

Editorial Manager(tm) for Health Services and Outcomes Research Methodology
Manuscript Draft

Manuscript Number: HSOR81R1

Title: Predicting the Frequency and Amount of Health Care Expenditures

Article Type: Manuscript

Keywords: Annual Expenditures model

Corresponding Author: Professor Edward W. Frees, Ph.D.

Corresponding Author's Institution: University of Wisconsin

First Author: Edward W. Frees, Ph.D.

Order of Authors: Edward W. Frees, Ph.D.; Edward W Frees, Ph.D.; Jie Gao, Ph.D.; Marjorie A Rosenberg, Ph.D.

Reply to Referee Number 1 on the Paper, “Predicting the Frequency and Amount of Health Care Expenditures”. (Manuscript Number HSOR81)

We thank each reviewer for detailed comments. To make sure that we have not missed anything, we re-produce the comments below (in small type) followed by our response. Because the comments are long, in some places we broke them up. We hope that this does not change the meaning nor intent of the remarks.

Referee Comment 1. Several features of this prediction exercise are intriguing. The standard approach to prediction is based on reduced form type equations (linear or nonlinear) in which the conditioning is with respect to variables that are exogenous, or (in a static framework) strictly exogenous. That is, marginal distributions are employed to generate predictions. If several different predictors are to be compared, then one assumes a given loss function and evaluates the performance relative to that loss function. Typically both the bias and the variance of the prediction error matters. In the exercise of this paper, however, prediction is based on conditional and not marginal distributions, and conditioning is with respect to variables that typically would not be assumed to be exogenous. Specifically, insurance status is a dubious conditioning variable when observational data are used, and there exists a large literature that analyzes the interdependence between health care utilization and insurance status.

Our response: This is a fair summary of the work – prediction based on conditional distributions in the focus of this work.

We agree that conditioning on endogenous variables such as insurance status is inappropriate for many models, particularly for analysts concerned with joint choices of insurance and health care utilization. However, for other analysts, particularly those who focus on insured populations, it is critical to condition on insurance status. We know that insured populations consume health care differently than non-insured populations. Although not exogenous, for some applications we can treat insured status as “sequentially exogenous” in the sense that the variable is known at the beginning of the year. This is our position for this application.

To tie this together, we now use insured status at the beginning of the year (actually, the month of January), not throughout the entire year as in the prior draft. Thus, you may note that all of the coefficients and predictions have changed slightly from the prior version, and we have altered the corresponding text in the paper. Qualitatively, the results are generally the same as in the prior version of the paper. We thank the reviewer for bringing this to our attention.

Referee Comment 1 - Continued. In the same vein, the factorization $f(y, N) = f_1(y|N)f_2(N)$ can be questioned. Strictly speaking, for the approach of the paper to be valid, we require that $f(y, N; \boldsymbol{\theta}) = f_1(y|N, \boldsymbol{\theta}_1)f_2(N; \boldsymbol{\theta}_2)$, and that $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are functionally independent. This is noted by the authors in the text following equation (2) on page 7. However, this is a very strong assumption because it would require that there are no latent factors that conditionally affect both the frequency and the intensity of health care service utilization. Because the authors' conditioning variables only include a set of relatively coarse health status variables, the possibility remains that the two variables are not conditionally independent as assumed. This argument invalidates the prediction equation.

Our response: We agree that we assume the parameters to be functionally independent. Of course, even if they are not, the relationship $f(y, N; \boldsymbol{\theta}) = f_1(y|N, \boldsymbol{\theta})f_2(N; \boldsymbol{\theta})$ holds, it is just that one needs to do the likelihood evaluation over both pieces simultaneously. One no longer has the convenience of analyzing the parts separately. The same argument holds for the two-part model.

As the reviewer notes, this relationship can also be confounded by a latent variable z , so that the factorization could be (1) $f(y, N, z; \boldsymbol{\theta}) = f_1(y, z|N, \boldsymbol{\theta})f_2(N; \boldsymbol{\theta})$, (2) $f(y, N, z; \boldsymbol{\theta}) = f_1(y|N, z, \boldsymbol{\theta})f_2(N, z; \boldsymbol{\theta})$ or (3) $f(y, N, z; \boldsymbol{\theta}) = f_1(y, z_1|N, z_2, \boldsymbol{\theta})f_2(N, z_2; \boldsymbol{\theta})$, where z_1 and z_2 are components of z . In this case, z may affect either the frequency, severity or both. The reviewer is correct in pointing out that the model that we propose may be incorrect in the presence of important omitted variables, z . Again, we do not believe that we are making assumptions that are more or less sensitive to this assumption than others in the literature.

Referee Comment 2. Even if the preceding argument is ignored for the moment, the correctness of the prediction equation is questionable. Observe that the factorization applies in the case of the second part where $y > 0$ and $N > 0$. Therefore, it seems to me that the factorization written more explicitly should be $f(y, N|y > 0, N > 0) = f_1(y|N > 0)f_2(N|N > 0)$. Hence the relevant distribution for N is truncated negative binomial rather than the untruncated version used. The fitted conditional mean of N from this truncated distribution would be different from that used by the authors and that difference would change both the point predictions and the simulated predictive distributions.

Our response: We use the same arguments as regularly applied in the two-part model except that now we allow the frequency portion, N , to be any count distribution, not just a Bernoulli distribution.

Specifically, let $\mathbf{Y} = (Y_1, \dots, Y_N)'$ and N be random. Use corresponding lower case symbols for arguments in the distribution function. Now, joint distribution functions are well-defined even when one variable/vector, such as \mathbf{Y} , has a continuous distribution, and the other, N , has a discrete distribution.

We can write the joint distribution as $\Pr(\mathbf{Y} \leq \mathbf{y}, N \leq n) = \sum_{k=0}^n F(\mathbf{y}, k)$, where $F(\mathbf{y}, k) = \Pr(\mathbf{Y} \leq \mathbf{y}, N = k)$. For $k = 0$, we interpret $F(\mathbf{y}, 0) = \Pr(N = 0)$. For $k > 0$, we have $F(\mathbf{y}, k) = \Pr(\mathbf{Y} \leq \mathbf{y}, N = k) = \Pr(Y_1 \leq y_1, \dots, Y_k \leq y_k | N =$

$k) \times \Pr(N = k)$. Assuming conditional independence of the Y 's, we summarize this as:

$$F(\mathbf{y}, k) = \begin{cases} \Pr(N = 0) & k = 0 \\ F(y_1|N = 1) \Pr(N = 1) & k = 1 \\ F(y_1|N = 2)F(y_2|N = 2) \Pr(N = 2) & k = 2 \\ \vdots & \vdots \\ \prod_{j=1}^k F(y_j|N = k) \Pr(N = k) & k \\ \vdots & \vdots \end{cases}$$

To go to the combination of joint probability density function and mass function used in likelihood equations, one uses the factorization

$$f(\mathbf{y}, n) = \prod_{j=1}^n f(y_j|N = n) \Pr(N = n),$$

as at the end of Section 2.1. This argument uses the usual (untruncated) version of the negative binomial, not the truncated one.

Referee Comment 3. There are several aspects of the data that should be at least mentioned in an exercise such as this. Expenditures are usually related to completed episodes of illness. Unavoidably, annual data will include some incomplete (censored) episodes. The literature on health expenditures also emphasizes the importance of end-of-life stage in generating high values. Unfortunately there is no mapping from N to the number of episodes.

Our response: Beginning on page 2 of the paper (second full paragraph), we reviewed the work on episodes. This is important in that our work could either be tied to events or to episodes; the statistical model takes advantage of replications and can be applied to either episodes or events. If someone wanted to analyze episodes (a very sensible thing to do in some contexts), then the analysis could use our statistical model but follow expenditures per episode, not expenditures per event. We investigated expenditures per event because our goal was to predict annual accounting expenditures.

The key point made by the reviewer is that this framework assumes “complete” data, that is, data that are not truncated nor censored. In contrast, episodes almost by definition cross calendar year boundaries. We view this as a problem that can be quite important in some contexts and yet is certainly not the main focus of our work. We also note that the standard two-part model also assumes complete data. We have added a paragraph in the discussion section to amplify this point.

Referee Comment 4. The authors use a large list of standard exogenous predictors. They could use even more. For example, MEPS provides information on age and number of chronic conditions, variables that are not used here even though they are known to have good predictive power.

