Research Brief

Use of Econometric Models to Estimate Expenditure Shares

Justin G. Trogdon, Eric A. Finkelstein and Thomas J. Hoerger

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. Theoretical investigation and secondary data analysis.

Data Collection/Extraction Methods. Condition files used to define the presence of 10 medical conditions.

Principal Findings. Incremental effects of conditions on expenditures, expressed as a fraction of total expenditures, cannot generally be interpreted as shares. When the presence of one condition increases treatment costs for another condition, summing condition-specific shares leads to double-counting of expenditures.

Conclusions. Condition-specific shares generated from multiplicative models should not be summed. We provide an algorithm that allows estimates based on these models to be interpreted as shares and summed across conditions.

Key Words. Health expenditures, cost of illness, expenditure share, attributable fraction

Policy makers and researchers are often interested in understanding the health and financial burden on society from diseases or risk factors. For example, research has focused on quantifying the contribution of obesity, smoking, and other preventable risk factors or diseases on mortality (Mokdad et al. 2004; Flegal et al. 2005). Other research has attempted to allocate disability-adjusted life years (DALYs) across medical conditions or quantify the relative contribution of a specific condition to overall medical expenditures (Murray and Lopez 1996; Finkelstein, Fiebelkorn, and Wang 2003, 2004; Finkelstein et al. 2005). All of these studies have one common feature; they are attempting allocate shares of a total "pie" to specific conditions.

      Econometric methods are increasingly being used to estimate these shares (Akobundu et al. 2006) and, because the underlying relationship between the condition of interest and burden (be it morbidity, mortality, or

economic) is likely to be nonlinear, sophisticated modeling strategies are required. With respect to medical expenditures, an extensive literature has developed alternatives to ordinary least squares (OLS). These models assume the relationship between conditions and expenditures is not additive and accounts for the special properties of expenditure data (see Jones 2000, for a review; Manning, Basu, and Mullahy 2005; Cantoni and Ronchetti 2006). Models that are multiplicative in levels of expenditures are increasingly recommended and estimated. These models include OLS on log (positive) expenditures, nonlinear least squares (Mullahy 1998), and generalized linear models (GLM).

This study is the first to identify the problem of double-counting, which is unique to commonly applied regression models of medical 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 percent and thus generally cannot be interpreted as shares. This adding up problem is not well understood and as a result, policy makers are often misinformed of the relative burden of select conditions.

In this analysis, we show that the fractions can only be interpreted as shares if (1) the true underlying generating function is additively separable or (2) conditions are mutually exclusive (i.e., only one condition per person). The fractions can be interpreted as indicating how much lower medical expenditures would be in the absence of particular conditions, all else constant. However, the implication for estimating 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.

We provide two potential solutions. The first constructs bounds for condition-specific expenditures. The second provides a point estimate by distributing predicted expenditures for jointly occurring conditions to constituent conditions. The normalization involved in this method avoids double counting. Although our focus is on medical expenditures, analogous issues arise with DALYs, mortality, absenteeism, or other measures of burden.

Address correspondence to Justin G. Trogdon, Ph.D., RTI International, 3040 Cornwallis Rd., PO Box 12194, Research Triangle Park, NC 27709-2194. Eric A. Finkelstein, Ph.D., and Thomas J. Hoerger, Ph.D., are also with RTI International, Research Triangle Park, NC.

ATTRIBUTABLE FRACTIONS: DEFINITIONS AND PROPERTIES

In the epidemiologic literature, attributable fractions (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 (Rothman and Greenland 1998; Rockhill, Newman, and Weinberg 1998; Rowe, Powell, and Flanders 2004). We use the term to refer 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:

$$AF_i = \frac{E[y|d_i] - E[y|d_i = 0]}{E[y|d_i]}$$

The numerator of the AF predicts expenditures (y) using observed conditions and subtracts from that predicted expenditures setting the condition of interest (d_i) to zero and leaving all other covariates and conditions as they are in the data. This is divided by predicted expenditures to express condition-specific expenditures as a fraction of total expenditures.

Table 1 shows a simple two-condition numerical example of the double-counting problem. There are four types of people in this model (column 1); we assume that there are 100 persons of each type. 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. The third column shows predicted expenditures in the observed population. The fourth column shows the counterfactual where the first condition is removed. This 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. The attributable expenditures associated with condition 1 are $2,400 ($3,600–$1,200), yielding an AF of 67 percent. The AF for condition 2 is also 67 percent, making the sum of these AFs 133 percent of total expenditures. Furthermore, if both conditions were removed, spending would still be $400. Adding this component (11 percent) to the sum of the condition-specific AFs results in 144 percent of predicted total expenditures.

