§</sup>		\$3,600 (100)	\$1,200 (0)	\$1,200 (0)	\$1,200 (0)	\$400 (0)	\$3,200 (100)	\$3,200 (100)	\$1,200 (0)
			Attributable Costs**	Attributable Fraction††	Attributable Costs‡‡	Attributable Fraction††		Attributable Costs**	Attributable Fraction††
Condition 1			\$2,400	67%	\$2,400	67%		\$400	11%
Condition 2			\$2,400	67%	\$800	22%		\$400	11%
Both conditions								\$2,400	67%
Sum of conditions			\$4,800	133%	\$3,200	89%		\$3,200	89%
Other expenditures <sup>§§</sup>			\$400	11%	\$400	11%		\$400	11%
Sum of conditions and other			\$5,200	144%	\$3,600	100%		\$3,600	100%

\* In Standard AF calculations, one condition is removed leaving other conditions as observed. Those with only the condition of interest become “Neither condition,” and those with both conditions become the other condition only.

† In Sequential AF calculations, conditions are removed sequentially. Removing condition 1 moves those with “Condition 1 only” to “Neither condition” and those with “Both conditions” to “Condition 2 only.” Subsequently removing condition 2 moves everyone to “Neither condition.”

‡ In Complete Cross-Classification, combinations of conditions are treated separately. Removing a condition moves those with that condition only to “Neither condition”; those with both conditions remain.

§ Total expenditures are calculated by multiplying the number of people of each type by the expenditures per type and summing.

\*\* Attributable costs are calculated by subtracting total expenditures after removing a condition from observed total expenditures.

†† Attributable fractions are calculated by dividing attributable costs by observed total expenditures.

‡‡ Attributable costs in Sequential AF calculations are calculated by subtracting total expenditures after removing an additional condition from predicted total expenditures after removal of all previous conditions.

§§ Other expenditures are expenditures when conditions 1 and 2 have been completely removed (i.e., 400 people spending \$1 each).

**Table 2. Alternative Estimates of Prescription Expenditure Shares for 10 Conditions\***

(1) Condition	(2) Prevalence (%)	(3) Prevalence with Other Conditions (%) <sup>†</sup>	(4) (5) (6) Per Person Expenditures (\$)‡			(7) (8) (9) (10) Attributable Fraction (%)			
			OLS	GLM <sup>§</sup>	GLM—Lower**	OLS	GLM <sup>§</sup>	GLM—Lower**	GLM—Cross <sup>††</sup>
Other MH/SA	12.36	51.89	668.93	1,132.82	524.81	14.82	22.36	10.36	16.88
Hypertension	13.10	71.84	585.66	908.30	474.66	13.74	18.99	9.92	12.32
Diabetes	4.67	82.49	1,130.46	1,879.42	729.70	9.45	14.00	5.44	9.44
Arthritis	10.04	69.96	442.19	601.54	257.14	7.96	9.64	4.12	5.58
Dyslipidemia	6.89	80.35	655.59	968.39	423.97	8.09	10.65	4.66	6.50
Heart disease	5.88	79.46	672.41	953.59	382.67	7.09	8.96	3.59	5.26
Asthma	4.78	45.92	668.47	1,178.52	482.52	5.73	9.00	3.69	6.22
Skin disorders	9.39	51.74	245.72	395.48	230.06	4.13	5.93	3.45	3.59
Depression	0.85	67.81	1,251.42	2,614.11	840.95	1.91	3.56	1.15	2.39
HIV	0.09	69.74	7,648.78	17,315.83	5,612.44	1.17	2.36	0.77	2.18
Sum of attributable fractions						74.10	105.45	47.14	70.36
Combined effect <sup>‡‡</sup>						74.10		70.36	

Note: OLS = ordinary least squares; GLM = generalized linear model; MH/SA = mental health and substance abuse.

\* Results from 2000–2003 MEPS (N = 125,052). The dependent variable is pharmaceutical expenditures. All regressions include the 10 conditions and age, age squared, gender, race, geographic region, education, income level, and insurance provider. All dollars are 2005 dollars.

<sup>†</sup> The prevalence with other conditions is the prevalence of at least one other condition, given that a person has the condition listed.

<sup>‡</sup> Per person expenditures are the average difference between predicted expenditures with and without the condition for those observed with the condition.

<sup>§</sup> The GLM estimates are based on a two-part model of expenditures with a logit for positive prescription expenditures and GLM with log link and gamma variance for positive prescription expenditures. The GLM estimates calculate the counterfactual leaving all the other conditions and covariates as observed in the data and are equivalent to an upper-bound estimate.

\*\* The GLM-Lower estimates remove all other conditions in the data before performing the counterfactual.

<sup>††</sup> The GLM-Cross estimates use complete cross classification and Equation 4 to redistribute costs associated with joint conditions to constituent conditions.

<sup>‡‡</sup> The combined effect removes all conditions at once in the counterfactual.