Our response: We agree wholeheartedly; thank you for this remark. We believe that we have done a reasonable review of the literature in this regard. It looks like our section on explanatory variable is small. However, we did that in response to prior reactions to the paper that advised us to keep the focus on the prediction part and not on the independent variables. We are certainly open to using additional variables if you have specific suggestions. Our main goal in this paper was to provide a different way of modeling the data and to use the variables as simply as possible.

Referee Comment 5. The research objective behind the main exercise of this article is not clear. If some of the preceding arguments are accepted then one cannot regard this paper as providing a correct and useful methodology for generating predictions from a two-part model. If such is the authors' objective, they should use a simpler variant of the TPM to implement their exercise. In a purely predictive exercise the discussion of the role of individual factors, such as that provided on pp. 12-13, is largely irrelevant and should be dropped. Reduced form equation coefficients are usually hard to interpret without knowing the underlying structural model. The issue of objective also affects how one views the exercise of comparing simulated "predictive distributions." It is unclear to the reader exactly what one can learn from Figure 1 or Table 6, where three very different distributions are presented. The differences could reflect population heterogeneity or enormous variability of predictions or enormous randomness in N. It is not useful to present these without giving the reader some indication of the source of variability.

Our response: The research objective of this paper is to provide methods that can be used to predict annual expenditures, given expenditures by event are available in a prior year on a cross-section of subjects.

We agree that the reduced form of the predictive model can differ from an underlying structural model. That does not, however, invalidate the usefulness of our model. It is comforting to users to be able to interpret coefficients of a predictive model although we grant that this is not the main focus, as in a structural model.

Figure 1 shows the application of the model to graphically represent its usefulness in predicting expenditures. It is possible to provide predictive distributions for segments of a population (thus reducing population heterogeneity). With our models, we could also provide predictive distributions for a single event ($N=1$) or a number of events. The goal was to provide an illustrative result of one way of providing information to managers of expenditures-type data.

Reply to Referee Number 2 on the Paper, “Predicting the Frequency and Amount of Health Care Expenditures”. (Manuscript Number HSOR81)

We thank each reviewer for detailed comments. To make sure that we have not missed anything, we re-produce the comments below (in small type) followed by our response. Because the comments are long, in some places we broke them up. We hope that this does not change the meaning nor intent of the remarks.

Referee Comment This manuscript presents an alternative method for predicting health care cost. As the authors state, the methods described are extensions of the two-part model commonly used in health economics. I must admit, it took me several reads to determine how the “new” models were implemented. While interesting, I question whether the extent the new models would be utilized. One strength of the two-part model is it can, operationally, be applied to any outcome variable (e.g., total expenditures, total medical expenditures, ER cost, IP cost, etc.). It appears to me that the “new methods” would apply only when the utilization associated with an expenditure could be enumerated. For example, outpatient cost and outpatient visits must both be captured. I would like the author to address this in a comparison of two-part models vs. the “new” approach.

Our response: We agree that the two-part model will be applicable in more cases than our proposed annual expenditure model. The new model is only available when expenditures by event are available. Our goal was to take the “gold standard,” the two-part model, and show how this could be improved upon when additional information is available.

As with the two-part model, this approach using expenditures could be applied to inpatient or outpatient expenditures, as we have done in our paper. It could also be applied to dental expenditures, or other types of expenditures. It requires more information as inputs (through data at the expenditure level) but we think provides additional information as outputs through additional predictive ability.

Referee Comment Not sure if the 1-part model is needed. If so, why not a GLM with log link?

Our response: We agree that the one-part model is not a strong alternative candidate. We included it as a basic benchmark that is still being used in some contexts. The one-part model could certainly be strengthened by, for example, using a gamma regression with a logarithmic link function. However, we felt that this has been amply investigated elsewhere in the literature. Our goal was to show how the two-part model could be improved, not the one-part model. We note that we did use the gamma regression model with a logarithmic link as part of our two-part model (see Table 5).

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

Predicting the Frequency and Amount of Health Care Expenditures

Edward W. Frees^{a*}, Jie Gao^a, and Marjorie A. Rosenberg^{ab}

^aSchool of Business, University of Wisconsin-Madison, Madison, WI 53706

^bDepartment of Biostatistics and Medical Informatics, School of Medicine and Public Health, University of Wisconsin-Madison, Madison, WI 53706

This article extends the standard two-part model for predicting health care expenditures to the case where multiple events may occur within a one-year period. The first part of the extended model represents the frequency of events, such as the number of inpatient hospital stays or outpatient visits, and the second part models expenditure per event. Both component models also use independent variables that consist of an individual's demographic and access characteristics, socioeconomic status, health status, health insurance coverage, employment status and industry classification. The second part model also includes a variable representing the number of events to predict the expenditure per event, creating a dependency between the first and second parts. This article introduces closed-form predictors of annual total expenditures and demonstrates how to create simulated predictive distributions for individuals and groups.

The data for this study are from the Medical Expenditure Panel Survey (MEPS). MEPS panels 7 and 8 from 2003 were used for estimation, panels 8 and 9 from 2004 were used to validate predictions. This annual expenditures model provided a better fit to the data than standard two-part models. The count variable was significant in predicting outpatient expenditures. The aggregate expenditures model provided better point predictions of held-out total expenditures than competing models, including the standard two-part model. The predictive distribution for aggregate expenditures for small groups is long-tailed, with both the variability and skewness decreasing as the group size increases, an important point for programs designed to manage expenditures.

Key Words: Annual Expenditures model

*Corresponding author Tel.: +1-608-262-0429; fax: +1-608-265-4195. *Email address:* jffrees@bus.wisc.edu (E.W. Frees).

1. Introduction

Monetary savings is an integral part of measuring the success of clinical interventions. Cost-effectiveness studies, that examine the difference between one group versus another, require cost estimates of health care. Disease management programs, similar in concept to cost-effectiveness studies, target certain high cost chronic diseases to measure whether interventions, such as a lifestyle change or drug introduction, can influence the course of a disease. Cost-effectiveness studies and disease management programs are two important topics in which the measurement of monetary values is a central issue. For health care data that are by nature long-tailed and have discrete and continuous components, prediction of the level of costs is not simple.

In modeling medical services utilization, two-part models (TPM) have been a widely used tool to model frequency and cost of medical services. The first part of a TPM typically models the distinction between users and non-users of service via a probit or logit regression. The second part describes the distribution of amount, conditional on some use, modeled either as a continuous distribution or as an integer-valued distribution (Duan et al. 1983; Jones 2000; Manning et al. 1987; Mullahy 1998; Pohlmeier and Ulrich 1995). The purpose of this paper is to extend the TPM by using more detailed information about the frequency of use and to improve the prediction of annual expenditures at an individual or group level.

Two-part models are motivated by important features of health care demand data. One feature is that data typically contain a mixture of zeros and a continuous distribution for non-zero values. A second feature is that the level of health care usage, once occurred, is largely unaffected by an individual's decision to seek treatment (Manning et al. 1987), as a physician mainly decides the intensity of expenditures as suggested in the principal-agent model (Zweifel 1981). This distinction between the decision to utilize and the subsequent level of services suggests that traditional censored regression models, such as the Tobit, are inappropriate.

The TPM is plausible for an analysis of a single illness spell, where individuals initiate

the first encounter of care and physicians decide on the intensity of the treatments during the rest of the illness spell. However, in an analysis of total health expenditures for a year, individuals can have multiple illness events during the year. The TPM's simplified modeling of the probability of care upon initiation of the first visit may not be defensible in an analysis of cross-sectional data for a year unless one believes that the initial visit for the year has some special characteristics (Deb and Trivedi 2002). Modeling total health care expenditures benefits from a methodology that uses frequency of multiple illness events and expenditures characteristics from each event rather than the simplified modeling treatment of binary outcomes for use and non-use.

This extension of the TPM is natural and certain aspects have been proposed in the non-healthcare services literature. In the actuarial and insurance literature, this extension is known as a "collective risk model" where it is typically modeled without explanatory variables (Klugman, Panjer and Willmot 2008). An exception is Pinquet (1998) that incorporates explanatory variables into a model of automobile insurance claims using Bayesian methods.