Table 1:   Calculating Attributable Fractions (AF)—Two-Condition Example of Multiplicative Model

Person Types (1)	Predicted Expenditures per Person (2)	Standard AF* Expenditures (Number of Persons)		
		Observed (3)	Remove Condition 1 (4)	Remove Condition 2 (5)
Neither condition	$1	$100 (100)	$200 (200)	$200 (200)
Condition 1 only	$5	$500 (100)	$0 (0)	$1,000 (200)
Condition 2 only	$5	$500 (100)	$1,000 (200)	$0 (0)
Both conditions	$25	$2,500 (100)	$0 (0)	$0 (0)
Total expenditures†		$3,600	$1,200	$1,200

	Attributable Costs‡	Attributable Fraction (%)§
Condition 1	$2,400	67
Condition 2	$2,400	67
Sum of conditions	$4,800	133
Other expenditures**	$400	11
Sum of conditions and other	$5,200	144

*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.

†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.

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

MULTIPLICATIVE MODELS WITH EXPONENTIAL MEANS

Medical expenditure data take on nonnegative values, often contain many zeros, and are heavily skewed. In the majority of models developed to deal with these issues, positive expenditures have an exponential conditional mean, $E(y \mid y > 0, X) = c \times \exp(X\beta)$, where c is a scale factor. These models include OLS on log (positive) expenditures, nonlinear least squares, and GLM with a log-link function; all are multiplicative.

Consider the following simple model of expenditures:

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

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 γ-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 $\beta_1 = \beta_2 = \ln(5)$.

　　　Because the spending associated with multiple conditions is greater than the sum of the spending associated with each of the conditions, the contribution to expenditures of the joint presence of the conditions will be attributed separately to both conditions in the condition-specific AFs (e.g., Table 1). The problem in interpreting AFs as shares is that both counterfactuals take credit for a large part of the reduction in expenditures for persons who have both conditions. Double-counting occurs even when the conditions are independent, as in Table 1. The conditions only need to occur jointly. Therefore, AFs in multiplicative models cannot be interpreted as shares of total expenditures due to each condition; summing condition-specific AFs will double-count attributable expenditures.

　　　The problem of double-counting in multiplicative models can be seen in the empirical example in Table 2. Consistent with prior studies (see Jones 2000), we estimated a two-part model of prescription expenditures using data from the 2000 through 2003 Medical Expenditure Panel Survey (MEPS; details available from the authors upon request). The two-part model is nonlinear and multi-plicative, implying interaction effects across the 10 conditions in the model; the intuition for multiplicative models above remains the same in a two-part model.

　　　Most of the 10 conditions have high prevalence rates and joint occurrence of conditions is common (Table 2 columns 2 and 3). High rates of joint conditions indicate that double-counting could be a major problem when summing condition-specific AFs in a multiplicative model. The standard AF calculations generated from the multiplicative model using only these 10 conditions add up to over 100 percent of total prescription expenditures, indicating that the AFs for these conditions cannot be interpreted as shares (Table 2 column 6).

RECOMMENDATIONS

Given the above discussion, how can researchers estimate shares via econometric models without double-counting?

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

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

*Results from 2000 to 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.

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

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

§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 γ 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.

††The GLM-cross estimates use complete cross classification and equation 1 to redistribute costs associated with joint conditions to constituent conditions.

‡‡The combined effect removes all conditions at once in the counterfactual.

GLM, generalized linear model; MH/SA, mental health and substance abuse.

Sequential Attributable Fractions

One approach 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 without the 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.

This will ensure no double counting occurs, but the order in which the conditions are removed is important for the sequential AFs. 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. In contrast, 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.

Unfortunately, there may 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 approach is demonstrated using the MEPS data in Table 2 (column 7). The lower-bound AFs are calculated by predicting expenditures with and without the condition, assuming all other conditions have been removed in the sample (i.e., dummy variables set to zero for these conditions). The lower-bound AFs are much smaller than the standard, upper-bound, AFs. On average, the AFs fall by more than 50 percent, highlighting the magnitude of the double counting problem.

The total impact of the entire set of conditions can be calculated by removing all conditions at the same time, leaving all else constant (Rowe, Powell, and Flanders 2004). The resulting AF accounts for the separate condition effects and the joint effects and provides 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. In the absence of all 10 conditions, expenditures would be lower by 70 percent. As expected, the

combined AF is in between the sum of AFs from the upper- and lower-bound calculations.