In the health economics literature, our extension of the TPM is related to the paper of Keeler and Rolph (1988) that organizes an individual's expenditures into episodes of treatment; each episode contains all spending associated with a given bout of illness, chronic condition or procedure. The main purpose of the Keeler and Rolph (1988) paper was to infer effects of health insurance plans, such as coinsurance, on expenditures. Keeler and Rolph (1988) use a random coefficients model to model logarithmic expenditures (their random coefficients account for intra-family correlations, not intra-individual correlations as in this paper). Further, they use count distributions to model the number of episodes, not simply binary variables as with the TPM. Keeler and Rolph found that almost all of the insurance plan effects on total expenditures were due to the number of episodes; the cost per episode was generally unrelated to insurance plan cost sharing features. Unlike our paper, Keeler and Rolph (1988) deals with episodes, not events. Although similar in the sense that these

1
2
3
4 are both ways of decomposing annual expenditures, by dealing with events we are able to
5 use the information in the frequency component to help predict expenditures (see Section 2).
6
7 The main purpose of this paper is to provide predictions of an individual's future healthcare
8 expenditures.
9

10
11
12 In other related work, Rosenberg and Farrell (2007) modeled inpatient utilization at
13 the individual-level with a Bayesian model, assuming the number of hospitalizations was
14 from a Poisson distribution with a lognormal prior for the mean number of hospitalizations.
15 Costs per hospitalization were assumed to have a gamma distribution. A follow-up paper,
16 Rosenberg and Farrell (2008), used this model to predict expenditures and utilization by
17 individual and as a group using Bayesian methods.
18
19

20
21 The primary focus of this paper is on prediction. Given a set of individual-level charac-
22 teristics such as demographic, economic attributes, and health history at the beginning of
23 a year, we predict the utilization of health care during the year for an individual or for a
24 group. We develop a point estimate of the prediction, as well as the predictive distribution,
25 to provide a better understanding of the range of possible expenditure levels for use in cost
26 effectiveness analyses and disease management programs.
27
28

29
30 In our expenditures model, we allow the expenditures per event to be a function of the
31 annual frequency of events. To illustrate, one might consider a situation where some latent
32 attribute of an individual simultaneously determines, for example, high frequency and low
33 expenditures. We point out in Section 6 that one could use a latent variable model to capture
34 this example. However, this paper focuses on modeling observable variables, frequency and
35 expenditures. To underscore the importance of this feature, in our data analysis we exam-
36 ine two data sets: hospital inpatient expenditures that are relatively infrequent, and more
37 frequently occurring outpatient expenditures. We find that using the frequency information
38 to predict expenditures is important for the outpatient case although not so for inpatient
39 expenditures.
40
41

42
43 The plan for the paper is as follows. We propose our extension of the TPM and review al-
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

ternative methods in Section 2. Section 3 presents our prediction strategies, including point predictions as well as simulated predictive distributions. Section 4 describes the Medical Expenditure Panel Survey (MEPS) data used for this study. Section 5 discusses the prediction results and Section 6 concludes with a few additional remarks.

2. Modeling Expenditures

We use separate models for inpatient and outpatient care. Let N_i be the number of events, for either inpatient stays or outpatient visits, and let y_{ij} , $j = 1, \dots, N_i$, be the amount per event. By convention, the set $\{y_{ij}\}$ is empty when $N_i = 0$. Let $\mathbf{y}_i = (y_{i1}, \dots, y_{i,N_i})'$ be the vector of individual expenditures. In addition to each expenditure per event, our interest is in *annual expenditures*, $S_i = y_{i1} + \dots + y_{i,N_i}$. For two-part models defined below, we also define r_i to be a binary variable indicating the presence of a claim (health care utilization), that is, $r_i = 1$ on the event $\{N_i > 0\}$ and is 0 otherwise. In traditional actuarial modeling, one assumes that the distribution of losses is, conditional on the frequency N_i , identical and independent over replicates j . This representation is known as the *collective risk model*, see Klugman et al. (2008, Chapter 9).

The individual-level explanatory variables are denoted by the vector \mathbf{x}_i . The observable data available consists of $\{N_i, y_{i,1}, \dots, y_{i,N_i}, \mathbf{x}_i, i = 1, \dots, n\}$

2.1. Annual Expenditures Model

For the frequency component, we use a negative binomial count regression model with N_i as the dependent variable and \mathbf{x}_{1i} as the set of explanatory variables with regression coefficients $\boldsymbol{\beta}_1$. For the expenditures amount component, we condition on $N_i > 0$ and use a mixed linear regression model with $\ln(y_{ij})$ as a dependent variable and a latent effect, α_i , to model the correlation among expenditures within an individual. In the usual notation, one can express this model as

$$y_{ij} = \alpha_i + \mathbf{x}'_{2i}\boldsymbol{\beta}_2 + N_i\beta_N + \varepsilon_{ij},$$

where α_i is mean zero and has variance σ_α^2 . The \mathbf{x}_{2i} variables are a subset of the explanatory variables, \mathbf{x}_i . By conditioning on frequency, we can use the count variable N_i as another potential explanatory variable, thus formally allowing for the possibility that the amounts may depend on the frequency.

From a likelihood perspective, one can motivate the annual expenditures model as follows. Suppressing the $\{i\}$ subscript, we decompose the joint distribution of the dependent variables as $f(N, \mathbf{y}) = f(N) \cdot f(\mathbf{y}|N)$, where $f(N, \mathbf{y})$ denotes the joint distribution of (N, \mathbf{y}) . With this notation, $f(N)$ represents the negative binomial frequency model and $f(\mathbf{y}|N)$ represents the conditional mixed linear regression model.

2.2. Alternative Models

We compare our annual expenditures model to classical one- and two-part models. The basic one-part model can be expressed through the linear regression model equation $S_i = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i$. This method is not always appropriate to use for health care expenditures that tend to be right-skewed and contain a large number of zeros. A popular alternative is to log-transform the dependent variable so that the data become less skewed and the normality assumption is more plausible. Further, the transform $\ln(1 + S_i)$ accommodates the presence of zeros.

Other one-part models include the well-known Tobit model for censored, or limited dependent variables, and the so-called ‘‘Tweedie’’ generalized linear model. The Tweedie model is used extensively in actuarial applications; it is based on a compound distribution that is a Poisson sum of gamma random variables. This distribution is a mixture of a positive mass at zero and a continuous component. As a member of the natural exponential family of distributions, it can readily be used as a special type of generalized linear model. See Smyth and Jorgensen (2002) for further details. Because it is well-known that one-part models are outperformed by two-part models for prediction (Mullahy, 1998), in our subsequent discussion we refer only to the basic linear regression model (with the dependent variable $\ln(1 + S_i)$) as our benchmark for performance among competing models.

In contrast to one-part models, a TPM handles the large number of zeros that are encountered in health care by decomposing overall expenditures into (i) zero versus non-zero expenditures and (ii) the amount of positive expenditures. The first part of the model assesses whether some utilization has occurred that we denote as $r_i = 1$. Typically, one uses a logit or probit model to represent this binary variable.

The second part of the model follows the one-part model. By conditioning on the event $r_i = 1$, we remove the zeros from the distribution of S_i . The second part may be based on (i) a linear regression, (ii) a log-transformed regression model or (iii) a generalized linear model. In a generalized linear model (GLM), the mean is a function of explanatory variables. In the health care literature, it is common to use a gamma regression model (e.g., Blough, Madden and Hornbrook, 1999; Manning, Basu and Mullahy, 2005). With this specification, it is customary to use logarithmic means as a linear combination of explanatory variables, so that $\ln(E S_i) = \mathbf{x}'_i \boldsymbol{\beta}$. The standard TPM is a special case of the annual expenditures model with N_i as a binary variable.

3. Prediction

We use estimates of coefficients, smearing factors and individual random effects based on 2003 data to predict expenditures for individuals from the 2004 MEPS survey. The smearing factor is the average of exponentiated residuals, an estimate of $E(e^\epsilon)$ (Duan et al. 1983). We compare alternative models using point predictions and then interpret the predictive distribution from the annual expenditures model.

3.1. Point Prediction: Annual Expenditures Model

Our primary goal is to predict annual expenditures S_i . For some important special cases, we are able to provide some closed-form point predictors. From the second part of the model, we observe that $E(\ln(y_{ij})|\alpha_i, N_i) = \alpha_i + \mathbf{x}'_{2i} \boldsymbol{\beta}_2 + N_i \beta_N$. This allows us to predict expenditure per event. For person i in the data set, a predictor for y_{ij} is $\hat{y}_{ij} = \exp\left(\hat{\alpha}_i + \mathbf{x}'_{2i} \hat{\boldsymbol{\beta}}_2 + N_i \hat{\beta}_N\right)$. For prediction of expenditures of person i not in the data set, we substitute zero as a predictor

of α_i .

A predictor of annual expenditures, as a function of N_i , is $\widehat{S}_i(N_i) = \widehat{y}_{ij}N_i$. However, this predictor is not useful at the beginning of the year because N_i is not known. Averaging this predictor over the estimated distribution of N_i yields our basic predictor of annual expenditures,

$$\widehat{S}_i = \exp\left(\widehat{\alpha}_i + \mathbf{x}'_{2i}\widehat{\boldsymbol{\beta}}_2\right) M'_{N_i}(\widehat{\boldsymbol{\beta}}_N). \quad (1)$$

Here, $M'_{N_i}(\widehat{\boldsymbol{\beta}}_N)$ is the derivative of the moment generating function of N_i evaluated at $\widehat{\boldsymbol{\beta}}_N$. Recall that the derivative of a moment generating function for a count random variable is computed as $M'_{N_i}(t) = \sum_k k \exp(kt) \widehat{\Pr}_i(N_i = k)$ and $\widehat{\Pr}_i(N_i = k)$ is the fitted frequency distribution from the first part of the model.

We consider three special cases to interpret $M'_{N_i}(\widehat{\boldsymbol{\beta}}_N)$. First, suppose there are no important covariates for the frequency model. In this case, the adjustment term $M'_{N_i}(\widehat{\boldsymbol{\beta}}_N)$ does not vary with i , similar to Duan's (1983) smearing factor adjustment. Second, suppose that $\widehat{\boldsymbol{\beta}}_N = 0$, indicating that the frequency component is not an important predictor for the amount regression. Then, the predictor reduces to

$$\widehat{S}_i = \exp\left(\widehat{\alpha}_i + \mathbf{x}'_{2i}\widehat{\boldsymbol{\beta}}_2\right) \widehat{E}(N_i), \quad (2)$$

that is, an expenditures component multiplied by an expected frequency. This is the classic predictor assuming independence between amounts and frequency (Klugman, Panjer and Willmot 2008). Third, for the negative binomial distribution, the Appendix shows the computation of the derivative of the moment generating function of N_i that yields an explicit form for the predictor \widehat{S}_i .

A variation is to use $\widehat{S}_i \times SME$ where the second term (SME) is the average of exponentiated residuals "smearing factor" adjustment from the expenditures regression model. (Duan et al. 1983)

3.2. Point Predictions: Alternative Models

For the one-part model that uses a logarithmic transformation accommodating zeros, we have $S_i = \exp(\mathbf{x}'_i\boldsymbol{\beta} + \varepsilon_i) - 1$. Assuming $\{\varepsilon_i\}$ are identically and independently distributed, then the resulting predictor is $\widehat{S}_i = \exp(\mathbf{x}'_i\widehat{\boldsymbol{\beta}}) \cdot SME - 1$. A simpler predictor is $\widehat{S}_i = \exp(\mathbf{x}'_i\widehat{\boldsymbol{\beta}}) - 1$ that results from assuming normality for $\ln(S)$ and using the median as a predictor.

For the two-part models, prediction of annual expenditures is the product of predictions of the two parts. The first part is straightforward; it is the predicted probability of an expenditure using a logistic regression model. We allow the second part to follow (i) a linear regression, (ii) a log-transformed regression model and (iii) a generalized linear model. For prediction with a log-transformed dependent variable, we used both Duan's smearing estimate and the median to predict expenditures. For the generalized linear model with the gamma distribution, we use $\widehat{S}_i = \exp(\mathbf{x}'_i\widehat{\boldsymbol{\beta}})$ as a predictor of expenditures.

3.3. Simulation of Predictive Distribution

To provide further intuition, we supplement the point predictions with simulation to illustrate the predictive distribution of annual expenditures. Through simulation, we can predict the number of events associated with each person, and given that number, simulate expenditures per event.

In many applications, it is not only of interest to provide predictions for specific individuals but also for a group of individuals. If S_i represents the random sum of 2004 total expenditures associated with the i th individual, then we might be interested in predicting $\sum_{i=1}^n S_i$ for individuals $i = 1, \dots, n$. For example, this set of individuals might represent employees of a firm or members of an association. Predicting distributions involving convolutions is straightforward with simulation, and demonstrates that the predictive distribution has discrete and continuous components. From our hold-out sample, we randomly selected groups of size 1, 5, 25, 50, 100 and 250 participants to form our groups. All simulations are based on 5,000 replications. We present results using the entire sample. We present predictions of the *average* expenditure (the predicted sum divided by group size) to enhance comparability

1
2
3
4 among the groups.
5
6

7 8 **4. Data** 9

10 The publicly available data collected in MEPS contains detailed information on each med-
11 ical care event by type of service including physician office visits, hospital emergency room
12 visits, hospital outpatient visits, hospital inpatient stays, all other medical provider visits,
13 and use of prescription medicine. The detailed information allows us to analyze utilization
14 and level of expenditures from inpatient admissions and outpatient visits. For estimation
15 purposes, we use panels 7 and 8 from the MEPS data from calendar year 2003. In this data
16 set, there were $n=18,735$ individuals between ages 18 and 65. For prediction purposes, we
17 use panels 8 and 9 from calendar year 2004. Specifically, we consider $n=9,472$ participants
18 from the 2004 MEPS panel 8 who were interviewed in 2003 as well as $n=9,657$ participants
19 from the 2004 MEPS panel 9 who were not interviewed in 2003.
20
21
22
23
24
25
26
27
28
29
30

31 Our dependent variables consist of utilization measures, number of visits and dollars of
32 expenditures, for both inpatient admissions and outpatient visits. Inpatient admissions
33 include persons who were admitted to a hospital and stayed overnight. Outpatient visits
34 include hospital outpatient department visits, office-based provider visits and emergency
35 room visits excluding dental services. Hospital stays with the same date of admission and
36 discharge are excluded from inpatient counts and expenditures, but are included in outpatient
37 count and expenditures. Payments associated with emergency room visits that immediately
38 preceded an inpatient stay are included in the inpatient expenditures. Prescribed medicines
39 that are linked to hospital admissions are included in inpatient expenditures.
40
41
42
43
44
45
46
47
48
49

50 Table 1 shows the frequency distribution of inpatient admissions and outpatient visits.
51 For inpatient admissions, we see that only 1.4% ($= 100 \times (18735 - 17322 - 1154)/18735$) had
52 more than one inpatient admission during the year, suggesting the use of a binary frequency
53 model such as used in the TPM. For outpatients visits, 54.8% of survey participants had
54 more than one visit during the year, making the TPM a less intuitively appealing choice to
55
56
57
58
59
60
61
62
63
64
65

represent this type of expenditure.

[Table 1 is about here]

Independent variables that help explain the variance of health care utilization were categorized as demographic, access such as geographic region and usual source of care, socioeconomic status, health status, health insurance coverage, employment status and industry classification. Our selection of independent variables follows a large healthcare literature that is summarized by Gao (2007).

Demographic factors include age, sex and ethnicity. We use census region to proxy the accessibility of health care services, and the overall economic or regional impact on residents' health care behavior. In addition to using census region as a proxy for availability of care, we include a variable indicating whether there is a particular place that an individual usually seeks care.

Socioeconomic factors that impact health care utilization include education, marital status, family size and income. In MEPS, education is represented by three categories: lower than high school, high school, and college or above. Marital status is another socioeconomic factor that gives rise to interpersonal differences in the need for health services and expected health care expenditures. Individuals are classified as married, widowed, divorced or separated, and never married. The measure of income is annual income as compared to the poverty line. Self-rated physical health, mental health and any functional or activity related limitations during the sample period are used as proxies for health status.

Health insurance coverage is a predictor of health care utilization as it reduces out-of-pocket expenditures and can induce moral hazard with increased use of health care. One modeling issue with the insurance variable is that the choice to obtain insurance coverage and lifestyle decisions that reflect health status, may be considered endogenous determinants of healthcare utilization (e.g., Goldman et al., 2002; Mello et al., 2002). The same issue of selection bias occurs with individuals who enrolled in managed care programs. Variables indicating individuals' health insurance coverage and enrollment in managed care tend to be

known at the beginning of the year and are useful predictor variables, especially when the purpose of the analysis is the prediction of expenditures.