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]). Not only does this procedure avoid double-counting of expenditures, it makes explicit the share of expenditures associated with all possible combinations of conditions. However, the number of unique combinations of conditions 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 joint conditions back to the constituent conditions. We recommend a procedure to allocate the joint expenditures that satisfies the following principles. First, 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; thus, all conditions must be present and contribute equally to the observation. Second, the joint expenditures are equal to the product of multiplicative factors; therefore it is reasonable to attribute a greater share of the joint expenditures to the condition with the larger coefficient in the main effect. Third, a condition with no main effect should receive zero share. Fourth, the shares must sum to unity.

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:

$$s_k = [\exp(\beta_k) - 1]\Big/\sum_{k=1}^{K} [\exp(\beta_k) - 1] \tag{1}$$

where the βs are exponential mean model parameters. Subtracting 1 in the numerator ensures that conditions with no discernible impact on expenditures (i.e., $\beta_k = 0$), will receive zero share of the joint expenditures. The denominator ensures that the sum of the shares across the joint conditions is unity.

The last column of Table 2 demonstrates the results of the procedure using MEPS data. We calculate the shares in Equation 1 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 are observed in the data). The redistributed 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 percent. Double-counting has been removed.

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 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 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 reporting the upper and lower bounds for the AF for each condition described above as well as our complete cross-classification weighting scheme. The bounds are more accurate but less precise; in contrast, the cross-classification method is more precise and allows interpretation of AFs as shares of total expenditures that can be summed to get the total impact of the set of conditions, but relies on additional assumptions.

One limitation of our weighting scheme is that it does not rely on clinical theory to estimate condition-specific expenditures. If condition 1 is only expensive in conjunction with condition 2 but alone is relatively cheap to treat compared with condition 2, our weighting scheme would incorrectly assign most of the expenditures associated with the joint conditions to condition 2. In many applications clinical theory for causal relationships may not be

available; in these cases our weighting scheme provides a reasonable alternative. When theory is available, complete cross-classification could be combined with alternative weighting schemes that incorporate clinically meaningful relationships among conditions.

Future research should explore the extent to which the bounds can be improved. It can also build on the intuition underlying the proposed strategies to deal with double counting in other measures of burden (e.g., morbidity, mortality, or productivity) and provide better estimates of the relative burden of conditions.

ACKNOWLEDGMENT

REFERENCES

Akobundu, E., J. Ju, L. Blatt, and C. D. Mullins. 2006. "Cost-of-Illness Studies: A Review of Current Methods." *Pharmacoeconomics* 24: 869–90.

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: 198–213.

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: 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)*: W3-219–226.

———. 2004. "State-Level Estimates of Annual Medical Expenditures Attributable to Obesity." *Obesity Research* 12: 18–24.

Flegal, K. M., B. I. Graubard, D. F. Williamson, and M. H. Gail. 2005. "Excess Deaths Associated with Underweight, Overweight, and Obesity." *Journal of the American Medical Association* 293: 1861–7.

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., A. Basu, and J. Mullahy. 2005. "Generalized Modeling Approaches to Risk Adjustment of Skewed Outcomes Data." *Journal of Health Economics* 24: 465–88.

Mokdad, A. H., J. S. Marks, D. F. Stroup, and J. L. Gerberding. 2004. "Actual Causes of Death in the United States, 2000." *Journal of the American Medical Association* 291: 1238–45. Erratum in: *JAMA* 2005 293 (3):293–4.

Mullahy, J. 1998. "Much Ado about Two: Reconsidering Retransformation and the Two-Part Model in Health Econometrics." *Journal of Health Economics* 17: 247–81.

Murray, C. J. L., and A. D. Lopez. 1996. *The Global Burden of Disease: A Comprehensive Assessment o f Mortality and Disability from Diseases, Injuries, and Risk Factors in 1990 and Projected to 2020. Global Burden of Disease and Injury Series*, Vol. 1. Cambridge, MA: Harvard School Public Health/WHO/World Bank.

Rockhill, B., B. Newman, and C. Weinberg. 1998. "Use and Misuse of Population Attributable Fractions." *American Journal of Public Health* 88: 15–9.

Rothman, K. J., and S. Greenland. 1998. "Causation and Causal Inference." In *Modern Epidemiology*, 2d Edition, edited by K. J. Rothman and S. Greenland, pp. 7–29. Philadelphia: Lippincott-Raven Publishers.

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: 243–9.