We use several variables to indicate the occupation of an individual. In MEPS, individuals report industry types which is consistent with the Standard Industrial Classification System. We also include a variable indicating the employment status of an individual.

Table 2 provides summary statistics of these variables. We observe that females had higher numbers of inpatient admissions and outpatient visits than males, yet lower average expenditures per event, presumably due to child-bearing. In terms of ethnicity, Asians use the least health care and account for only 4.6% of the total sample size. The effect of region is not clear; individuals from the Midwest have the highest utilization. More educated persons and higher income individuals have fewer but more expensive inpatient admissions, and have a greater number and less expensive outpatient visits. Higher income individuals use more outpatient services and have less hospital admissions. Poorer self-rated health status, physical, mental health and activity related limitations lead to more utilization of both types of services. The effect of industry and managed care are not clear. Employed individuals use less health care. Uninsured individuals use less inpatient and outpatient services than those with insurance coverage.

[Table 2 is about here]

5. Modeling Results

5.1. Inpatient Admissions In-Sample Analysis

Table 3 summarizes the fit of the hospital inpatient admissions. For the frequency component, Table 3 compares a logistic regression fit to a more complex count model, the negative binomial. As anticipated, there is relatively large agreement in the coefficients of these two models and the results are largely in accord with those found in the literature. Demographic factors such as age and sex are statistically significant determinants of hospital admissions. Access factors such as usual source of care had significantly positive effect on hospital ad-

missions. Social economic factors such as marital status, family size and income (middle and high) are strongly significant. Health status factors such as self-reported physical health status and any activity limitation are significant determinants. Employment factors such as unemployment status, education or health services industry classifications are statistically significant. Insurance coverage led to more hospitalization. Results for demographic factors such as ethnicity, access factors such as geographic region, social economic factors such as education, health status factors such as self-rated mental health status and insurance factors such as enrollment in managed care were mixed although coefficient estimates are consistent with the summary statistics shown in Table 2. Variables for Asian and Black ethnicity, Midwest and South region are statistically significant. After controlling for other covariates, the coefficient associated with individuals enrolled in managed care programs was not statistically significant.

[Table 3 is about here]

For the amount per event, Table 3 shows the analysis for inpatient amounts, given that there is a positive expenditure. This analysis is based on a sample of $n=1,797$ using the annual expenditures model with logarithmic expenditures as the dependent variable. We do not report expenditures for the TPM, as the results are largely consistent with those of the annual expenditures model. For expenditures, only a few variables were statistically significant. Age, ethnicity equal to black, geographic region equal to south, income variables and insured status are strongly significant, whereas usual source of care (p -value 0.059) is somewhat statistically significant. Although the mixed linear model was initially specified, there was not sufficient heterogeneity in the data to support including a variance component (α_i) and only a linear regression model is reported. Interestingly, the frequency portion has no significant effect on expenditures (p -value of COUNT_IP is 0.323).

The amount component had fewer variables that significantly impact expenditure per event when compared to the number of significant explanatory variables affecting the count process. We interpret this as individuals having a greater amount of latitude in choosing

when to see a provider as compared to decisions concerning the level of care required. This result is consistent with the two-stage decision making process in the Rand Health Insurance Experiment in using TPMs (Keeler and Rolph 1988; Manning et al. 1987) and the principal-agent model suggested by Zweifel (1981).

5.2. Outpatient Visits In-Sample Analysis

Table 4 summarizes the fit of the outpatient visits. As has been well-documented in the literature, one-part models paint a very different picture than models that decompose the frequency and amount. For example, the one-part model shows a significant positive effect of FEMALE on (logarithmic) expenditures. In contrast, the annual expenditures model decomposes the impact on total expenditures of the frequency component and the expenditure per event component. There is a significant positive effect for FEMALE for the number of stays in the first part, but a significant negative effect for FEMALE for expenditures per event in the second part. Moreover, Table 4 shows important differences between the logistic and negative binomial regression models. For example, the logistic regression shows strong statistical significance for the MIDWEST variable that is not evident in the negative binomial model.

Table 4 also shows the analysis for outpatient event-level expenditures based on a sample of $n=12,970$ expenditures. Here, the annual expenditures model employs a mixed linear model with logarithmic expenditures as the dependent variable. Unlike the inpatient analysis, there were sufficient observations to support including a variance component to account for correlations among expenditures within a person. The variance component estimate was 0.251 with a standard error of 0.006. Thus, the intra-class correlation was 19.3% ($= 0.251/(0.251+1.050)$).

Table 4 shows that all but near poor income level variables for the expenditure per event are positively statistically significant. Those with poor, fair or good physical health (relative to excellent) or some activity limitation also spend significantly more per event. Further, age, ethnicity equal to Native, geographic region equal to south and midwest and insure

status are statistically significant.

Interestingly, Table 4 shows that the number of outpatient visits (COUNT_OP) is a statistically significant negative influence on the typical expenditure per visit, whereas the count variable for inpatient stays was not significant in Table 3.

[Table 4 is about here]

5.3. Point Prediction

Table 5 summarizes the predictive ability of competing models by comparing the held-out 2004 expenditures, S_i , to point predictors, \hat{S}_i . For consistency, both S_i and \hat{S}_i are expressed in terms of dollars for each model (despite the transformations used in the modeling). To assess the proximity of \hat{S}_i to S_i , we focus on the mean absolute percentage error, $MAPE = n^{-1} \sum_{i=1}^n |S_i - \hat{S}_i| / \hat{S}_i$. There are several other measures that one could use; to underscore this point, Table 5 also gives results for the mean absolute error $MAE = n^{-1} \sum_{i=1}^n |S_i - \hat{S}_i|$.

For the inpatient panels 8 and 9, four models generating the smallest values of $MAPE$ were one version of the one-part model, one version of the TPM and two versions of the annual expenditures model. For the one-part model and TPM, these were using a logarithmic expenditure as the dependent variable and a smearing adjustment for prediction. For the annual expenditures model, these were the base version and the “classic” version (represented in equations 1 and 2 respectively), both with a smearing adjustment. Of these four, the annual expenditure predictions were smallest but the difference among the four is small; the largest difference is between the classic annual expenditures with a smear adjustment and the one-part model with smear adjustment and this amounts to only a 7 - 8% difference depending on the panel.

For the outpatient panels 8 and 9, two versions of the TPM and the same two versions of the annual expenditures model generated the smallest values of $MAPE$, again with the annual expenditure predictions being the most accurate. For the TPM, the two versions were assuming logarithmic expenditures as the dependent variable with smearing adjustment and gamma distributed expenditures. As anticipated, the differences are larger; the largest

1
2
3
4 difference is between the classic annual expenditures with a smear adjustment and the TPM
5 with log of expenditures smear adjustment of 13 to 17% depending on the panel. Intuitively
6 the difference is larger for the outpatient models as compared to the inpatient models due
7 to the inpatient distribution having a greater mass at zero and thus placing lower weight on
8 modeling the entire remainder of the frequency distribution.
9

10
11 The annual expenditure predictions compared favorably for both panels 8 and 9. In
12 panel 9, because there was no 2003 experience for these individuals, we used zero to predict
13 individual random effects. In panel 8 with 2003 experience, we used standard best linear
14 unbiased predictors (known by the acronym BLUP).
15

16
17 Comparing results for inpatient and outpatient models, we found it surprising that the
18 “classic” predictor (in equation 2) outperformed our base version (in equation 1). For inpa-
19 tient data, Table 3 shows that the count is not significantly related to amount and so one
20 would expect the classic version to do at least as well as our base version. However, for
21 outpatient data, Table 4 shows that the count is significantly related to amount; the base
22 predictor uses this information unlike the classic version. Nonetheless, the two estimators are
23 close and our comparison is based on this single application. We feel that both estimators
24 can be useful.
25

26
27 Although they provide some useful information, the mean absolute error results in Table 5
28 are more difficult to interpret. Average expenditures are close to the one-part *MAE* because
29 the average expenditures model has predictions close to zero, especially for inpatient data.
30 This is the reason for the large values of *MAPE* for the one-part model. Using the average
31 expenditures model is not a serious contender - we include it as a benchmark for comparison
32 purposes. However, particularly for inpatient, the one-part model does have the smallest
33 value of *MAE*. This underscores the difficulty of using point predictions to assess alternative
34 models where the distribution is a mixture of a discrete and continuous component.
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

5.4. Predictive Distributions

We present only distributions for inpatient admissions, as these events were more infrequent than outpatient services and thus more difficult to predict. Table 6 shows that the distribution for small group sizes have large masses at zero, combined with continuous components. As the group size increases, the probability of the average expenditure equaling zero diminishes. Figure 1 complements Table 6 by showing the distribution for large group sizes. There, we see the clear effects of the central limit theorem; the distribution of average expenditures becomes less skewed and more normally distributed as group size increases. Figure 1 also shows that the variability of the average decreases as group size increases.

Table 6 also summarizes the performance of our point predictors. The predictor in Section 3.2 is labeled as “point” and when multiplied by the smearing factor, it is labeled as “smear.” Table 6 underscores the fact that, for small group sizes, a single point predictor is not helpful, as it does not capture the discrete and continuous components of the predictive distribution. For larger group sizes, the smearing estimate does a better job at summarizing the center, at least as measured by the mean and median of the predictive distribution.

[Table 6 is about here]

[Figure 1 is about here]

6. Discussion

The annual expenditures model is intuitive when individuals accumulate expenditures in a fixed period of time through multiple health care events. Similar concepts have been used by researchers from other fields in the actuarial and insurance literature (Klugman, Panjer and Willmot 2008; Pinquet 1998) and joint modeling of purchase decision and amount in the marketing literature (van Praag and Vermeulen 1993; Boatwright, Borle and Kadane 2003). This research models total health expenditures with the consideration of multiple events. Our empirical analysis based on the MEPS validates the difference in the frequency and

1
2
3
4 expenditures characteristics between the inpatient and outpatient utilization and expendi-
5 tures. The use of a predictor of counts for the amount portion was a significant predictor
6 for outpatient expenditures. While using a predictor of counts was not significant for inpa-
7 tient expenditures for these data, the approach may be useful for predicting expenditures
8 for chronic diseases, where more hospitalizations occur per individual per year.
9

10
11 The structure of the data will remind some readers of clustered, or multilevel, data, (Rau-
12 denbush and Bryk 2002), as each person may have multiple expenditures. However, unlike
13 the usual multilevel model, the number of replications is random. To reflect this fact, van
14 Praag and Vermeulen (1993) used the phrase “endogenous recording of observations” in their
15 study that examined the amount of food expenditures.
16

17
18 When conducting our four regression analyses (inpatient/outpatient, frequency/amount),
19 we used the same set of covariates. This was done to minimize issues of model selection
20 so we could focus on the predictive features of our annual expenditures model. Clearly,
21 independent analyses of inpatient and outpatient expenditures do not require the same sets
22 of covariates. Moreover, a strength of our hierarchical model structure is that one does not
23 need the same set of covariates for the frequency and amount components.
24

25
26 The in-sample analysis based on the annual expenditures model provides support for the
27 principal-agent model. We found that explanatory variables such as demography, education,
28 regional, health status and economic factors significantly explained the variation in counts of
29 inpatient admissions and outpatient visits; however, most of the variables were not significant
30 in explaining expenditures per visit. This is consistent with the episode approach in Keeler
31 and Rolph’s (1988) study. A limitation of our study is that we did not differentiate insurance
32 coverage variables with respect to types of services. As with the TPM, a limitation of our
33 approach is that we assume that there are no omitted variables that may affect the frequency
34 and amount distributions.
35

36
37 We assume that event expenditures data are complete in the sense that they are not
38 truncated nor are they censored. Another approach would be to link the events to episodes
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

of illness. The clear strength of this approach is that the illness provides a context for the explanation of medical expenditures. The limitation is that episodes, by definition, are not restricted to calendar years. At the beginning of a calendar year, complete expenditures may be reported for episodes that occurred in the prior calendar year. At the end of a year, there may be episodes that are just beginning for which complete expenditures are not available. When restricted to a specific financial time period, such as a calendar year, episode data are both truncated and censored, thus requiring more complex statistical methodology. We leave this a potential topic for future research.

Our statistical models, although complex, allow us to retain the important features of health care expenditures when making predictions. Our predictive distributions show large probabilities of zeros corresponding to no health care utilization, long-tail behavior for individuals and small groups, as well as readily predictable, symmetric, distributions for large groups. Because they are intuitively plausible, these predictive distributions can serve as important tools for managing healthcare expenditures.

Using more detailed event-level expenditures, our annual expenditures model is flexible and can readily be modified for other data needs. With the hierarchical structure that we propose, one could easily use alternative count models such as Poisson, zero-inflated Poisson or latent variable models (e.g., Cameron and Trivedi 1998). For event-level expenditures, one could also entertain a random coefficients generalized linear model (e.g., Frees 2004) as an alternative to our linear mixed model. Here, the random coefficient would account for the intra-individual correlation of expenditures and the generalized linear model specification would address the long-tail nature of claims. The focus of this paper was on the simpler specification of linear mixed models that permit closed-form predictors but one could easily see how these extensions would be worthwhile for different data sets.

7. Acknowledgements

This research was supported by the Agency for Healthcare Research and Quality, Grant Number R03 HS16519, the National Science Foundation, Grant Number SES-0436274, and the Assurant Health Professorship in Actuarial Science.

APPENDIX

This appendix shows the first derivative of moment generating function of the negative binomial distribution and how to estimate it in the context of generalized linear models.

If the probability mass function is $\Pr(N = n|r, p) = \binom{n+r-1}{r-1} p^r (1-p)^n$, then the moment generating function is

$$M_N(t) = E(e^{tN}) = \sum_{n=0}^{\infty} e^{tn} \Pr(N = n|r, p) = p^r [1 - (1-p)e^t]^{-r}.$$

Thus, the first derivative of $M_N(t)$ is

$$M'_N(t) = p^r r (1-p) e^t [1 - (1-p)e^t]^{-r-1}. \quad (\text{A.1})$$

Straightforward calculations show that the mean is $E(N) = M'_N(0) = (1-p)r/p$. Similarly, the variance can be expressed as $\text{Var}(N) = M''_N(0) - (E(N))^2 = (1-p)r/p^2$.

In generalized linear modeling, the negative binomial distribution is specified by the dispersion parameter, σ , and the mean parameter, $\mu_i = \exp(\mathbf{x}'_i \boldsymbol{\beta}_1)$, through the log-link assumption. With this specification, the mean of N_i is given by $E(N_i) = \mu_i$, and the variance by $\text{Var}(N_i) = \mu_i + \sigma \mu_i^2$.

Equating the means and variances from these two relations, we see that $\sigma = 1/r$ and that p related to the mean through $(1-p)/p = \mu\sigma = \exp(\mathbf{x}' \boldsymbol{\beta}_1) \sigma$.

References

- Blough, D.K., C.W. Madden, and M.C. Hornbrook. 1999. Modeling Risk Using Generalized Linear Models. *Journal of Health Economics* 18: 153-171.
- Boatwright, P., S. Borle, and J. Kadane. 2003. "A Model of the Joint Distribution of Purchase Quantity and Timing." *Journal of the American Statistical Association* 98: 564-572.
- Cameron, A.C., and P.K. Trivedi. 1998. *Regression Analysis of Count Data*. ambridge University Press.
- Deb, P., and P.K. Trivedi. 2002. "The Structure of Demand for Health Care: Latent Class versus Two-Part Models." *Journal of Health Economics* 21: 601-625.
- Duan, N.H., W.G. Manning, Jr., C.N. Morris, and J.P. Newhouse. 1983. "A Comparison of Alternative Models for Demand for Medical Care." *Journal of Business & Economic Statistics* 1(2): 115-126.
- Frees, Edward W. 2004. *Longitudinal and Panel Data: Analysis and Applications for the Social Sciences*. Cambridge University Press.
- Gao, J. 2007. *Modeling Individual Healthcare Expenditures by Extending the Two-Part Model*. Unpublished Dissertation, University of Wisconsin, Madison.
- Goldman, D.P., Hosek, S.D., Dixon, L.S., Sloss, E.M. 2002. The effects of benefit design and managed care on health care costs? *Journal of Health Economics* 14, 401-418.
- Jones, A.M. 2000. "Health Econometrics." In *Handbook of Health Economics*, Vol. 1A. edited by A.J. Culyer and J.P. Newhouse, pp. 265-344. Amsterdam: Elsevier.
- Keeler, E.B., and J.E. Rolph 1988. "The Demand for Episodes of Treatment in the Health Insurance Experiment." *Journal of Health Economics* 7: 337-367.
- Klugman, S., H. Panjer and G.E. Willmot. 2008. *Loss Models: from Data to Decisions, Third Edition*. New York: Wiley.
- Lohr, K.N., R.H. Brook, C. Kamberg, G.A. Goldberg, A. Leibowitz, J. Keeseey, D. Reboussin, and J.P. Newhouse. 1986. "Use of Medical Care in the Rand Health Insurance Experiment: Diagnosis- and Service-specific Analyses in a Randomized Controlled Trial." *Medical Care* 24(9): S1-S87.
- Manning, W.G., J.P. Newhouse, N. Duan, E.B. Keeler, A. Leibowitz, and M.S. Marquis. 1987. "Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment." *The American Economic Review* 77(3): 251-277.
- Manning, W.G., A. Basu, and J. Mullahy. 2005. "Generalized Modeling Approaches to Risk Adjustment of Skewed Outcomes Data." *Journal of Health Economics* 24(1): 65-88.
- Mello, M.M., Stearns, S.C., Norton, E.C. 2002. Do medicare HMOs still reduce health services use after controlling for selection bias? *Health Economics* 11, 323-340.
- Mullahy, J. 1998. "Much Ado About Two: Reconsidering Retransformation and the Two-Part Model in Health Econometrics." *Journal of Health Economics* 17: 247-281.
- Pinquet, J. 1998. "Designing Optimal Bonus-Malus Systems from Different Types of Claims." *ASTIN Bulletin* 28: 205-220.
- Pohlmeier, W., and V. Ulrich. 1995. "An Econometric Model of the Two-Part Decisionmaking Progress in the Demand of Health Care." *The Journal of Human Resources* 30: 339-361.

- 1
2
3
4 van Praag, B.M.S., and E.M. Vermeulen. 1993. "A Count-Amount Model with Endogenous Recording of
5 Observations." *The Journal of Applied Econometrics* 8: 383-395.
6
7 Raudenbush, S.W., and A.S. Bryk. 2002. *Hierarchical Linear Models: Applications and Data Analysis Meth-*
8 *ods*. 2nd eds. London: Sage.
9
10 Rosenberg, M.A., and P.M. Farrell. 2007. "Impact of A Newborn Screening Program on Inpatient Utilization
11 for Children with Cystic Fibrosis." *Technical Paper*.
12
13 Rosenberg, M.A., and P.M. Farrell. 2008. "Predictive Modeling of Costs for a Chronic Disease with Acute
14 High Cost Episodes." *North American Actuarial Journal* 12(1), 1-18.
15
16 Smyth, G. K. and B. Jorgensen (2002). "Fitting Tweedie's Compound Poisson Model to Insurance Claims
17 Data: Dispersion Modelling." *Astin Bulletin* 32, 143-157.
18
19 Zweifel, P. 1981. "Supplier-Induced Demand in a Model of Physician Behavior." In *Health, Economics, and*
20 *Health Economics*, edited by J. van der Gaag and M. Perlman, pp. 245-267. Amsterdam: North-Holland.
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

Table 1. Frequency of Inpatient Admissions and Outpatient Visits

Count	0	1	2	3	4	5-9	10-19	20-49	50-99	100-401	Total	Average
Inpatient	17,322	1,154	184	46	17	12					18,735	0.097
Outpatient	5,765	2,712	1,194	1,352	1,108	2,848	1,873	967	164	32	18,735	5.472

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

24
Table 2. Covariate Descriptions, Mean Counts and Expenditures

Category	Variable	Description	Percent	Mean Counts		Mean Expenditures	
				Inpatient	Outpatient	Inpatient	Outpatient
Demography	FEMALE	1 if female	54.0	0.13	6.75	7,399.25	227.33
		0 if male	46.0	0.06	3.97	12,371.41	267.02
Ethnicity	ASIAN	1 if Asian	4.6	0.04	3.33	6,694.69	201.76
	BLACK	1 if Black	14.8	0.13	5.03	8,341.67	255.25
	NATIVE	1 if Native American	1.0	0.12	6.26	8,753.93	347.22
	WHITE	1 if White, multiple races and native Hawaiian	80.0	0.09	5.67	9,026.21	238.39
Region	NORTHEAST	1 if Northeast	14.8	0.08	6.32	9,515.30	216.03
	MIDWEST	1 if Midwest	19.9	0.11	6.29	10,467.00	256.81
	SOUTH	1 if South	39.0	0.11	4.90	7,970.48	255.38
	WEST	Base category	26.3	0.07	5.21	8,526.70	221.87
Access to care	USC	1 if have Usual Source of Care	71.1	0.11	6.83	8,958.72	242.43
		0 if otherwise	28.9	0.05	2.12	8,302.56	225.53
Education	COLLEGE	1 if college or higher degrees	26.6	0.08	6.59	9,985.99	222.64
	HIGHSCH	1 if high school degree	44.8	0.09	5.41	7,740.56	249.05
Marital Status		Base category is lower than high school degree	28.6	0.11	4.53	9,502.30	253.62
	MARRIED	1 married	55.5	0.09	5.50	9,078.02	242.70
	WIDOWED	1 if widowed	2.3	0.22	9.07	7,998.20	291.06
	DIVSEP	1 if divorced or separated	14.0	0.14	7.66	9,192.31	243.60
Income compared to poverty line		Base category is never married	28.2	0.07	4.03	8,259.00	229.81
	HINCOME	1 if high income	32.7	0.07	6.16	9,892.03	237.96
	MINCOME	1 if middle income	29.7	0.08	5.21	11,545.05	238.60
	LINCOME	1 if low income	15.4	0.10	4.26	7,673.91	260.67
Self-rated Physical health	NPOOR	1 if near poor	5.5	0.12	4.79	7,027.50	243.54
		Base category is poor/negative	16.8	0.17	5.94	6,782.33	235.12
Self-rated mental health Any activity limitation	POOR	1 if poor	4.0	0.46	15.85	8,709.11	292.24
	FAIR	1 if fair	11.1	0.15	8.86	10,850.09	248.75
	GOOD	1 if good	28.5	0.10	5.87	8,668.59	238.86
	VGOOD	1 if very good	30.8	0.07	4.39	8,302.01	226.81
Self-rated mental health Any activity limitation		Base category is excellent health	29.3	0.05	3.28	7,676.99	218.24
	MNHPOOR	1 if poor or fair mental health	7.7	0.23	11.40	10,695.57	248.73
		0 if good, very good, excellent mental health	92.3	0.09	4.98	8,442.07	239.05
	ANYLIMIT	1 if any functional or activity limitation	22.3	0.21	11.51	11,191.70	262.43
	0 if otherwise	77.7	0.06	3.74	6,631.79	221.44	

Continued on next page

Table 2. – continued from previous page

Category	Variable	Description	Percentage		Mean Counts		Mean Expenditures	
			Inpatient	Outpatient	Inpatient	Outpatient	Inpatient	Outpatient
Industry Classification	NATRESOURCE	1 if natural resources	1.6	2.52	0.05	2.52	11,077.89	684.66
	MINCONST	1 if mining or construction	4.9	3.55	0.07	3.55	7,941.41	230.53
	MANUFACT	1 if manufacturing	8.8	4.28	0.05	4.28	10,447.52	267.11
	SALES	1 if sales	10.0	4.41	0.07	4.41	8,426.83	234.51
	TRANSINFO	1 if transportation, utilities and information	4.7	5.00	0.06	5.00	7,874.48	216.04
	FINANCE	1 if finance, insurance, or real estate	4.3	5.50	0.07	5.50	6,127.85	227.63
	PROFSERV	1 if professional services	7.0	4.51	0.05	4.51	9,237.99	225.19
	EDUCHEALTH	1 if education, health and social services	14.6	6.19	0.09	6.19	7,333.76	232.42
	LEISURE	1 if leisure and hospitality	5.9	3.29	0.07	3.29	11,948.42	211.62
	PUBADMIN	1 if public administration	3.4	6.70	0.07	6.70	9,175.54	215.44
	MILITARY	1 if active military	0.2	4.19	0.02	4.19	4,510.44	384.88
	OTHERSERV	1 if other services	3.8	3.68	0.04	3.68	13,925.95	272.95
		Base category is inapplicable, uncertain or industry unknown	29.8	7.29	0.17	7.29	8,986.25	249.90
	Employment Status	UNEMPLOY	1 if ever unemployed during 2003 0 if otherwise	23.0	8.40	0.20	8.40	9,416.56
Insurance coverage	INSURE	1 if covered by public or private health insurance in January, 2003	77.0	4.60	0.07	4.60	8,350.30	234.85
		0 if covered by public or private health insurance in January, 2003	72.9	6.51	0.11	6.51	9,420.34	242.06
Managed Care	MANAGED	0 if have no health insurance in Jan 2003	27.1	2.67	0.06	2.67	5,962.72	220.23
		1 if enrolled in an HMO or a gatekeeper plan 0 if otherwise	60.0	6.09	0.10	6.09	9,252.08	237.79
			40.0	4.56	0.10	4.56	8,251.83	246.36

Table 4. Outpatient Visits In-Sample Models

Parameter	One Part		Two Part Model		Annual Expenditures Model			
	Estimate	<i>p</i> -value	Logistic Regression		Negative Binomial		Mixed Linear Model	
			Estimate	<i>p</i> -value	Estimate	<i>p</i> -value	Estimate	<i>p</i> -value
Intercept	0.507	0.001	-1.607	0.001	-0.221	0.010	4.415	0.001
AGE	0.013	0.001	0.005	0.009	0.003	0.027	0.004	0.001
FEMALE	1.027	0.001	0.912	0.001	0.514	0.001	-0.087	0.001
ASIAN	-0.733	0.001	-0.436	0.001	-0.473	0.001	-0.053	0.131
BLACK	-0.291	0.001	-0.231	0.001	-0.148	0.001	0.019	0.335
NATIVE	0.055	0.787	-0.248	0.190	-0.021	0.888	0.301	0.001
NORTHEAST	0.233	0.001	0.179	0.004	0.022	0.654	-0.011	0.573
MIDWEST	0.353	0.001	0.248	0.001	0.066	0.134	0.090	0.001
SOUTH	0.091	0.069	0.077	0.105	-0.141	0.001	0.065	0.001
USC	1.661	0.001	1.213	0.001	0.685	0.001	0.003	0.887
HIGHSCH	0.218	0.001	0.117	0.013	0.134	0.001	0.031	0.072
COLLEGE	0.420	0.001	0.249	0.001	0.302	0.001	-0.003	0.893
MARRIED	0.375	0.001	0.290	0.001	0.210	0.001	0.001	0.948
WIDOWED	0.077	0.582	0.046	0.757	0.013	0.899	0.025	0.534
DIVSEP	0.177	0.012	0.099	0.150	0.111	0.034	0.021	0.349
FAMSIZE	-0.223	0.001	-0.175	0.001	-0.129	0.001	0.001	0.854
NPOOR	-0.125	0.194	-0.101	0.259	-0.068	0.361	0.022	0.502
LINCOME	0.017	0.814	0.019	0.774	-0.107	0.050	0.068	0.006
MINCOME	0.093	0.157	0.076	0.215	-0.002	0.967	0.086	0.001
HINCOME	0.267	0.001	0.241	0.001	0.069	0.211	0.068	0.004
POOR	1.833	0.001	1.254	0.001	0.866	0.001	0.191	0.001
FAIR	1.286	0.001	0.824	0.001	0.621	0.001	0.106	0.001
GOOD	0.744	0.001	0.451	0.001	0.452	0.001	0.079	0.001
VGOOD	0.438	0.001	0.330	0.001	0.207	0.001	0.011	0.544
MNHPOOR	0.284	0.001	0.184	0.041	0.232	0.001	0.005	0.815
ANYLIMIT	1.131	0.001	0.812	0.001	0.721	0.001	0.045	0.005
MINCONST	-0.204	0.017	0.141	0.319	-0.110	0.107	-0.063	0.066
EDUHEALTH	0.248	0.001	-0.119	0.123	0.088	0.048	0.038	0.036
UNEMPLOY	0.302	0.001	0.025	0.774	0.201	0.001	-0.028	0.104
INSURE	0.882	0.001	0.189	0.002	0.382	0.001	0.046	0.021
MANAGEDCARE	0.407	0.001	-0.071	0.455	0.152	0.001	0.025	0.105
COUNT_OP							-0.007	0.001
Dispersion	6.979	se=0.072			1.513	se=0.030	1.050	se=0.005
Variance Component		89,666.4					0.251	se=0.006
-2 Log Likelihood				17,956.53		-139,824.8		290,919.4

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

Table 5. Out-of-Sample Point Prediction Accuracy

	Inpatient		Outpatient	
	Panel 8	Panel 9	Panel 8	Panel 9
Mean Absolute Proportional Error				
One Part	15,916.471	1,863.901	12.841	23.534
One Part with Smear	1.857	1.763	1.275	1.421
TPM with Log Expenditures	2.619	2.474	2.277	2.768
TPM with Log Expenditures with Smear	1.783	1.711	1.220	1.342
TPM with Gamma Expenditures	1.870	1.802	1.183	1.295
Ann Expenditures	2.509	2.382	1.967	2.324
Ann Expenditures Classic Version	2.462	2.340	1.798	2.101
Ann Expenditures with Smear	1.733	1.669	1.101	1.168
Ann Expenditures Classic Version with Smear	1.709	1.647	1.060	1.109
Average Expenditure	848.42	795.73	1,201.55	1,501.28
Mean Absolute Error				
One Part	849.40	796.72	1,057.07	1,314.55
One Part with Smear	1,402.72	1,360.55	4,725.42	4,923.85
TPM with Log Expenditures	1,190.07	1,134.87	1,064.37	1,293.00
TPM with Log Expenditures with Smear	1,507.00	1,457.91	1,337.00	1,454.61
TPM with Gamma Expenditures	1,471.16	1,430.21	1,301.74	1,433.00
Ann Expenditures	1,202.58	1,147.53	1,082.29	1,305.41
Ann Expenditures Classic Version	1,216.29	1,161.13	1,088.68	1,295.60
Ann Expenditures with Smear	1,529.72	1,480.81	1,367.44	1,469.93
Ann Expenditures Classic Version with Smear	1,557.14	1,509.07	1,538.47	1,607.54

Table 6. Predictive Distribution, by Group Size

Group Size	Simulated Predictive Distribution							Point Predictions	
	Percentiles							Point	Smear
	1st	10th	25th	50th	75th	maximum	average		
1	0.00	0.00	0.00	0.00	0.00	85,525.94	322.93	162.21	306.46
5	0.00	0.00	0.00	0.00	576.89	44,083.00	888.94	437.24	826.09
25	0.00	0.00	36.58	268.13	804.75	19,865.10	640.97	321.07	606.59
50	27.59	263.01	548.02	1,083.53	1,833.99	14,851.41	1,370.29	669.81	1,265.48
250	286.85	489.59	663.74	900.49	1,194.43	6,943.49	972.92	474.00	895.53
9,657	825.13	894.42	933.99	979.88	1,029.71	1,395.39	983.53	479.26	905.47

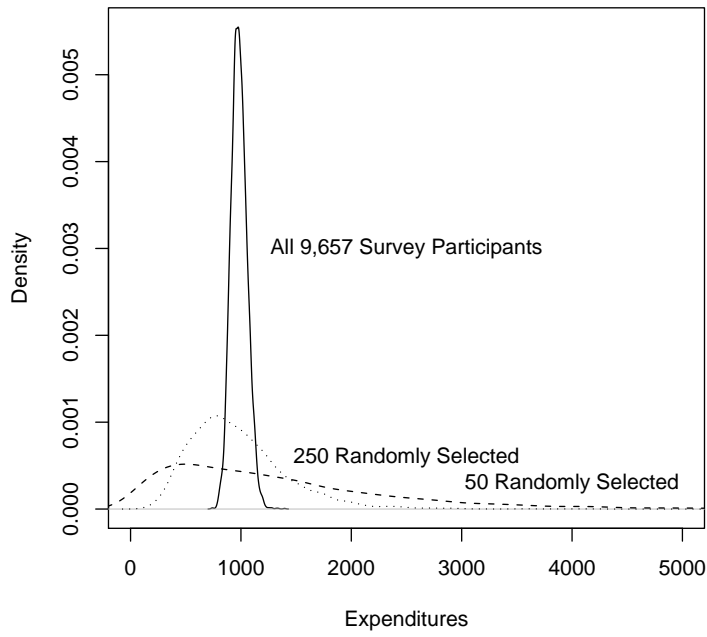


Figure 1. Comparison of Predictive Distributions for Group Sizes 50, 250 and 9,657